SESSION

MEDICAL SYSTEMS AND DEVICES + MONITORING TOOLS + RELATED METHODOLOGIES AND ALGORITHMS

Chair(s)

TBA

ISBN: 1-60132-437-5, CSREA Press ©

ISBN: 1-60132-437-5, CSREA Press ©

Multidimensional Mobility Metric for Continuous Gait Monitoring using a Single Accelerometer

Ik-Hyun Youn[1], Deepak Khazanchi[1], Jong-Hoon Youn[1], and Ka-Chun Siu[2]
[1]College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE, USA
[2]College of Allied Health Professions, University of Nebraska Medical Center, Omaha, NE, USA

Abstract – *Mobility has been measured in the form of gait analysis – a method used to assess the physical functioning of a human. Many health-related clinical and pathological gait analysis applications have been proposed in the last decades. Gait analysis research has been predominantly conducted in laboratories in a discipline-specific manner. The main measurement approach is to monitor intrinsic gait patterns in natural settings such as home or work. Our proposed mobility assessment approach is to assess gait along three major gait dimensions: intensity, symmetry, and variability. This multidimensional metric uses a single accelerometer to assess the three dimensions so that the gait metric is continuously applied to mimic the natural human gait in daily life. Initial experimental results demonstrate its practical use in different gait patterns with various gait speeds.*

Keywords: Mobility metric, gait metric, accelerometer, continuous monitoring, gait speed

1 Introduction

Mobility in the recent past has predominantly been studied using gait analysis. Gait analysis is used to assess individual conditions that affect a person's ability to walk. Gait research broadly incorporates quantification and interpretation of gait patterns. Gait analysis was initially developed to assess the subtle changes in human gait patterns that were usually not noticeable by the naked eye. Many measurement approaches and technologies such as infrared cameras and pressure sensor platforms have been used to analyze human gait. For example, in typical gait analysis studies, high-speed digital cameras placed around a walkway are used. Several reflexive markers are attached at many different locations of the body to track motions of the body while walking. Another well-known application of human gait analysis focuses on pathology. The pathological gait analysis measures reflect underlying symptoms of diseases such as cerebral palsy and stroke. This analysis can also be applied to rehabilitation engineering, sports training to improve performance, and biometric identification.

In the last decade, gait analysis has become more popular in the general population, particularly because gait is now well known as a fundamental physical activity for maintaining health levels [2]. With the advent of wearable devices such as Fitbit, many researchers are trying to quantify gait patterns outside of laboratories [13]. Continuous gait monitoring using wearable devices enables us to quantify gait parameters using natural gaits. Many of the available wearable sensors try to measure physical activity by using parameters such as duration, number of steps, and energy consumption. The main advantage of wearable devices is better portability for assessment of gait patterns. For instance, measuring the number of steps using a wearable device on the waist or wrist is a popular way to connect gait to better human physical health. Although gait features from wearable devices are able to represent mobility patterns, the information is still insufficient to apply to a comprehensive understanding of human mobility to make useful predictions of activity [7].

A major problem for people who investigate health issues is to continuously monitor gait functioning and to timely interpret the data to prevent loss of gait functional abilities and improve the quality of life [17]. Although proposed wearable sensor-based gait analysis approaches have introduced great portability to assess gait patterns, many of them are discipline-specific. Moreover, the lack of attempts to comprehensively depict gait patterns has been observed [15]. To address this gap, in this paper we propose a comprehensive mobility evaluation metric for continuous gait monitoring using a single wearable sensor. The mobility metric consists of three gait dimensions: symmetry, variability, and intensity.

The rest of the paper is organized as follows. Section 2 introduces the proposed mobility evaluation metric, which uses wearable sensor-based gait recognition and feature extraction techniques. In section 3, we describe a technique to acquire data on the multidimensional components of mobility. In the next section, we report on an experimental study to demonstrate the efficacy of this metric. In the last section, we discuss the experimental results and their implications.

2 Comprehensive Mobility Measurement

Human gait involves complicated sequential commands by neural control of locomotion to the body muscle [5]. As the center of gravity moves forward based on bipedal motion, potential energy is converted to kinetic energy. Maintaining efficient gait requires minimizing this complicated mechanism smoothly [6]. Based on the literature, human gait is generally

defined as a bipedal movement with dynamic balance control [1 and 14]. We propose a comprehensive mobility assessment metric for continuous monitoring using a single accelerometer-based system. We conceptualize that the mobility patterns of a person can be described in terms of the three key dimensions described in Table 1.

Table 1. Mobility dimensions.

Dimension	Definition	Functionality
Intensity	Walking dynamism	Locomotive ability to walk
Symmetry	Bilateral gait balance control	Bilateral balance
Variability	Stride-to-stride variation	Natural fluctuation

2.1.1 Intensity

Intensity describes the active locomotion of a person. As opposed to smooth wheel motion, bipedal human movement generates dynamic and cyclic behaviors. We define active locomotion as a degree of dynamic gait. This functioning eventually enables a person to reach an intended location with a certain degree of dynamic locomotion. By considering the activeness of gait, we are able to measure degrees of active locomotion that describe how active the gait is. If we solely consider gait intensity, a higher degree of intensity demonstrates a better ability to actively propel a center of gravity forward.

2.1.2 Symmetry

Symmetry is a dimension that reflects the degree of bilateral balance of gait. Symmetry has been defined as a good balance between the actions of the limbs [3 and 11]. When a person has the capability to locomote between two locations, there will be patterns of locomotion. The bilateral balance of each stride is one of the gait functionalities that represents bilateral balance while walking. Bilateral balance is considered to be an efficacious method to analyze the balance control ability of humans. Gait symmetry includes the effect of limb dominance and different levels of muscle contributions in gait in order to achieve control and propulsion in gait mechanisms [10]. Gait symmetry has been used to understand the risk for falls. When a difference between two successive bilateral gait parameters increases, we can interpret this tendency as a more asymmetric gait pattern in a stride.

2.1.3 Variability

As a person keeps maintaining gait cycles, there will be a natural fluctuation over multiple strides. The intrinsic stride-to-stride fluctuation is natural due to bipedal movement, and it is also an important aspect of human gait characteristics. The gait cycle variation is considered as a signature of individual gait patterns or an indicator of malfunction of balance control. The dimension of variability can then be defined in terms of the stride-to-stride fluctuation in human gait patterns and the natural variations that occur in locomotive performance for multiple gait cycles [8]. Variability in gait is a useful indicator of impaired locomotor control in a clinical study and a parameter in the evaluation of mobile abnormality [4]. The ability to mitigate gait variability within certain criteria is important in order to maintain effective mobility. Generally, increased variability in a movement pattern indicates less cooperative behavior among the components of the underlying control system. Decreased variability generally indicates highly stable and cooperative behavior [12].

3 Measurement of Mobility Dimensions

Each mobility dimension is measured using attributes collected from our wearable sensor system. Because a single accelerometer is used for the wearable measurement, attributes are derived from three-dimensional acceleration to assess gait parameters.

3.1.1 Intensity

The vector magnitude (VM) of 3-D acceleration is used to identify the intensity of gait. Although step time easily represents locomotive capability to get to a destination, measuring time to reach the intended destination is inappropriate for a continuous monitoring design. Instead of time to get to a destination, we propose to directly use the magnitude of acceleration while walking. In particular, by taking the vector magnitude of acceleration, we can measure pure acceleration generated for center of gravity propulsion. A person who generates greater intensity in each gait cycle will tend to have a higher center of gravity propulsion.

3.1.2 Symmetry

The symmetry index (SI) proposed by Robinson et al. is used to assess asymmetry of gait [9]. Using SI, we quantify left and right step time gap as a degree of gait symmetry. Although the symmetry relies on complicated mechanisms from several components of the body, comparing a temporal gait parameter that is the left and right step time of each stride is an effective way of measuring symmetry.

3.1.3 Variability

Stride time fluctuation is used to evaluate gait variability. Since we define gait variability as a natural inconsistency of gait over multiple strides, the natural variation of a temporal gait parameter is efficiently represented by the accelerometer-based sensor system used to measure gait. The standard deviation (SD) of multiple stride times is used to calculate the degree of variability in this study.

4 Mobility Metric Computation

Mobility in terms of the gait metrics proposed in the previous section is measured using a single sensor wearable system. The three-dimensional acceleration measured by the sensor is used to recognize each step and extract associated gait features. Initially, a gait recognition algorithm is applied to

determine the initiation timing of each step. Once the algorithm detects gait cycles, gait parameters are extracted from each gait cycle. In order to make this work self-contained, we briefly revisit the technical achievements in our preliminary studies and relevant background information [16].

4.1 Gait Recognition

For accurate gait cycle recognition, the gait recognition technique searches peaks stemming from heel-strike actions. In Figure 1, we observe regular peaks from heel strikes. The actions also substantially change jerk, which is defined as a change in the rate of acceleration as shown in Figure 2.

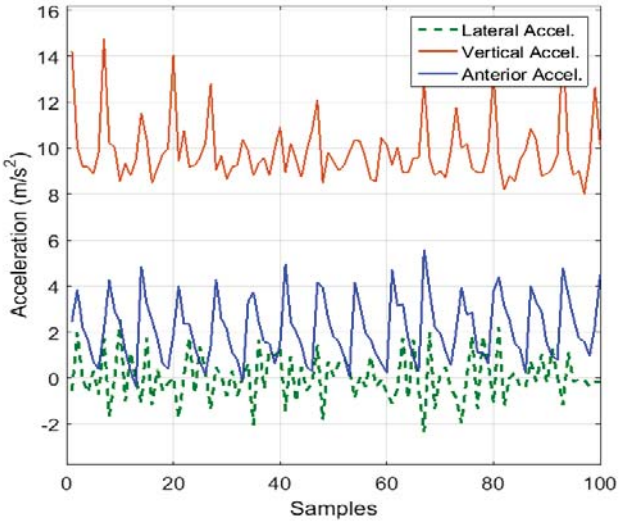


Figure 1. Raw acceleration data.

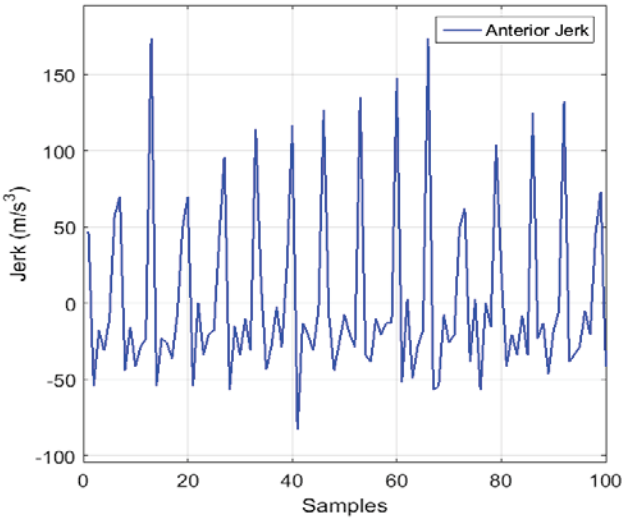


Figure 2. Jerk data from acceleration.

Compared to an acceleration pattern, a jerk pattern shows clear peak moments as shown in Figure 2. In particular, the anteroposterior directional jerk is used to identify each heel strike, rather than other directional jerks. In Figure 3, recognized gait cycles are shown. The vertical dotted lines indicate calculated gait cycles. Horizontal distances of each vertical line as step times are computed at this phase in order to extract gait dimension.

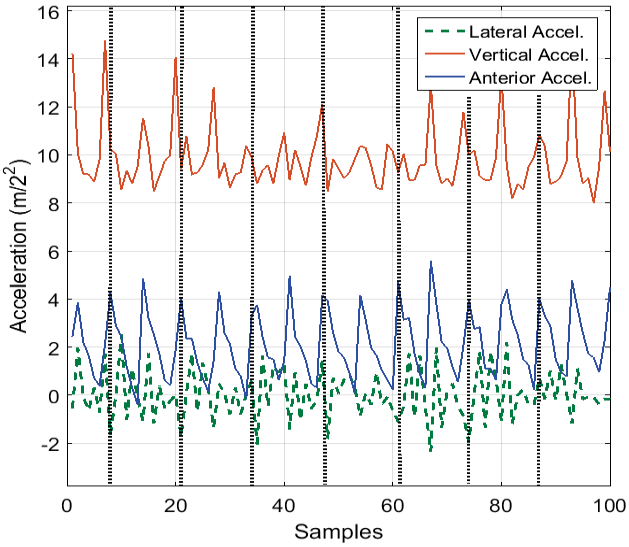


Figure 3. Recognized steps with vertical dotted lines.

Basic gait parameters listed in Table 2 are used to determine gait dimensions in the proposed dimensions for mobility. The gait parameters are directly calculated from raw acceleration data.

Table 2. Basic gait parameters extracted from sensor.

Type	Name	Description
Acceleration	Step acceleration	3-D acceleration of each step
Step time	Step duration	Individual step time
Stride time	Standard deviation	Duration of two successive step times

4.2 Mobility Dimension Extraction

The three mobility dimensions are eventually obtained at the end of the following extraction process. Using mathematical approaches briefly discussed in section 2 regarding gait attributes, the basic gait parameters are translated to associated mobility dimensions using the following equations (refer to Table 3).

Table 3. Attributes of gait metric.

Name	Attribute
Intensity	Vector magnitude of 3-D acceleration
Symmetry	Symmetry index of successive step time
Variability	Standard deviation of stride times

Intensity is defined as an average vector magnitude of 3-D acceleration for each step. Equation 1 below is used to calculate intensity using 3-D step acceleration. Equation 1 is used to calculate the average vector magnitude of each step, where ***m*** is the number of acceleration samples in each step and *x, y,* and *z* represent each directional acceleration.

$$VM = \frac{1}{m}\sum_{i=1}^{m}\sqrt{\left(x_i^{\,2} + y_i^{\,2} + z_i^{\,2}\right)} \qquad (1)$$

Symmetry is assessed using the symmetry index [Robinson, 1987] as shown in Equation 2. Symmetry of gait is calculated using two consecutive step times, where T_{odd} is the odd number step time and T_{even} is the even number step time.

$$SI = \frac{T_{odd} - T_{even}}{\frac{1}{2}\left(T_{odd} + T_{even}\right)} \times 100 \qquad (2)$$

Finally, variability is computed using recognized stride times. We previously defined gait variability as a standard deviation of stride times. Stride times are entered into Equation 3 to compute the variability of gait. In this equation, T denotes stride time and $\overline{T}$ is the mean of the stride times.

$$SD = \sqrt{\frac{\sum (T - \overline{T})^2}{N - 1}} \qquad (3)$$

5 Experimental Study

We conducted an experimental study using the proposed multidimensional mobility metric. We chose varied walking speeds as an experimental protocol to demonstrate the practical use of the proposed gait metric. Because previous studies found that gait patterns changed when walking speed changed, we hypothesized that the proposed gait metric would reflect previous findings at various walking speeds.

5.1 Protocol

Table 4. Gait speed table.

Gait Speed	Gait Speed Control
Slowest	40 percent slower speed than the preferred
Slower	20 percent slower speed than the preferred
Preferred	Self-determined gait speed
Faster	20 percent faster speed than the preferred
Fastest	40 percent faster speed than the preferred

Twelve healthy subjects volunteered to participate in our experiment. Five male and seven female subjects with ages between 21 and 33 participated in the experiment. Subjects were required to walk a 30-meter long walkway in a building. They wore comfortable shoes for the experiments. Before the data collection, the subjects walked the walkway twice in order to determine their preferred walking speeds. Table IV shows five different gait speeds that were computed from their preferred gait speeds. A single accelerometer was worn at the lower back of the trunk in each subject.

5.2 Results

Using the data collected from our experiment, we analyzed gait acceleration for five different gait speeds. Figure 4 shows the typical acceleration patterns from the slowest, preferred, and fastest gait speeds as an example.

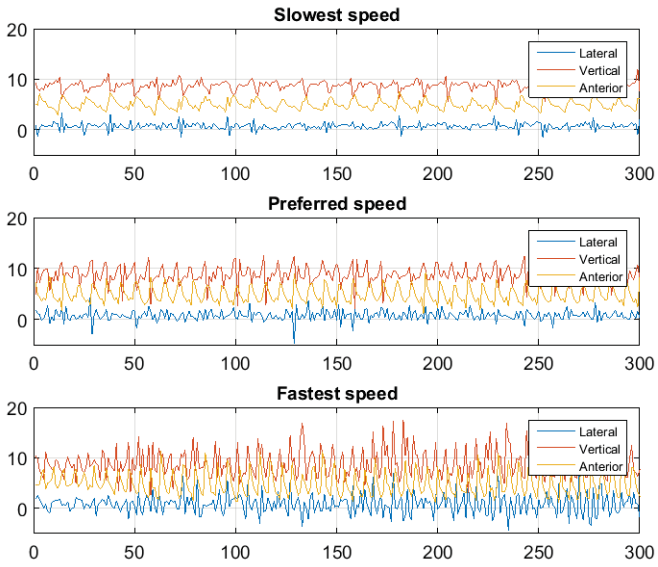


Figure 4. Acceleration from three different gait speeds.

The proposed comprehensive mobility measurement evaluated each gait pattern, Figures 5-7 illustrate these experimental results. As shown, five different gait speeds showed a distinctive status of gait using our proposed metric.

Figure 5 shows the intensity pattern for each gait speed. Intensity highly correlates with gait speeds with clear boundaries between different speeds. For example, the intensity value for the fastest and faster gait speed distributes independently without overlap between them. Figures 6 and 7 show similar patterns along with gait speeds. Typically, both gait dimensions yield the highest values of symmetry and variability at the lowest gait speeds. The results reflect previous findings; slow walking requires more tight dynamic balance control than faster gait, hence a slow gait likely has more inconsistency.

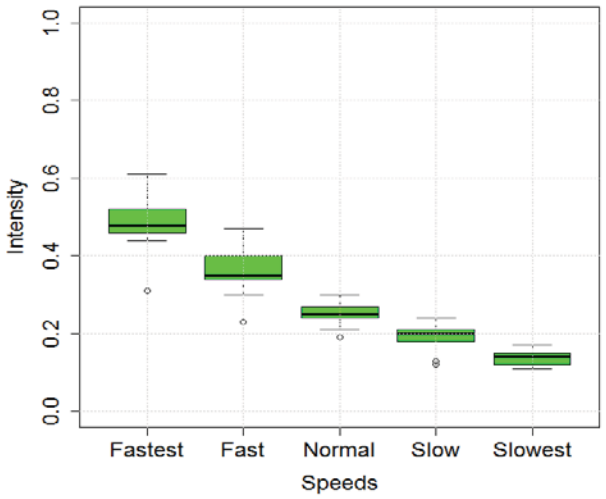


Figure 5. Intensity result for five different gait speeds.

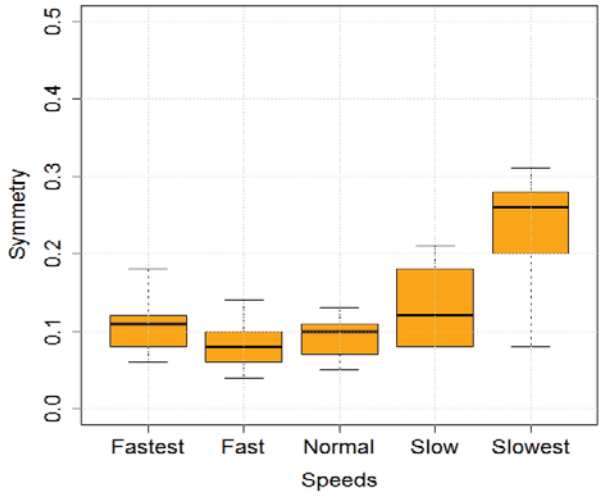


Figure 6. Symmetry result for five different gait speeds.

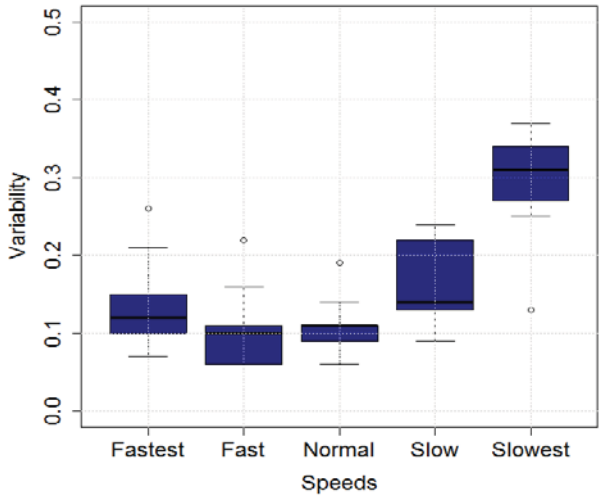


Figure 7. Variability results for five different gait speeds.

6 Conclusion

Gait analysis has attracted a great deal of interest in the domain of clinical and biomechanical research. Gait analysis using wearable devices provides continuous and repeatable results in daily activity with inexpensive cost and convenient portability [13]. However, using basic temporal and spatial gait parameters such as the number of steps, step length, and step duration are insufficient to get informative gait interpretation.

We propose a multidimensional view of mobility that uses a gait metric that includes notions of balance control and gait agility. We conducted an experimental study for various gait speeds using the proposed concept. The results illustrate that different speed gait is efficiently identified using our comprehensive mobility metric. It addresses the inherent spatial limitation of laboratory-based gait analysis. Moreover, the mobility metric can be extended and applied to continuously tracking human gait in daily living and work settings.

We expect that our wearable sensor-based mobility metric can be adopted for long-term mobility monitoring to monitor general gait patterns. As a next step of the study, the multidimensional mobility concept is being applied to a physical activity classification system in order to capture various types of gait in natural setting to assess daily and weekly gait pattern changes.

7 References

[1] S. Collins, A. Ruina, R. Tedrake and M. Wisse, "Efficient bipedal robots based on passive-dynamic walkers," Science, vol. 307, pp. 1082-1085, Feb 18. 2005.

[2] U.S. General, "Surgeon General's report on physical activity and health. From the Centers for Disease Control and Prevention," JAMA, vol. 276, pp. 522, 1996.

[3] G. Giakas and V. Baltzopoulos, "Time and frequency domain analysis of ground reaction forces during walking: an investigation of variability and symmetry," Gait Posture, vol. 5, pp. 189-197, 1997.

[4] J.M. Hausdorff, "Gait variability: methods, modeling and meaning," Journal of Neuroengineering and Rehabilitation, vol. 2, pp. 1, 2005.

[5] F. Lacquaniti, Y.P. Ivanenko and M. Zago, "Patterned control of human locomotion," J. Physiol. (Lond.), vol. 590, pp. 2189-2199, 2012.

[6] D.W. Marhefka and D.E. Orin, "Gait planning for energy efficiency in walking machines," in Robotics and Automation, 1997 IEEE International Conference on, pp. 474-480, 1997.

[7] C. Mummolo, L. Mangialardi and J.H. Kim, "Quantifying dynamic characteristics of human walking for comprehensive gait cycle," J.Biomech.Eng., vol. 135, pp. 091006, 2013.

[8] M. Plotnik, N. Giladi and J.M. Hausdorff, "Bilateral coordination of gait and Parkinson's disease: the effects of dual tasking," J. Neurol. Neurosurg.Psychiatry, vol. 80, pp. 347-350, Mar. 2009.

[9] R.O. Robinson, W. Herzog and B.M. Nigg, "Use of force platform variables to quantify the effects of chiropractic manipulation on gait symmetry," J. Manipulative Physiol. Ther., vol. 10, pp. 172-176, Aug. 1987.

[10] H. Sadeghi, "Local or global asymmetry in gait of people without impairments," Gait Posture, vol. 17, pp. 197-204, 2003.

[11] H. Sadeghi, P. Allard, F. Prince and H. Labelle, "Symmetry and limb dominance in able-bodied gait: a review," Gait Posture, vol. 12, pp. 34-45, 2000.

[12] N. Stergiou, R.T. Harbourne and J.T. Cavanaugh, "Optimal movement variability: a new theoretical perspective for neurologic physical therapy," Journal of Neurologic Physical Therapy, vol. 30, pp. 120-129, 2006.

[13] W. Tao, T. Liu, R. Zheng and H. Feng, "Gait analysis using wearable sensors," Sensors, vol. 12, pp. 2255-2283, 2012.

[14] M. Vukobratovic, B. Borovac, D. Surla and D. Stokic, Biped locomotion: dynamics, stability, control and application, Springer Science & Business Media, 2012.

[15] S.C. Webber, M.M. Porter and V.H. Menec, "Mobility in older adults: a comprehensive framework," Gerontologist, vol. 50, pp. 443-450, Aug. 2010.

[16] I. Youn, S. Choi, R. Le May, D. Bertelsen and J. Youn, "New gait metrics for biometric authentication using a 3-axis acceleration," in Consumer Communications and Networking Conference (CCNC), 2014 IEEE 11th, pp. 596-601, 2014.

[17] W. Zijlstra and K. Aminian, "Mobility assessment in older people: new possibilities and challenges," European Journal of Ageing, vol. 4, pp. 3-12, 2007.

Recognition of Smoking Gesture Using Smart Watch Technology

Casey A. Cole[1], Bethany Janos[2], Dien Anshari[3], James F. Thrasher[3], Scott Strayer[4], and Homayoun Valafar[1*]

[1]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA

[2]Department of Biomedical Engineering, University of South Carolina, Columbia, SC 29208, USA

[3]Department of Public Health, University of South Carolina, Columbia, SC 29208, USA

[4]University of South Carolina School of Medicine, Columbia, SC 29208, USA

[*] Corresponding Author Email: homayoun@cec.sc.edu Phone: 1 803 777 2404 Fax: 1 803 777 3767

Mailing Address: Swearingen Engineering Center, Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA

Abstract *– Diseases resulting from prolonged smoking are the most common preventable causes of death in the world today. In this report we investigate the success of utilizing accelerometer sensors in smart watches to identify smoking gestures. Early identification of smoking gestures can help to initiate the appropriate intervention method and prevent relapses in smoking. Our experiments indicate 85%-95% success rates in identification of smoking gesture among other similar gestures using Artificial Neural Networks (ANNs). Our investigations concluded that information obtained from the x-dimension of accelerometers is the best means of identifying the smoking gesture, while y and z dimensions are helpful in eliminating other gestures such as: eating, drinking, and scratch of nose. We utilized sensor data from the Apple Watch during the training of the ANN. Using sensor data from another participant collected on Pebble Steel, we obtained a smoking identification accuracy of greater than 90% when using an ANN trained on data previously collected from the Apple Watch. Finally, we have demonstrated the possibility of using smart watches to perform continuous monitoring of daily activities.*

Keywords: smoking cessation, smart watch, machine learning, neural networks, pattern recognition

1 Introduction

In the past decade, measures have been taken to warn the population about the dangers of smoking. While the smoking rate has decreased significantly since then, smoking remains the leading preventable cause of death throughout the world. Additionally, youth tobacco use has increased as the popularity of products such as e-cigarettes and hookah has risen[1] . In America, 53.4% of college students have smoked at least one cigarette and 38.1% reported smoking in the past year[2] . Even though the hazards of smoking are generally accepted, there remains many smokers who struggle to quit. Those who try to quit are typically middle aged and beginning to feel the adverse effects of smoking. Yet, on average, smokers relapse four times before successfully quitting[3] . Many smokers do not realize that it is normal to require multiple attempts to quit smoking and therefore need recurring intervention and support to aid them. Constant support from an individual's community is shown to increase the likelihood of quitting[2] . The existence of an application (housed on a smart phone or watch) that would provide this constant support could greatly increase a person's fortitude to abstain from smoking.

The first step in making such an application is the ability to detect when a person is smoking so that the appropriate intervention can be initiated. Previous works have shown the possibility of detecting smoking gestures using in-house designed wearable devices[4,5]. These techniques have shown great promise with both high accuracy (95.7-96.9%) and low false positive rates (<1.5%). However, they require the use of devices not commonly found in a typical household such as multiple 9-axis inertial measurement units (IMU's), respiration bands that must be worn across the chest and two-lead electrocardiograph worn under the clothes. The use of these uncommonly and relatively expensive devices severely limits mass deployment for daily use.

Smart watches are becoming increasingly prevalent in common households. According to Apple's website, over 5 million Apple Watches were sold in 2015 alone and projections into 2016 show promising growth. Other smart watch companies like Asus and Pebble have seen similar growth patterns from as well. By contrast to the previous methods, the method explored in this study relies solely on the use of a smart watch's built-in accelerometer, effectively eliminating the need to use more uncommon detection devices. In addition, the pairing of a smartwatch with a smart mobile device enables immediate alerting, engagement and recruitment of social support groups to prevent or alter one's smoking behavior. As the first step in

this process we have examined the feasibility and complexity of detection of smoking gesture using smart wearable devices. Our investigation has included minimum data requirement and an exploration of most informative dimension of accelerometer sensors. Prior knowledge of the problem complexity will allow for a smoother transition into actual deployment of our detection mechanism on smart watches in the future.

2 Background and Method

The overall view of our study consisted of three major stages: data collection, training of multiple artificial neural networks for pattern recognition, and evaluation. The following sections provide a more detailed outline for each of these stages.

2.1 Data Collection

Data in this study were acquired by a non-smoking participant utilizing an Apple Watch (version 2.1). Using the application PowerSense (available in App Store for iOS) a number of individual smoking gestures (also referred as a puff) and continuous smoking sessions (a session that consists of multiple puffs) were recorded and analyzed. All samples were measured at a 50Hz sampling rate. Due to minor fluctuations in the duration of each gesture, the number of data points varied for each gesture. Each isolated puff pattern was represented by 200 interpolated data point in order to create a uniform size input set for the pattern recognition stage. The resulting smoking gestures are shown below in Figure 1. The three differently colored line clusters represents each of the three dimensions of the accelerometer (X in blue, Y in red and Z in green). Each cluster is an overlay of all 20 individual smoking sessions used in the training of the neural networks. Based on visual inspection, it is clear that the smoking pattern is very well conserved across each of the samples.

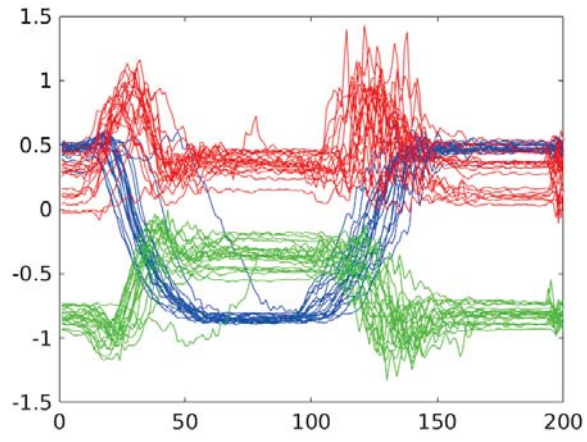


Figure 1. Overlay of all single smoking gestures with the X dimension in blue, Y in red and Z in green.

In addition to smoking sessions, several non-smoking gestures were also recorded. These gestures (seen in Figure 2) included drinking, scratching one's nose, yawning, coughing, brushing hair behind one's ear and rubbing one's stomach. The selection of these patterns were based on activities that may be similar to smoking gesture, or ones that may be present during most common smoking sessions. These gestures were including in both the training test and testing set.

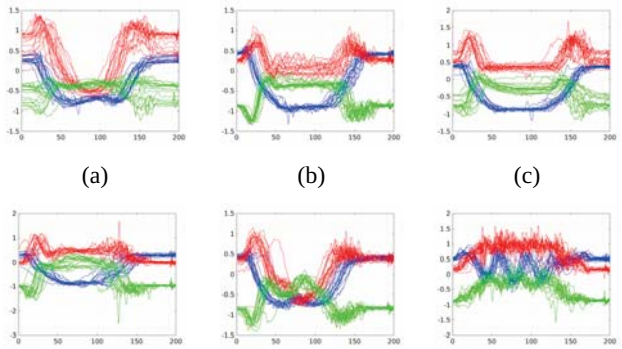


Figure 2. Overlay of all patterns collected for the following non-smoking gestures: (a) drinking, (b) scratching one's nose, (c) yawning, (d) coughing, (e) brushing hair behind one's ear, and (f) rubbing one's stomach.

In some of these gestures, such as scratching one's nose and yawning in Figure 2(b) and Figure 2(c) respectively, the movement of the hand and arm clearly resemble an individual smoking gesture (seen in Figure 1). Inclusion of these gestures into the data set will allow for studying how well the proposed method can distinguish smoking gestures from other very similar gestures.

In addition to individual gestures, longer continuous sessions were recorded. The continuous smoking sessions consisted of approximately 7 to 10 gestures per session. The non-smoking sessions included three common activities: eating, drinking and putting on chapstick/lipstick. Each session was divided into 200 time step segments using a rolling window approach. As seen in Figure 3, a continuous smoking session is no more than a combination of individual smoking gestures seen in Figure 1.

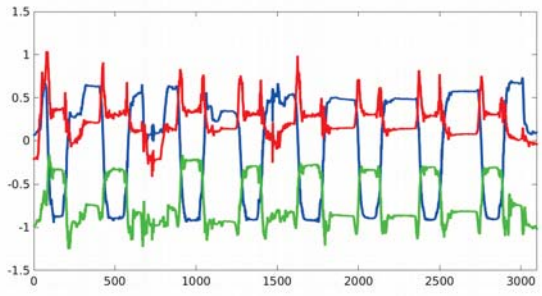


Figure 3. Example of continuous smoking session.

Table 1 summarizes the total number of smoking and non-smoking gestures in each of the data sets and breaks down the exact amount used in each phase of the investigation.

Table 1. Summary of data utilized in analyses.

	Non-smoking	Smoking
Training Set	120	20
Testing Set	30	10
Extended Sessions	5	5

In order to test the applicability of the presented method across other wearable platforms, as well as other participants, one smoking session was acquired by a different participant (than the original set of data used for training) on the Pebble Time Steel used on Android platform. The Pebble Time Steel was selected as a second test watch due to its long lasting battery life, durability, and reasonable price. A continuous smoking session was recorded using the AccelTool (http://mgabor.hu/accel/) App at a sampling rate of 50 Hz.

2.2 Pattern Recognition Via Artificial Neural Networks

The neural network toolbox in Matlab (version R2016a) was utilized during this phase of our study. Levenberg-Marquardt backpropagation[6–8] was selected as the training algorithm in all sessions. For each training session, the data were randomly partitioned into 3 sets: 70% in the training set, 15% in the validation set and 15% in the testing set. The networks were then trained, validated and then rigorously tested for accuracy. The procedures for the training and validation/testing phases are outlined below.

Training – The interpolated raw data consisted of information from three dimensions (X, Y and Z). In order to fully identify and evaluate useful information in the data, all three dimensions were utilized both individually and in combination with each other. In total, five ANNs were created for use in this study—one for each of the three dimensions (X, Y and Z), one for the combination of all three dimensions (referred to as XYZ) and one for the average of the three dimensional data (referred to as AVG). The number of inputs for the X, Y, Z and AVG data sets was 200, while the input size for the XYZ data set was 600. In each of the neural networks the hidden layer consisted of 10 hidden neurons. A single output neuron was used, with zero denoting a non-smoking gesture and one signifying a smoking gesture.

Validation/Testing – Validation of the appropriate level of training was accomplished using 15% of the excluded training dataset. The networks were further subjected to testing using several different data sets. The first of which was the remaining 15% of the data excluded from the original training set. Next, a new set of individual gestures (not included in the original training set) was presented to the networks. To test the method on more realistic cases, continuous smoking and non-smoking sessions were also presented to the networks. Lastly, a continuous smoking session from a different smart watch was tested on each of the ANNs. The results of each test set are reported in Section 3.

2.3 Evaluation

To measure the success of the proposed method specificity, sensitivity and total accuracy of each trial were observed. Specificity describes the rate at which the method is able to correctly classify a non-smoking event. Specificity is calculated by use of Eq (1), where *TN* and *FP* denote the number of true negatives and false positives, respectively.

$$Specificity = \frac{TN}{TN + FP} \quad (1)$$

Sensitivity refers to the rate at which the method correctly identifies a smoking event and can be calculated using Eq (2), where *TP* represents the number of true positives and *FN* denotes the number of false negatives.

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

Total Accuracy is then a measure of how often the method correctly classifies both smoking and non-smoking gestures and is calculated by Eq (3).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

In this work a cutoff threshold of 0.8 (or 80%) was used above which was considered a successful detection.

3 Results and Discussion

In the following sections the results of testing the neural networks are reported. In each section the data set that performed the best and worst are discussed.

3.1 Accuracy on Training Set

For each of the data sets the corresponding neural networks was independently trained 10 times and the network with the highest accuracy was chosen to be used for future test sets. The accuracies, specificities and sensitivities described in Figure 4 represent the performance of the final trained neural networks on their respective training sets.

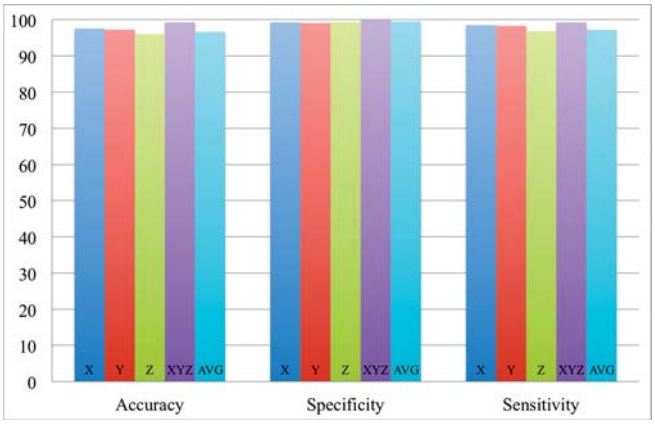


Figure 4. Accuracy, specificity and selectivity of the neural

networks during training. The bars are individually labeled based on their respective training sets.

All of the neural networks exhibited high accuracy and specificity (> 90%) with respect to the training set of data. In each case the network trained with the XYZ data performed the best and the network trained with just the Z data performed the worst. A visual comparison of the individual smoking gestures and the non-smoking gestures clearly explains Z's poorer performance in correctly classifying smoking gestures. The Z dimension (in green Figures 1 and 2) exhibits very similar patterns across both smoking and non-smoking gestures. However, it is worth noting that the Z dimension still obtained a high specificity (nearly 100%) which means that it still may carry some complementary information, especially when identifying a non-smoking gesture.

3.2 Individual Gesture Detection

In this experiment, a new set of individual gestures (smoking and non-smoking) were presented to the previously trained neural networks. The accuracy, specificity and sensitivity are reported in Figure 5. Accuracies were measured by forward propagating each of the new samples in the corresponding neural network and then recording the number of correct and incorrect predictions. A threshold of 0.5 was used in interpretation of the neural network output, that is, any output larger than 0.5 was considered as smoking and any output lower than 0.5 was considered as non-smoking.

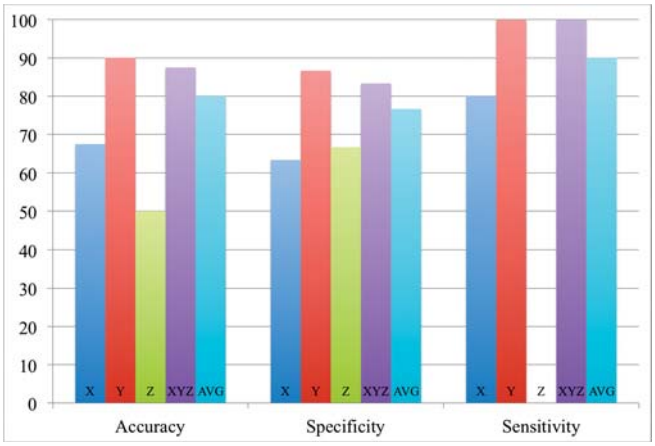


Figure 5. Accuracies, specificity and sensitivity in the individual gesture detection trials.

As shown in Figure 5, the Y dimension performed the best with not only the highest accuracy, but also the highest specificity and a 100% sensitivity. The XYZ and AVG data sets also performed well especially in their ability to identify smoking gestures. Consistent with the previous section, the Z dimension performed the worst with both a low accuracy and specificity as well as a 0% sensitivity rate indicating utilization of the Z dimension results in identification of 0/10 smoking gestures.

To better understand the nature of false-positive classifications, contribution of each individual gesture was recorded and results are shown in Figure 6. Based on the results shown in this figure, the non-smoking pattern that caused the most false positives was coughing followed by scratching of nose and yawning. These results were expected due to the high degree of visual similarity of these gestures to an individual smoking gesture.

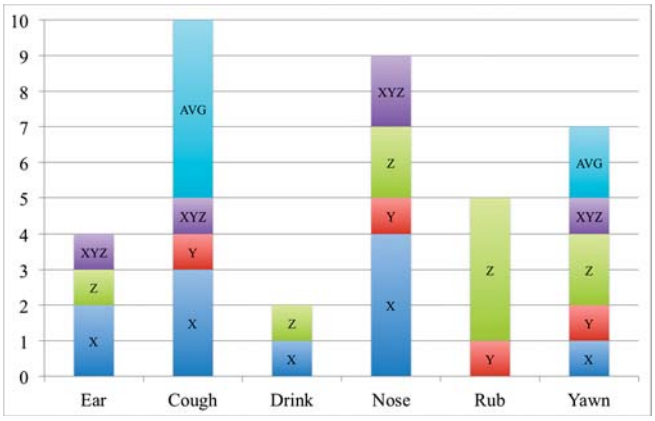


Figure 6. Total number of false positives created by each non-smoking gesture. Each segment is labeled based on the dimension of the accelerometer data being used.

3.3 Continuous Gesture Detection

In this section the results for detection on continuous monitoring of gestures are reported. To accomplish this objective the neural networks trained on static gestures were utilized. Using a running window of size 200 (without any interpolation) the continuous gestures were parsed into data sets for input into the networks. The classification result for each running window (0 denoting non-smoking and 1 denoting smoking) is plotted at the beginning of each running window. Figure 7 illustrates an example output (in purple before converting to a binary representation) of the neural network trained on X. Visual inspection of this figure clearly confirms the correlation between spikes in detection pattern over the regions where an apparent smoking gesture. However, there is not a trivial way to quantify the network's success because it is not immediately clear where should constitute the start and end of a gesture. Therefore, in order to be sure to encompass the entire gesture, generous ranges were handpicked to describe each smoking gesture. As in the previous section, a cutoff of 0.5 was chosen where any output greater than 0.5 was considered smoking and anything below was counted as non-smoking. Specificity was measured by considering all other sections of the continuous gesture not within the smoking ranges.

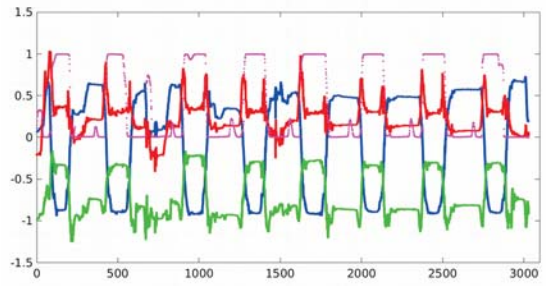


Figure 7. Example of continuous smoking session superimposed with the output of the neural network trained on the X dimension.

The averaged results across all five continuous sessions are presented in Figure 8 along with error bars representing the minimum and maximum of each averaged result.

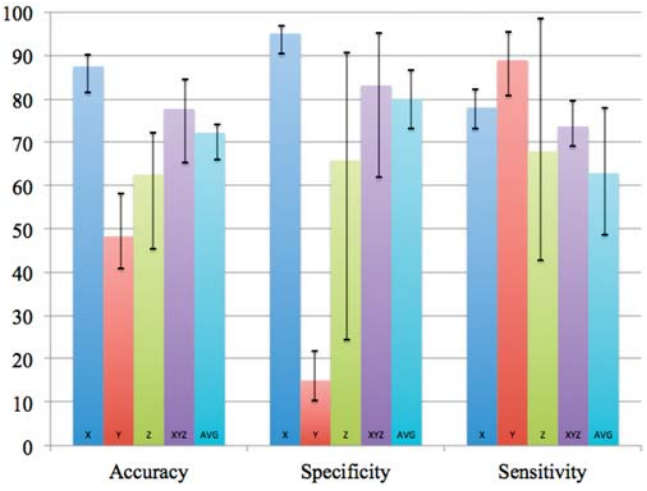


Figure 8. Averaged accuracies, specificities and sensitivities across the five continuous smoking gesture detection experiments with error bars representing the respective min and max of each value.

In the continuous smoking sessions, the X dimension performed better than all other dimensions. The Y dimension had a better sensitivity score but this can be disregarded due to its low specificity score. A high sensitivity coupled with a low specificity score denotes that the Y dimension classifies practically everything as smoking and therefore it's high sensitivity should be ignored. In this sense, Y performed the worst overall.

In the non-smoking sessions selectivity becomes inapplicable and specificity is equivalent to total accuracy. therefore, only accuracies these sessions are presented in Figure 9.

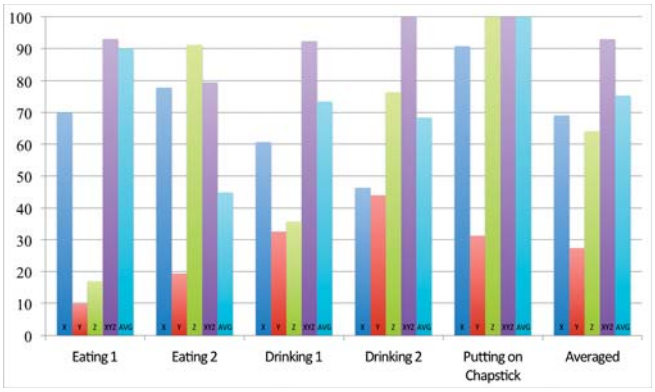


Figure 9. Accuracies for five continuous non-smoking session trials.

Despite it's superior performance in classification with continuous smoking gestures, the X dimension performed with an average accuracy of under 70% in the non-smoking continuous sessions. It seems that in the presence of more complex continuous gestures (like eating and drinking) the XYZ data set seems to contain the most useful information for correct classification. Eating sessions seemed to pose the most difficulty for XYZ. This could signify that eating is one of the closest gestures to smoking and can therefore lead to confusion in the network. In accordance with previous results, the Y dimension performed the worst across all the non-smoking sessions.

3.4 Exploration of dependency on the wearable device

As described in the Section 2.1, a continuous smoking gesture was recorded using a Pebble smart watch. Results were collected using the pre-existing neural networks that was trained on data from the Apple Watch from a different participant. Figure 10 shows the smoking session recorded using the Pebble watch, which exhibit significant similarity to patterns shown in Figure 1. The outputs of the neural network using the X data set are shown in purple. Again, there is good correlation between the smoking gestures and the spikes in the output of the neural network. Figure 11 shows the resulting accuracies, specificities and sensitivities for this smoking session.

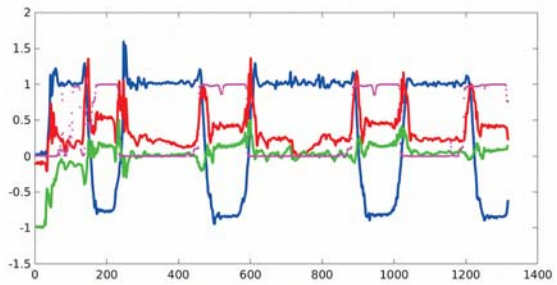


Figure 10. Output of the neural network for the X dimension superimposed to the original smoking session.

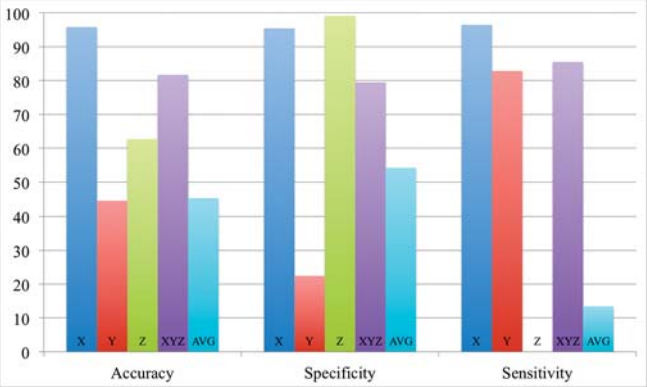


Figure 11. Results for the Pebble smart watch.

As in the previous cases of continuous smoking sessions, the X dimension performed the best in classifying the gestures. In the case of the Pebble watch, the Y and Z dimensions did equally poorly. The Y dimension classified almost everything as smoking and the Z dimension classified nearly everything as non-smoking.

4 Conclusion

The general summary of our work supports the feasibility in detection of smoking gestures using typical sensors available in smart watches. Based on our experiments, pattern recognition via Artificial Neural Networks applied to the sensor data obtained from smart watches can produce performances very comparable to previously reported work. However, the use of a smart watch is far more pragmatic in general population studies over the other existing technologies. Results shown in section 3.4 suggest the possibility of delivering an application capable of detecting smoking gesture across the general population of smokers. This universal Artificial Neural Network eliminates the need to customize a training session per user.

Our exploration of efficacy of individual sensor data in detection of gesture has produced unexpected results. The neural network trained with data from just the X dimension performed the best in the presence of continuous smoking gestures but when faced with more complex non-smoking motions, it fails and a more complete set of data is needed to distinguish smoking gestures. Across all the testing sets the neural network trained with data from all three dimensions (XYZ) did consistently well. However, the XYZ data set requires 600 inputs to the network whereas the X data set only requires 200. This is a significant reduction in data requirement which directly impacts the computational time of the method. For this reason, the viability of both data sets will continue to be investigated.

Additional investigations are required before general deployment of such approaches. Continuous monitoring of data may be outside of power limitations of such devices and may act as a technological barrier. Our future investigations will include optimization of sampling rate, minimization of bluetooth communication between the smart watch and the companion phone, and better assessment of the universality of the trained ANN.

5 Acknowledgments

Funding for this work is provided by ASPIRE-II grant from the University of South Carolina Research Foundation.

6 Bibliography

1. Li, K. *et al.* Smoking and Risk of All-cause Deaths in Younger and Older Adults: A Population-based Prospective Cohort Study Among Beijing Adults in China. *Medicine (Baltimore).* **95,** e2438 (2016).
2. Rooney, B. L., Silha, P., Gloyd, J. & Kreutz, R. Quit and Win smoking cessation contest for Wisconsin college students. *WMJ* **104,** 45–9 (2005).
3. Khati, I. *et al.* What distinguishes successful from unsuccessful tobacco smoking cessation? Data from a study of young adults (TEMPO). *Prev. Med. reports* **2,** 679–85 (2015).
4. Saleheen, N. *et al*. puffMarker. in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15* 999–1010 (ACM Press, 2015).
5. Parate, A., Chiu, M.-C., Chadowitz, C., Ganesan, D. & Kalogerakis, E. RisQ: Recognizing Smoking Gestures with Inertial Sensors on a Wristband. *MobiSys ... ... Int. Conf. Mob. Syst. Appl. Serv. Int. Conf. Mob. Syst. Appl. Serv.* **2014,** 149–161 (2014).
6. Levenberg, K. A method for the solution of certain problems in least squares. *Q. Appl. Math.* **2,** 164 – 168 (1944).
7. Marquardt, D. W. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *J. Soc. Ind. Appl. Math.* **11,** 431–441 (1963).
8. Ranganathan, A. The Levenberg-Marquardt Algorithm. (2004).

64 Channel Digital Signal Processing Algorithm Development for 6dB SNR improved Hearing Aids

S. Jarng[1], M. Samuel[2], Y. Kwon[2], J. Seo[3], and D. Jarng[4]
[1]Department of Electronics Eng., Chosun University, Gwangju, The Republic of Korea
[2]Heaing Impairment and Hearing Rehabilitation Research Institute affiliated with AlgorKorea Co. Ltd, Seoul, The Republic of Korea
[3]Dept. Otolaryngology-HNS, The Catholic University of Korea, Seoul, Korea
[4]Dept. of Computer Science, Biola University, Los Angeles, CA, USA

Abstract - *No two people with hearing loss will have a similar audiogram. Hearing loss affects people at varying frequency ranges. It is therefore imperative to design a hearing aid that can amplify/modify sounds with a high frequency resolution. Our 64 channel hearing aid has a 125 Hz resolution which ensures that we provide a very precise control over the 8000 Hz frequency spectrum with non-linear compression. A probabilistic noise reduction algorithm combined with feedback cancellation and directionality enhances the speech clarity while minimizing undesirable sound sources.*

Keywords: 125 Hz resolution; non-linear compression; 64 channel spectrum; Noise reduction; feedback cancellation

1 Introduction

Most hearing aids that are currently available in the market have a frequency resolution larger than 500 Hz. When you observe an audiogram of a person with hearing loss, in certain cases, there are varying levels of hearing within a 500 Hz range across an 8 kHz spectrum. Applying a generic 'fix' across a 500 Hz would be sufficient enough to make sounds audible again, but the clarity of speech would remain elusive. Therefore a higher resolution becomes a necessity. To attain a high quality of speech, an 8000 Hz frequency spectrum is divided into 64 channels with each channel demarcated by a 125 Hz step size. To further improve speech quality, environmental noises and inherent feedback that may exist in a hearing aid needs to be nullified. A probabilistic noise reduction algorithm was designed to suppress ambient noises while preserving speech. A feedback algorithm based on the energy level of the input sounds was used to suppress feedback. A dual microphone design was used in the hearing aid to have a directional hearing aid which can supplement the existing speech clarity.

1.1 Hearing loss

Hearing losses can be classified into three main categories [1],

1) Conductive Hearing loss
2) Sensorineural Hearing loss
3) Mixed Hearing loss

Each type of hearing loss has inherent unique qualities that differentiate one from another. Standard audiometric tests do not provide very precise measurements of a person's hearing loss irrespective of its type.

We therefore, designed our own In-Situ test method where a customer can be tested for all 64 channels with a 125 Hz precision. Based on the intensity of the sound levels, three levels of measurement in the In-Situ design are assigned for each channel namely Loud, Normal and Soft. These three levels of measurement provide an explicit audiometric measurement of a customer's hearing level. We use this measurement to design a fitting curve that would be suitable for a customer.

2 64 channel non-linear compression

Dynamic range of hearing refers to the range between the maximum and minimum audible level of hearing. It is common knowledge that with Sensorineural hearing loss, the dynamic range of hearing is reduced. Softer sounds are barely audible, while loud sounds remain loud [2]. This can lead to situations where conversational speech is inaudible or disconcerting. To solve the issue, merely amplifying sounds mean speech and softer sounds are audible, but loud sounds are excruciatingly louder. Therefore sounds have to be compressed to maintain a balance between loud and soft sounds so that they fall within the dynamic range of a person with hearing loss.

There are two distinct types of compression

1) Linear compression
2) Non-linear compression

Linear compression refers to an amplification type where there is consistent amplification for all levels of sounds as shown in Figure 1.

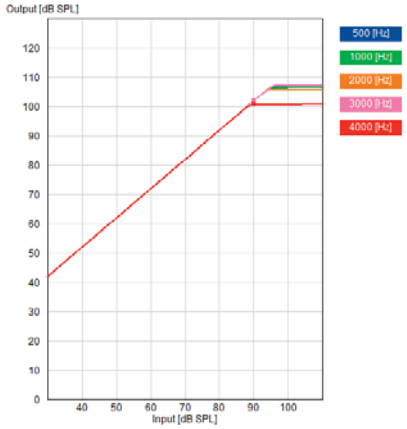


Figure 1: Linear compression

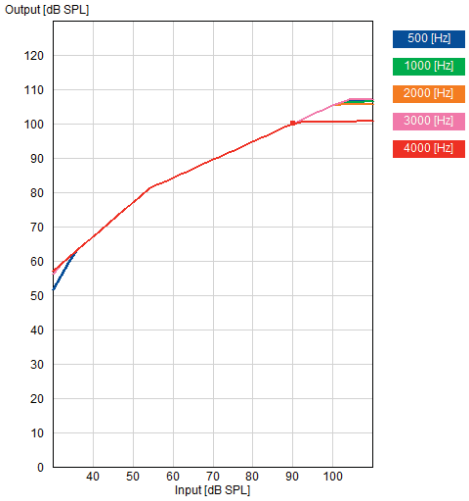


Figure 2: Non-Linear compression

Within Linear compression, the maximum output is limited to a certain value and all output sounds are less than or equal to that value albeit with a uniform gain.

However, in non-linear compression, varying amplification is applied to different levels of sound; for example, weak sounds get a higher amplification, while louder sounds get a lower amplification to fit all sounds within the dynamic range of the user, as shown in Figure 2.

There are four segments within the compression region, namely, Squelch (s1), Linear (s2), Non-linear (s3) and Saturation (s4).The Squelch Threshold (SQTH) point refers to the end of the Squelch region (s1) where weak input sounds are suppressed to reduce noise levels. Segment 2 (s2), also known as the linear region is bordered by SQTH and the Lower Threshold point (LTH). Within the linear region, a uniform gain is applied to the input signal. The Non-linear region, Segment 3 (s3), is where varying levels of amplification are applied to the input signal. The upper threshold (UTH) point signals the point of maximum amplification that can be applied to an input signal. A low level gain (LLG) controls the gain threshold within the saturation region. Beyond UTH, lies the saturation region where the output is limited to a saturation level. By varying the High Level gain (HLG), the saturation level can be set to any desired level. Based on the threshold points, the compression parameters are calculated for the entire frequency spectrum.

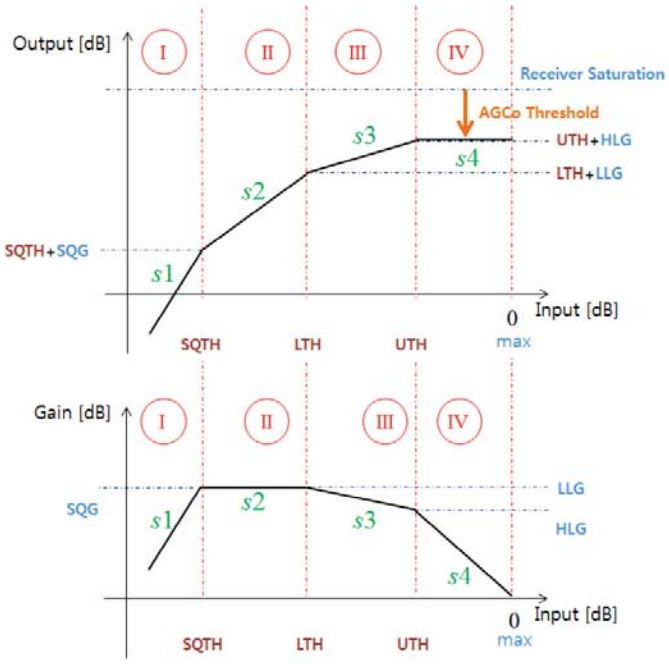


Figure 3: Compression region

2.1 Calibration

The key aspect in the above compression scheme is the frequency resolution that provides our design an edge over others. We split the 8000 Hz frequency range into 64 distinct regions/channels, each channel with a frequency resolution of 125 Hz. The precision begins even before compression is implemented. We start by obtaining an accurate measurement of a customer's hearing ability. This is done by performing an In-Situ test.

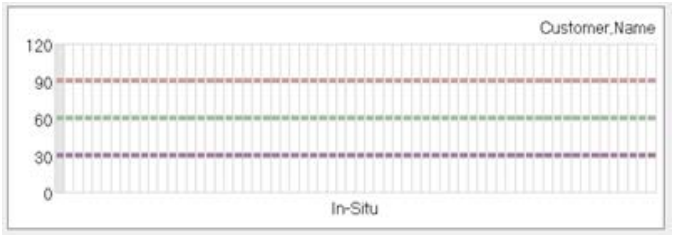


Figure 4: In-Situ Test

There are three levels of measurement, Loud, Comfortable and Soft within an In-Situ test as shown in Figure 4 by the three distinct dotted lines with each dot representing a single

channel. A tone played at these three levels of measurements helps determine a customer's hearing level across the entire 64 channels. Coupled with the customer's hearing ability, both the microphone and receiver used in the hearing aid are calibrated to the same 125 Hz precision which means we get an accurate representation of the capabilities of the microphone and receiver as well. The front and back microphone are calibrated individually. During a receiver calibration, both a sensitive and a saturated receiver calibration are done.

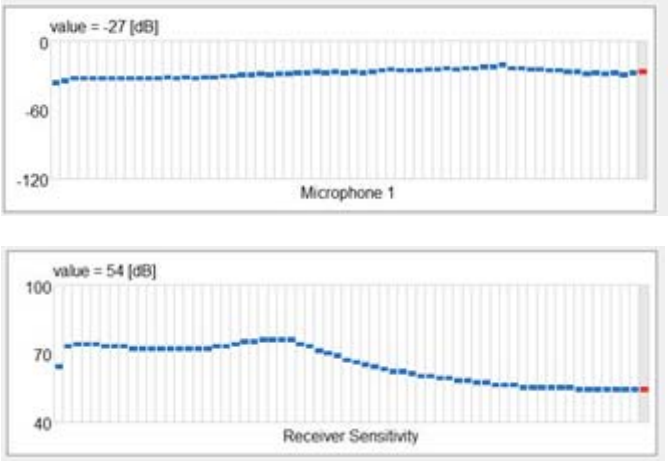


Figure 5: Microphone & Receiver calibration

With the microphone and receiver calibrated data and an accurate audiogram, we calculate the compression parameters required to design an I/O curve as shown in the figure 6 based on the formula for a slope,

$$\text{Gain} = X * S + B \quad (1)$$

Where,

$$X = x_n + TM - 29.8973,$$
$$S = 2.0 * s_n * C$$
$$B = b * 2.0 + 240$$

C = calibration constant which is a combination of Microphone and Receiver sensitivity values.

$$b = (y - s_n * X) * C$$

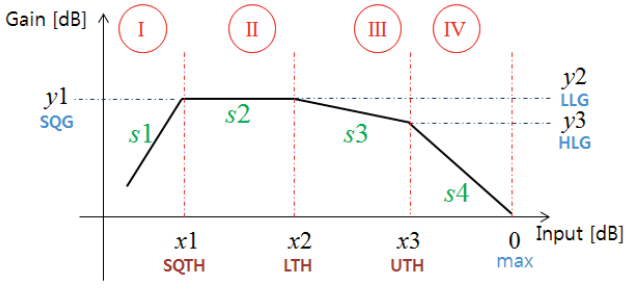


Figure 6: Compression parameters

Due to the 125 Hz precision, the gain calculated for each of the 64 channels ensures that the gain applied is smooth over the entire frequency spectrum and best fits a hearing aid user's requirement. This peerless fitting for each customer provides the best possible clarity a customer can hope for.

3 Noise Reduction

In spite of the precise non-linear compression applied, environmental noises in the signal can be such a nuisance for someone with hearing loss. People with normal hearing can filter out environmental disturbances whereas people with hearing loss are not capable of doing the same. Therefore a Noise Reduction algorithm plays the part of a filter to remove noises present in the signal [3]. In this algorithm, we determine the minimum power in a predefined window of samples and use that information to calculate a smoothing parameter which helps us get the noise estimate. The estimate is used to calculate the SNR (Signal to Noise Ratio), which determines the presence or absence of speech, which in turn determines the gain required to suppress noise and amplify speech. The output has minimal distortions and is clear enough for practical purposes.

3.1 Background

Consider a noisy speech signal, x(n).

$$x(n) = s(n) + d(n) \quad (2)$$

Where, s(n) and d(n) denote discrete time signals of clean speech and noise respectively. The noisy speech signal, x(n) is split into overlapping frames using a window function. A Discrete Short Fourier Transform (DSTFT) written as,

$$X(k,l) = \sum_{n=0}^{N-1} x(n + lM)h(n)e^{-j(\frac{2\pi}{N})kn} \quad (3)$$

is used to analyze the overlapping frames. Where,

k = frequency bin index,
l = time frame index
h = Analysis window of size N,
M = framing step (Number of samples separating two Successive frames)

Using DSTFT analysis, (1) can be represented as,

$$X(k,l) = S(k,l) + D(k,l) \quad (4)$$

3.2 Noise reduction algorithm

It is assumed that a noisy signal is a combination of a desired signal mixed with undesired sounds and signals. We therefore have to segregate the desired sounds from the noise within the noisy speech signal. It is also assumed that the intensity of the noisy signal has the behavior of Gaussian distribution in each frequency band, and that the noise is not correlated with the voice signal.

We now apply a short time (ST) Fast Fourier Transform (FFT) and inverse FFT as a fundamental analysis tool in order

to analyze the spectral characteristics of the noisy signal. We aim to estimate the optimal spectral gain, $G(\mathrm{k}, l)$ so that

$$\hat{X}(k,l) = G(k,l)Y(k,l) \tag{5}$$

Where k = frequency bin index
l = time frame index
Y = ST FFT of the noisy signal
X = optimal spectral amplitude of the voice signal

If X is the original spectral amplitude of the voice signal, we aim to derive G by means of minimizing $\mathrm{E}\{(|\hat{X}(k,l)| - |X(k,l)|)\}$. From the Gaussian distribution hypotheses, the conditional Probability Density Function (PDF) of the observed signal are given by,

$$pdf\left(Y\left(\mathrm{k}, \ell\right) \mid voice\ absence\right) = \frac{1}{\pi\sigma(k, \ell)^2} exp^{-\frac{\left|Y\left(\mathrm{k}, \ell\right)\right|^2}{\sigma(\mathrm{k}, \ell)^2}} \tag{6}$$

$$pdf\left(Y\left(\mathrm{k}, \ell\right) \mid voice\ presence\right) = \frac{1}{\pi\left(\sigma_d(k, \ell)^2 + \sigma_x(k, \ell)^2\right)} exp^{-\frac{\left|Y\left(\mathrm{k}, \ell\right)\right|^2}{\left(\sigma_d(\mathrm{k}, \ell)^2 + \sigma_x(\mathrm{k}, \ell)^2\right)}} \tag{7}$$

Where $\sigma_d(k,l)^2$ and $\sigma_x(k,l)^2$ are the variances of the signal without voice and of the signal with voices respectively.

The noise reduction algorithm is briefly described below. Initially, the minimum of the local energy, $S_{min}(k,l)$ is tracked as follows,

$$S_{min}\left(k, \ell\right) = min\left\{S_{min}\left(k, \ell - 1\right), \left|Y\left(k, \ell\right)\right|^2\right\} \tag{8}$$

$$S_{tmp}\left(k, \ell\right) = min\left\{S_{tmp}\left(k, \ell - 1\right), \left|Y\left(k, \ell\right)\right|^2\right\} \tag{9}$$

If $\mathrm{mod}(k,l) = 0, S_{tmp}(k,l)$ is re-initialized.

$$S_{min}\left(k, \ell\right) = min\left\{S_{tmp}\left(k, \ell\right), \left|Y\left(k, \ell\right)\right|^2\right\} \tag{10}$$

$$S_{tmp}\left(k, \ell\right) = \left|Y\left(k, \ell\right)\right|^2 \tag{11}$$

Figure 7 shows an example of the noisy speech spectrum and its minimum of the local energy at any time frame. The noisy speech changes all the time, while the minimum of the local energy varies slowly. The noise spectrum is estimated based on the following procedure:

The ratio between $|Y(k,l)|^2$ and its derived minimum of the local energy is denoted as,

$$S_r\left(k, \ell\right) \triangleq \left|Y\left(k, \ell\right)\right|^2 / \left|S_{min}\left(k, \ell\right)\right| \tag{12}$$

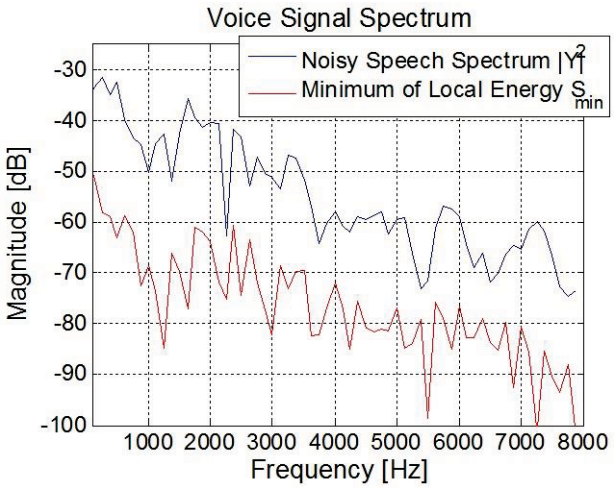


Figure 7: An example of the noisy speech spectrum and the minimum local energy

Then, the conditional speech presence probability is estimated as

$$\hat{p}'\left(k, \ell\right) = \alpha_p \hat{p}'\left(k, \ell - 1\right) + (1 - \alpha_p) I\left(k, \ell\right) \tag{13}$$

Where $\mathrm{I}(k,l) = 1$, $S_r(k,l) > \delta$ and $\mathrm{I}(k,l) = 0$. $\propto_p$ is 0.2 and δ is 5.
The noise spectrum $\hat{\lambda}_d(k,l)$ is recursively estimated while a smoothing parameter $\hat{\alpha}_d(k,l)$ is updated with the conditional speech presence probability, $\hat{p}'(k,l)$.

$$\tilde{\alpha}_d\left(k, \ell\right) \triangleq \alpha_d + (1 - \alpha_d)\hat{p}'\left(k, \ell\right) \tag{14}$$

$$\hat{\lambda}_d\left(k, \ell\right) = \tilde{\alpha}_d\left(k, \ell\right)\hat{\lambda}_d\left(k, \ell - 1\right) + [1 - \tilde{\alpha}_d\left(k, \ell\right)]\left|Y\left(k, \ell\right)\right|^2 \tag{15}$$

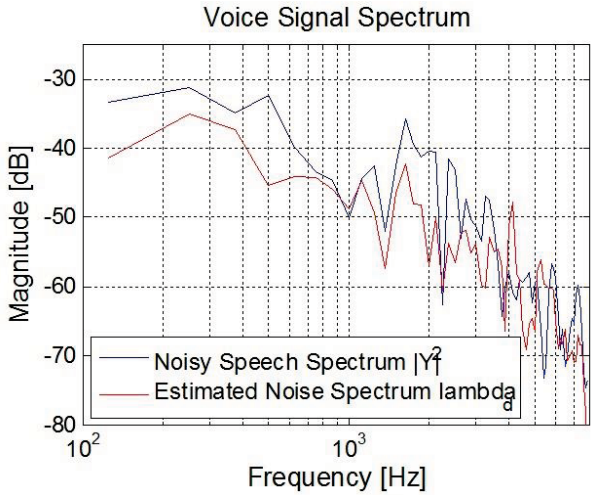


Figure 8: An example of the noisy speech spectrum and the estimated noise spectrum

Figure 8 shows an example of the noisy speech spectrum and the estimated noise spectrum. If the noisy speech spectrum is

higher than the estimated noise spectrum at a particular frequency band, we assume that the noisy speech spectrum should be amplified at that particular frequency band. However if the noisy speech spectrum is lower than the estimated noise spectrum, the noise spectrum should be suppressed at that particular frequency band. The estimated spectral gain, $G(k,l)$ is calculated as follows,

$$\gamma\left(k,\ell\right) \triangleq \min\left[\frac{\left|Y\left(k,\ell\right)\right|^2}{\hat{\lambda}_d\left(k,\ell\right)}, 1000\right] \quad (16)$$

$$\hat{\xi}\left(k,\ell\right) = \alpha\chi\left(k,\ell-1\right) + (1-\alpha)max[\gamma\left(k,\ell\right)-1,0] \quad (17)$$

$$\hat{\xi}'\left(k,\ell\right) = \hat{\xi}\left(k,\ell\right)/\left(1+\hat{\xi}\left(k,\ell\right)\right) \quad (18)$$

$$v\left(k,\ell\right) \triangleq \gamma\left(k,\ell\right)\hat{\xi}'\left(k,\ell\right) \quad (19)$$

$$G\left(k,\ell\right) = \hat{\xi}'\left(k,\ell\right)\exp\left(\frac{1}{2}\int_{v\left(k,\ell\right)}^{\infty}\frac{e^{-t}}{t}dt\right) \quad (20)$$

$$\chi\left(k,\ell\right) = G^2\left(k,\ell\right)\gamma\left(k,\ell\right) \quad (21)$$

Where α is 0.9899.

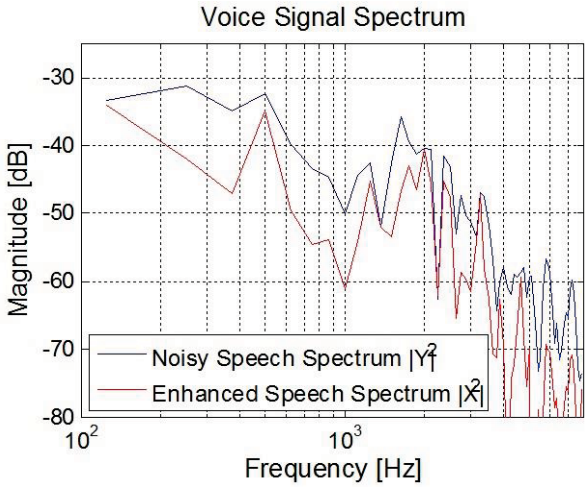


Figure 9: An example of noisy speech spectrum and the enhanced speech spectrum

Figure 9 shows an example of the noisy speech spectrum and the enhanced speech spectrum $\hat{X}(k,l)$ after multiplying $Y(k,l)$ with $G(k,l)$, that is, $\hat{X}(k,l) = G(k,l)\,Y(k,l)$.

3.2.1 Noise reduction results

The noise reduction algorithm was implemented in MATLAB and tested. The results were observed on an oscilloscope. It was later translated to assembler code and implemented on an Ezairo 7110. A vacuum cleaner in the background served as the noise source and a few random words were spoken. The input speech signal coupled with noise passed through the hearing aid microphone before it was processed by the Ezairo 7110 chip and sent out through the hearing aid receiver. A standard 2cc acoustic coupler transmitted the receiver output sound to a standard measuring amplifier and the output displayed on a digital oscilloscope.

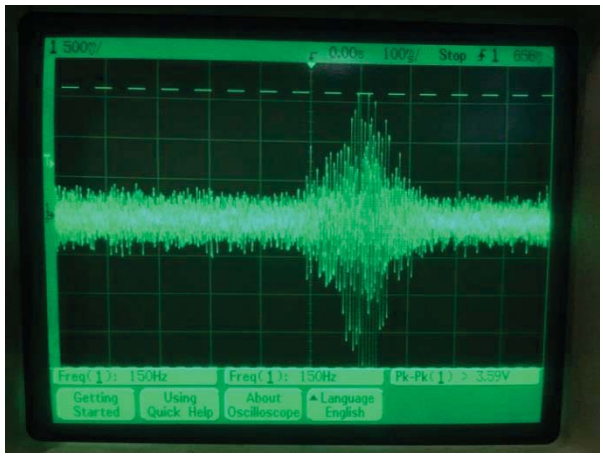


Figure 10: A word "One" was spoken with a noise source (vacuum cleaner running) in the background. At this instance, the Noise reduction algorithm was switched OFF.

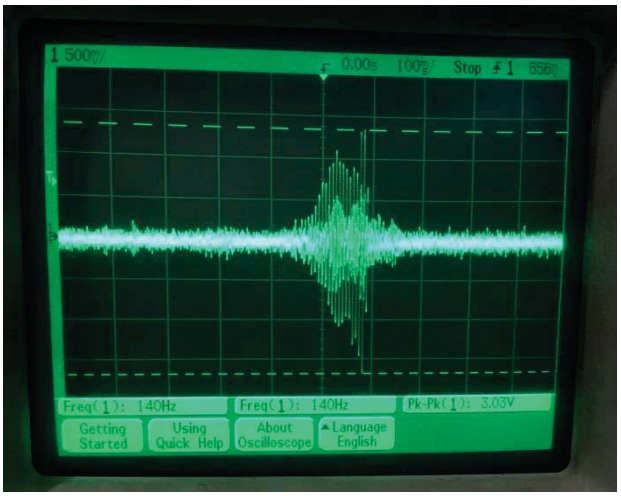


Figure 11: A word "One" was spoken with a noise source in the background. The Noise Reduction algorithm was turned ON.

Observing Figure 10 and Figure 11 affirms the effect of the Noise Reduction algorithm in a noisy environment. We were able to obtain a 6dB voice Signal-to-Noise Ratio (SNR) improvement. It should be noted that the speech signals itself remains more or less around the same amplitude while the noise levels were suppressed.

4 Feedback cancellation

Hearing aid users often experience a high pitched tone emanating from their hearing aids at different scenarios. Output sounds from the hearing aid receiver that loop back into the hearing aid microphone produce this extremely

disturbing high pitched tone referred to as a feedback tone. For example, if the hearing aid itself is a not a custom design mold, but a generic design, then subtle facial movements can cause the hearing aid to unsettle from its ideal position causing feedback to occur. If the fitting for the hearing aid is poorly done, then feedback can occur as well. Therefore a feedback cancellation algorithm is implemented to avoid this untoward disturbance.

4.1 Feedback cancellation design

The Energy (squared magnitude) of the complex (frequency domain) data of the input signal is used to determine if there is feedback present in the signal or not.

$$E = real(x^2) + imag(x^2) \quad (22)$$

The minimum of the local energy indicates the presence of feedback present within the system. We did a series of tests for feedback present in a hearing aid and measured the energy present within the signal. The results of the test are shown below.

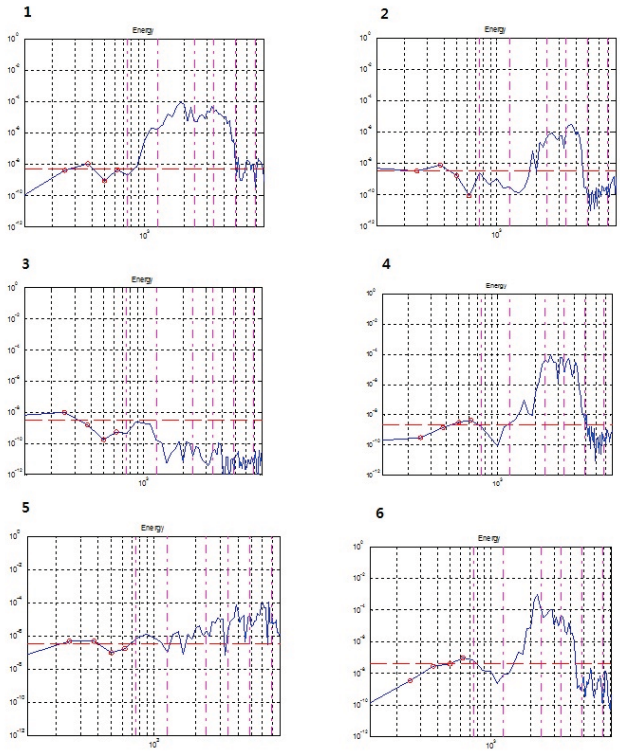


Figure 12: 1-Strong Feedback; 2-Weak Feedback; 3-No feedback; 4-Feedback in Ear; 5-Voice only; 6-Voice with feedback

To suppress the feedback, we split the frequency spectrum into five distinct regions and check for the energy level within each region. The energy present within each indicates the presence of feedback. When feedback is detected within a particular region, we suppress the feedback by minimizing the local energy within that specific region. This causes the feedback to disappear. The algorithm implemented in the Ezairo 7110 successfully achieves the purpose of the reducing feedback with gains up to 30 dB present. In the near future, a separate paper focusing on Feedback Cancellation will be written to emphasize the capabilities of our Feedback cancellation algorithm.

5 Conclusions

A 64 channel hearing aid with non-linear compression provides an unrivalled clarity among all the different hearing aids available due to its 125 Hz precision. The clarity arises as a combination of the compression with the background design setup which also functions on a 125 Hz precision. To aid the hearing aid user further, a probabilistic noise reduction algorithm and a feedback cancellation algorithm are implemented to suppress noise and remove feedback respectively.

Acknowledgment

This study was supported by research fund from the ministry of commerce, industry and energy (MOCIE Korea) 2015 core medical device commercialization technology development project (smartphone controlled 64 channel digital hearing aid with 6dB voice SNR (Signal to Noise Ratio) improvement: project number 10054678) in 2015.

6 References

[1] Hearing Instrument Science and Fitting practies. Second Edition. International hearing Society. "LaTeX: A Document Preparation System". Addison-Wesley Publishing Company, 1986.

[2] Shilpi Banerjee. "The Compression Handbook". Third Edition. Starkey Hearing Research & Technology.

[3] Soon Suck Jarng, Carl Swanson, Frank Lee and Joseph Zou, "64 Bands Hybrid Noise Reduction and Feedback Cancellation Algorithm for Hearing Aid", International Journal of Control and Automation, Vol.7, No.1, pp.427-436, 2014.

Research on Representation and Similarity Measurement of ECG Series*

Chunkai Zhang[1], Jingwang Zhang[2], Haodong Liu[3],and Longfei Chen[4]
[1234]Department of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, China
[1]ckzhang812@gmail.com
[2]zhangjingwang@outlook.com
[3]1187624554@qq.com
[4]bluesmile2013@hotmail.com

Abstract *–In this paper, we propose a new method to improve the traditional ECG automatic recognition system, which can assist us to utilize KNN to identify the query sequence in the database with no medical knowledge. The method consists of IPDT_PLR and SegMode_DTW. IPDT_PLR inherits the advantages of the ITTP_PLR to retain the important trend turning point and it can effectively pick out the important points in the sequence through the top-down method. SegMode_DTW improves the original DTW. First, it reduces the searching range and improves the time efficiency by MSP. Second, it introduces the concept of line segment mode distance in DTW framework, which effectively overcomes the shortcoming of DTW without considering the tending of sample points. Third, it proposes the sub pattern distance based on DTW, which ensures the alignment of the two strong feature points, so that the trend of each segment is clearer, and the error matching of DTW is reduced effectively. At the same time, it also speeds up the measurement. Experiments show that the method we propose can be widely and efficiently applied to the similarity query of ECG time series data sets and it in most of the data sets can maintain the accuracy above 85%.*

Keywords: ECG, time series, representation method, similarity measure, mode distance, DTW

1 Introduction

The World Health Organization disclosure global three percent of the deaths are attributed to cardiovascular disease, heart disease has become an increasingly troubled human health is one of the major ills, and that it can be efficiently, accurate diagnosed in contemporary society has important practical significance. The traditional ECG automatic diagnosis system is a recognition algorithm which distinguishes specific waveform, extracts waveform features and then is classified [1]with the mature classification algorithm. However, the system requires developers have a strong medical knowledge, but also need to use large data sets to train the model, and also need regularly updating and maintenance the model. With the development of medical information, the database has been a huge amount of storage, all kinds of ECG data. We do not have to stick on the waveform feature extraction in this way, but we can directly utilize the ECG database to find a sequence with the most similar to the ECG waveform, and to similar class of waveforms and at last determine categories of query waveform.

The ECG sequence is a super high dimension data, without reprocessing directly using it will not only make the similar degree of change more slowly, and introduced noises may affect the final measure precision. This leads to an important research field of time series, which is the representation of time series. Time series representation is to extract more abstract features from the sequence. This part of the mainly two purposes: dimensionality reduction and noise reduction. In recent years, time series representation of the main representatives of: transforming time domain to frequency domain representation method (DFT, DCT), symbolic representation method (SAX [1]), piecewise linear partition method (PLR [2]). Transforming time domain to frequency domain representation method is generally said that we retain the low frequency part to fit the original sequence, so that you can retain a small amount of low frequency coefficients to represent the original sequence, but this method will lead to local information (such as extreme point and inflection point) and other important information loss. Symbolic representation method with some mapping rules predefined splits the original sequence into an orderly character set, then the string matching techniques measure the similarity between two strings, namely the two time series. This method is simple and intuitive, the disadvantage is that a large number of parameters need to be adjusted. Piecewise linear

* This work is supported in part by National High Technology Research and Development Program of China (No. 2015AA016008)

representation (PLR) is that with a cut-off rule predefined the important points in the sequence are picked out, and then sequentially connected, and at last the original sequence of piecewise linear representation comes out. This method can effectively preserve the shape of the original sequence, reduce the dimension, and high efficiency. The disadvantage is that the method is only applicable to the processing of some types of time series data and has a weak generality. The method is one of the common methods used in time series data representation. Similarity measure of time series is the core problem in time series data mining. However, due to high dimension, complexity and the driftage of the time series data, now the time series data lack universally applicable method to measure similarity. However, the characteristics of the ECG data are not aligned. So how to measure the two time series similarity? It also introduces another problem in the research field of time sequence, time series similarity measure. Euclidean distance [3] is a relatively simple method, it has the advantages of intuitive, simple calculation, low time complexity. The disadvantage is that the two time series must have the same length, and it is sensitive to drifting time axis. Cosine distance [4] is to compare the angle between the two vectors in the high dimension space. If the time sequence is regarded as a high dimensional vector, then we can calculate the similarity with the two angle vector time series. The advantages of this method are also easy to calculate, easy to understand. But this method does not apply to the time series data of severe turbulence. Edit distance [5] is using the idea of dynamic programming to convert a string into another string to the minimum number of operations. With application to the time series, we compare two adjacent elements in the time series, if the deviation is greater than a certain value, it will be set to 1, otherwise it to 0, so that we get a binary sequence representing the original sequence. The advantage of this method is a good use of string matching technique. The disadvantage is that there is need to be set to map time series discretization rules for string, time cost is larger. Longest common subsequence [6] and the edit distance is the string matching technique is used to measure the time series similarity. The longest common subsequence is first time series discretization as a string, and then compare two string sequence of the longest common sub on relative to the length ratio of the longer sequence length of the string, then use it as the similarity of two time sequences. Dynamic time warping distance [7] is based on the time axis stretching and compression to explain longitudinal data, allowing comparison between the time series length is not consistent, but the computing complexity is very high, many experts and scholars put forward many improved algorithm based on dynamic time warping algorithm to improve the computational efficiency, but not all the data on the set of applicable, commonly used method is limits on a curved path. The SegMode_DTW algorithm based on dynamic time warping algorithm is proposed. Experimental results show that it can effectively measure the similarity of two ECG time series.

2 Preliminaries

2.1 IPDT_PLR

Extreme point and inflection point of ECG time series often contains important information, are visually more important point and are the important basis on which doctors diagnose. Piecewise linear representation method is able to effectively retains the original sequence of shape, and ITTP_PLR is the outstanding representative. In this paper the IPDT_PLR is proposed based on the method of ITTP_PLR.

ITTP_PLR algorithm is divided into two parts: first part by traversing the original sequence it recognizes from all the trend of turning points, which are similar in the mathematical sense of extreme points, and this part of the sequence is viewed as potentially important points. The second part is to identify the key points from potential important points. In the algorithm at the bow and stern endpoints are important points, first of all the first point of the sequence is added to the important points set, and then while traveling trend turning point sequence, the distance D between the current point and the line consisting of the front point and the later point can be got, and if D is more than the predefined threshold ε, the current point will be added to the important point set, or the current point will be abandoned. The points in the important point set are connected in proper order, and in the end a polygonal line can be got, which is the representation of the original sequence.

The advantage of this method are to retain the important information of extreme points and piecewise high efficiency (time complexity is O(n)), but there are three major problems:

(1) When the shape of the sequence is monotonic, the method does not fit well. As shown in Figure 1 below:

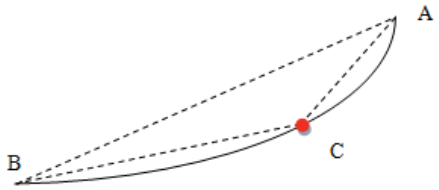


Figure 1 an example of ITTP_PLR

Seen from the diagram, the sequence (solid line ACB) is a monotonically increasing sequence and there is no extreme points. Applying ITTP_PLR PLR, we can obtain only fore A and the end B. Obviously the line AB represents the original sequence, fitting effect of which is poor. In the ECG time series data, there are many similar parts, so with it fitting results certainly are relatively poor. Through the above analysis, an ITTP_PLR algorithm fitting effect is poor and it does not consider inflection point information. As shown above, C is a turning point in the sequence, if the point C will be added to the important point set, to represent the original sequence by using the line BC and line CA, fitting effect will significantly improve.

(2) While selecting important points, ITTP_PLR does not consider the trend information of the parts.

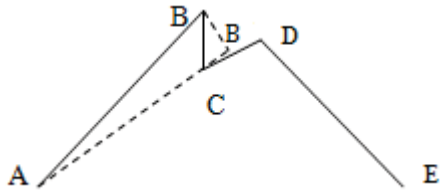


Figure 2 ITTP_PLR select the important point

From the figure 2, the trend of the sequence of the turning point has 3 points: B, C, D. ITTP_PLR identifies turning points in these three trends, but in the end B, C may be abandoned, and the important point set includes A, D, E. However the point B is more important (because B has a larger amplitude than D). Eventually join the important point sequence is only a little D, A, E. However, from the figure can see that B is a more important point (because the B amplitude higher than D), but D but was added to an important point sequence, and B is abandoned. This is because selecting important points, ITTP_PLR does not consider the trend of the current part information.

(3) ITTP_PLR does not consider the time span of the trend of turning point, as shown in Figure 3:

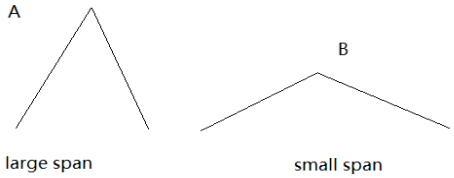


Figure 3 ITTP_PLR trend turning point in time span

In the ITTP_PLR PLR, the part in the figure 2.3 (A) has a higher priority to segment than the part in the figure 2.3 (B), because of the greater 2.3 (A) a point and the deflection distance. However, to reduce the fitting error, B can better reduce fitting error, because B trend turning a longer time span. The area below the B point is clearly larger than the area below the A point. In terms of reducing the fitting error, it is better to choose a large flat wave with a smaller span than the span. If piecewise fitting error as a condition for judging whether need to be segmented, rather than simply relying on deflection distance of the trend better segmentation.

In this paper, we propose a segmentation method based on the important points of the trend deflection distance, IPDT_PLR method. The method effectively improves the ITTP_PLR PLR methods fitting error shortcomings, the three problems of the method for ITTP_PLR PLR.

(1) In order to solve the ITTP_PLR problem that it cannot fit the trend of monotone sequences, while travelling the sequence of the original, IPDT_PLR retains the information of extreme points and inflection point.

(2) In order to solve the problem that ITTP_PLR does not consider the piecewise trend information in the choice point, IPDT PLR adopts a top-down selection important point. IPDT_PLR first connects the head and rear endpoint in the sequence, and the line consisting of the head and the rear point is regarded as the trend line. Then we calculate the piecewise on each point to the trend line distance and find the deflection trend point of the maximum distance. We take this point as the piecewise point, and the sequence is segmented into two sub part. In each sub segment to the implementation of the algorithm until it meets the conditions where the fitting error is less than the predefined threshold.

(3) In order to solve that ITTP_PLR in the segment does not consider the turning point of the trend of the time span, IPDT_PLR calculates every point to trend sequences and the deflection distance, and calculate deflection trend distance as the fitting error. If fitting error is greater than the threshold, it will continue segment; if not, then stop segmentation.s

The IPDT_PLR method can be effective to fit the ECG time series data through the improvement of the above three aspects. But because the method is top-down selection an important point, so the time complexity is $O(n \log n)$, where n is a turning point in the trend of sequence number.

2.2 SegMode_DTW

DTW algorithm is the most widely used one of the algorithms which measure time series similarity. It has a time allows dynamic axle bending, and many other advantages. However, there are still many problems in this algorithm.

(1) The computational complexity of similarity measure ($O(n^2)$): DTW uses dynamic programming to solve in order to meet the constraints of optimal bending path, so that the two time series similarity measure is a feature matching a feature (peak vs. peak, trough vs troughs). When the dimension of time series is very large, the time complexity of DTW algorithm is not acceptable.

(2) DTW algorithm does not guarantee that the obtained path must be the optimal path. Keogh pointed out that the DTW algorithm can't make all of the time series in a way to compare feature alignment. As shown in Figure 2.4:

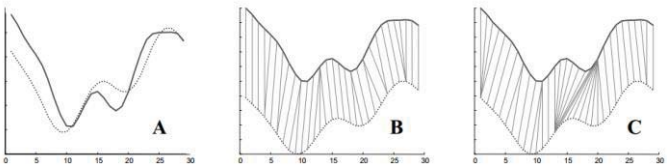


Figure 4 DTW algorithm can't be aligned with the characteristics of the example

Can be seen from the figure 4, matching the two time series of the original nature should be as shown in Figure 4 (b). But after DTW calculation that matching path is as shown in Figure 4 (c) is shown, from the figure can be seen, the characteristics of the two time series is not in alignment. This is because the DTW algorithm is in the constraint conditions, the calculation of the global optimal matching path. This may cause the value in order to obtain a smaller curved path, leading to a sequence of single point and another sequence of sequence matching a piece, which will lead to such as Figure 4 (c) above.

(3) DTW algorithm is based on the difference between the point and the point, but does not consider the shape difference.

The algorithm only considers the difference in the amplitude between the point and the point under the constraint condition. However, for single points of a sequence, each point has the time and amplitude information, but trend information. Suppose there are two points x, y which are located respectively

in the sequence Q and Sequence C, and x and y have a longitudinal axis of the same value, but the point x in a period of rising trend sequences in that point y in a period of decline trend series, DTW algorithm will give priority to the two point matching, because the minimum cost. But, intuitively, we don't want to match the two different trends.

Based on the DTW algorithm, the algorithm combines the characteristics of ECG time series data. The specific steps of the algorithm are as follows:

(1) Main sub pattern matching: find the main sub part in a sequence of unknown origin, and use the main sub model of the unknown sequence to match and find the approximate sequential approximation of M candidate in the database, so as to narrow the search scope.

(2) Piecewise linear representation of time series: with IPDT_PLR, segment candidate approximation sequences into candidate representations, and record the position of the peaks and troughs in the sequence of segments.

(3)Segmenting the sequence according to the extreme point: according to wave crests and troughs, the sequence is divided into three segment trend more nearly monotone sequence, which has three purposes. Firstly, it improves the computational efficiency, because time complexity of DTW is $O(n^2)$. With the increasing sequence length n, time complexity degrees is approximately exponential growth, so piecewise can improve efficiency. Secondly, the accuracy is improved, because DTW in monotone sequences have better effect. Thirdly, the matching error is reduced, because the wave crest and trough are strong features, DTW forces alignment to reduce global optimization and matching errors.

(4) Calculating of the fragment mode distance: it will be introduced explicitly later.

(5) Classification using KNN classifier. Find the most similar sequence of K. The K sequences of the label vote determine the subsequence category.

The sub sequence between the maximum point and the minimum point of the sequence is called the main sub pattern. The main sub model of ECG time series is more important in the visual piece sequence, and is also a sequence trend beginning to change a sequence. That main sub pattern matching narrows the scope of retrieval is based on the assumption of that the main pattern of similar time series is similar. The converse-negative proposition of this assumption is that if the main sub sequence modes are not similar, the time series will not be similar sequence.

Why do you want to choose the sequence between the maximum and the minimum as the main sub pattern? Because the main sub model is usually a shorter sub sequence in the entire time sequence, and changes in trend is most significant. Trend varying clearly is easy to distinguish, and short length determines the algorithm can possible complete the main sub pattern match in the shortest time.

The global maxima and minima point can be found, while travelling the sequence, and these two points are strong features of the ECG sequence. By the two vertices of the original time series to segment, there are the following two advantages: one is that after segmenting every part of the time series trend seems monotone. DTW in monotone sequences has an excellent performance, and can reduce the nonoccurrence of correct matching; the other one is DTW's time complexity is $O(n^2)$, when the time series dimension is large, the time complexity becomes approximately exponential growth. Therefore, the time complexity can be greatly reduced.

The length of the main sub mode in the two time series is no likely to the same. Therefore, it cannot be used directly with Euclidean distance to measure the main sub pattern similarity. If the main sub pattern of two time series in length have a big difference or opposite trend, it will return directly to infinity. Otherwise, it will be matching sequences of the main sub model according to the main sub model of unknown origin sequence length attainment, isotonic position if not an integer, aliquots of position (that is, the time axis digital) rounded up five into an integer position. The integer position as the matching points and the corresponding points of the query sequence, and then calculate the difference between them.

The main sub mode distance between the two series is the sum of the deviation of the corresponding points in the two main sub mode and smaller it is, and more similar the main sub patterns are. Distance Find the most similar K sequences to the query sequence, and the K sequences as candidate sequence. It only need to retrieve the most similar sequences in the K sequences, without having to search the most similar sequences in the database.

Line mode distance is the similarity measure of the distance between two line segments, segment pattern distance measure as follows:

Assuming there are two linear subsections, a and b. Segment of a represented by ($t_a^{start}, t_a^{end}, k_a, b_a$), and b by ($t_b^{start}, t_b^{end}, k_b, b_b$), where t^{start}, t^{end} respectively express starting and ending time, and k and b are respectively the slope and intercept. Their span of time are recorded as $len\_a = t_a^{end} - t_a^{start}$ $len\_b = t_b^{end} - t_b^{start}$ (assuming len Len_a<len_b). The length of the two line segments recorded as side_a and side_b.

We shift two segments to left aligned, and shifting horizontally distance is not considered to be counted, because DTW, for time axis stretching and compression, is not sensitive. The distance of the line segment is measured in 3 cases, as shown in Figure 5:

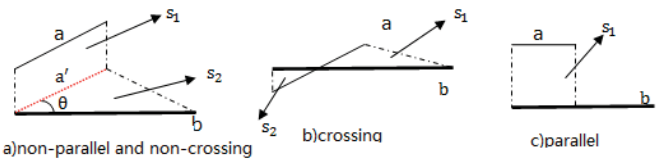


Figure 5 mode distance of the three cases

In the first case, as shown in Figure (a), the segment of line_a is moved to line_a' position. The vertical translation area is assigned as s_1, and line a' and line B angle area assigned as s_2, the distance of the line segment model for D=$s_1 + s_2$. It is explained that the similarity distance between line_a and line_b is converted into the sum of the distance of line_a and line_a' and the distance of a' and b. Where $s_1$and s_2 can be calculated by the following steps:

The height between the line segment a and the line segment a' can be calculated by the formula (1):

$$\boldsymbol{h} = \frac{\left|b_a - \mathrm{b}_{a'}\right|}{\sqrt{1+\mathrm{k}^2}} \quad (1)$$

Thus it can be concluded that the area of s_1 is as shown in formula (2):

$$\boldsymbol{s_1 = side\_a' \times h} \quad (2)$$

The sine H value of the included angle between the two line segments can be determined by the formula (3):

$$\boldsymbol{sin\,\theta} = \frac{|k_a - k_b|}{\sqrt{(1+k_a^2+k_b^2+k_a^2k_b^2)}} \quad (3)$$

The angle between the line segment a' and the line segment b can be calculated by the formula (4) in the area of the s_2:

$$\boldsymbol{s_2 = \frac{1}{2} sin\,\theta \times side\_a \times side\_b} \quad (4)$$

In the second case, as shown in Figure 5 (b), first to get the crossing point, then calculates the length of each truncation line, using the area formula obtained s_1, s_2 area, concrete can obtained by the following steps:

Intersection of two line segments, can be calculated by the formula (5) two line segments of the intersection position.

$$\boldsymbol{k_a \times t + b_a = k_b \times t + b_b} \quad (5)$$

The position of the intersection point is $\left(\frac{b_b - b_a}{k_a - k_b}, \frac{k_a b_b - k_b b_a}{k_a - k_b}\right)$ obtained by the formula (5).

Using formula (4) s_1, s_2 can be obtained, thus we can get line pattern distance D= s_1+ s_2.

For the third situation, as shown in Figure 5 (c), what only need to calculate is the vertical translation area s_1.

However, considering the difference of the length between two line segments, we also need to add a length difference and a relatively small gamma, s_1 area such as formula (6):

$$\boldsymbol{s_1 = side\_a \times |b_b - b_a|} \quad (6)$$

The line pattern distance of line segment a and line segment a is shown in the formula (7):

$$\boldsymbol{d = s_1 + \gamma \times (side\_b - side\_a)} \quad (7)$$

Through the above steps we can calculate the pattern distance of two lines, but in the algorithm, what we expect is two line segments of nearly equal length are matched, so line mode distance is multiplied by a weight, which is set to the absolute value of the difference between the two segments of length. The pattern distance of the line segments is shown in the formula (8):

$$\boldsymbol{SegDistance = d \times (len_b - len_a)} \quad (8)$$

Because the number of the corresponding segment of the line may be different, and DTW can measure the similarity of sequences of different lengths, in this paper DTW is applied to calculate the segmental pattern distance. In other words, in the outer loop of the algorithm is still use the DTW, while in the inner element it is the calculation of the distance of the line segment model, instead of computing the difference between points.

Suppose two time series segmented with the extreme points by the IPDT PLR, two line segment lists can be got, respectively denoted segs_list1, segs_list2. The length of segs_list1 is m, and that of segs_list2 is n. Due to needing to record matching path, so we assign a $(m+1) \times (n+1)$ cumulative distance matrix as M. At the beginning of the algorithm, the starting poing in one series corresponds to the starting point in the other one, and the end to the end. That is, DTW should start from the upper left corner of the matrix to the lower right corner. So the first row and the first column in the matrix M in addition to the upper left corner of the elements are initialized to infinite number and the upper left corner is set to 0, which can insure that the algorithm is starting from the upper left corner. What the M[i,j] represents is the similarity distance between the first i segments in one sequence and the first j segments in the other sequence. The algorithm scans the matrix with the line priority, with formula (9) to update the $\boldsymbol{M}[i, j]$ the value.

$$\boldsymbol{M[i, j] = d[i, j] + \min\{ M[i-1, j-1],\ M[i, j-1], M[i-1, j] \}} \quad (9)$$

The d[i, J] in the formula (9) is segment mode distance between the i-th line segment in the first sequence and the j-th one in the second sequence. Until M[m, n] is updated, the algorithm will be completed. The greater the value of M[m,n], the lower the similarity of the two time series, and vice versa.

The specific algorithm is shown in algorithm 1:

Algorithm 1 SegMode_DTW

Input: *segs_list1*: sequence Q linear segment
segs_list2: sequence C linear segment

Output: *dist*: distance between Q and C

```
CAL_DTW_DIST(segs_list1， segs_list2):
(1)  m = seqs_list1.length
(2)  n = seqs_list2.length
(3)  M is a matrix of m × n
(4)  for i =1 to m
(5)      for j = 1 to n
(6)          M[i,j]=SegDistance
     (segs_list1[ i ],segs_list2[ j ])
(7)  for i =1 to m
(8)      M[ i，  1 ] = INF
(9)  for j = 1 to n
(10)     M[ 1，  j ] = INF
(11) M [ 1，  1 ] = 0
(12) for i = 2 to m
(13)     for j = 2 to n
(14)         M[i，  j] += min(M[i-1，j-1],M[i, j-1],M[i-1，
     j])
(15) return M[m，  n]
```

3 Conclusions

In this section, there are time series similar measurement methods which are commonly used in several methods for comparison, and these similarity distance measure method are

respectively ED, DTW, DTW (c), EDR, LCSS, Swale, Spade [9], TSBF [10], CID [11], FSH [12], SegMode_DTW. These methods cover all categories of similarity measure method. Among them, ED, Swale are based on lock step, DTW, DTW (c) and CID are based on elastic, LCSS, EDR are based on mode, Spade is based on threshold, FSH is based on model. Evaluation criteria are used to classify the data set with the correct rate and time consuming, and the results are from [13] and [14].

Table 1 Comparison among algorithms above

Similarity method	CinC_ECG_torso	ECG200	ECGFiveDays	TwoLeadECG
ED	0.949	0.840	0.882	0.871
DTW	0.835	0.780	0.846	0.967
DTW(c)	**0.994**	0.850	0.878	0.930
EDR	0.989	0.790	0.889	0.935
LCSS	0.943	0.830	0.768	0.854
Swale	0.943	0.830	0.710	0.851
Spade	0.850	0.740	0.735	**0.983**
TSBF	0.748	0.860	0.817	0.954
CID	0.916	0.890	0.782	0.768
FSH	0.836	0.770	**0.996**	0.910
SegMod_DTW	0.961	**0.920**	0.877	0.889

In addition, the experiment also measures the performance of the proposed algorithm from the time efficiency. The comparison includes the time efficiency of ED, DTW, and DTW (c), which are the most representative. ED's time complexity is O(n), DTW for time complexity is$O(n^2)$, and DTW (c) 's time complexity is O (sn), s is a constant. They are shown in table 3.2:

Table 2 time consuming results for searching similar sequences of various algorithms (s)

Similarity method	CinC_ECG_torso	ECG200	ECGFiveDays	TwoLeadECG
ED	24.4	0.4	1.0	0.9
DTW	60085.0	223.4	877.2	457.6
DTW(c)	3069.5	18.2	51.1	7.5
SegMode_DTW	119.6	9.4	7.5	5.4

From the table 2 we conclude that SegMode_DTW is better than DTW and DTW(c) in time efficiency. With respect to ED's linear time complexity, the running time of SegMode_DTW is just a constant times of ED's. So SegMode_DTW can be used to quickly query similar sequences in large data sets.

SegMode_DTW in ECG200 data sets performs best, because the data contain more noise points, and the methods based on point matching is easily affected by noise. SegMode_DTW is based on line distance of the pattern, the original sequence after piecewise representation is compressed in dimension and denoised effectively, so it avoids the interference of noise. In TwoLeadECG, DTW has a far better performance than ED, indicating that the data set for the time axis of migration and compression is more sensitive. SegMode_DTW is based on DTW, and it has a certain tension compression capability. However, in the calculation of the distance of the line segment, the time axis cannot be dynamically bent. So, in this data set SegMode_DTW performance weak. In ECGFiveDays, DTW and SegMode_DTW do not work very well, indicating that this flexible way in this type of data does not arrive in what we want. In CINC_ECG_torso, ED is far better than the performance of DTW. Because the data set is basically feature alignment. Applying DTW to it, we will get a false match. While SegMode_DTW often has a good classification effect. In fact, we can explain that the SegMode_DTW is a similarity measure algorithm with limited bending capacity.

The original intention of SegMode_DTW is to improve the time efficiency of the DTW. There are three aspects to make DTW better. Firstly, the piecewise linear representation of time series, can effectively reduce the dimension. Secondly, is the calculation of line distance of the pattern, it is no longer the point matching method but the line pattern distance. Thirdly, according to the maximum and minimum values of a sequence segment, the time complexity of SegMode_DTW is O (k^2), where k is the number of the line segment representation.

The above experimental results, we can see, both in terms of accuracy and in time efficiency, SegMode_DTW than most similarity algorithm has more advantages.

From the accuracy point of view, DTW is based on the distance between points, and it did not consider the sequence of shape. The outer frame of SegMode_DTW is DTW, but in the inner algorithm the calculation is line mode distance, which consider the length of a line segment, ramp rate and amplitude of three aspects. It can be seen that SegMode is a kind of similarity measure method based on shape instead of point.

The main sub pattern matching gets 8 candidate sequences which are the approximations of the query sequence, and in the 8 candidates, we will find the most similar waveform. The experimental results are shown in table 3.3.

Table 3 Comparison of the classification accuracy of the main sub pattern matching.

DataSet	SegMode_DTW	Sub_SegMode_DTW
ECG200	92.0%	87.0%

CinC_ECG_torso	96.1%	87.2%
ECGFiveDays	87.7%	82.8%
TwoLeadECG	88.9%	85.6%

SegMode_DTW don't have a main sub partern, but Sub_SegMode_DTW has one. In addition, we also compared computation time with and without the main mode, the experimental results are shown in table 3.4.

Table 4 Time consuming comparison of the computation of the main sub pattern matching (s)

DataSet	SegMode_DTW	Sub_SegMode_ DTW
ECG200	9.4	1.3
CinC_ECG_torso	119.6	62.0
ECGFiveDays	7.5	4.9
TwoLeadECG	5.4	2.9

In table 3, with the main sub pattern matching the classification accuracy has declined, indicating that the main sub pattern match may do some underreporting behavior. But in table 1, although main sub pattern matching method classification does not work best, it not worst in all of the methods. The method in most of the data sets can maintain the accuracy above 85%. And combined with advantages of the method in the similarity metric efficiency, the method is completely used to assist physicians in the diagnosis of disease.

In the computational time-consuming, the main sub pattern matching, greatly reduced search scope, which can in a shorter period of time to realize similarity measure. In table 4 we can also see the main sub pattern matching algorithm computational complexity degree decrease obviously.

Above experimental results analysis, we can draw that SegMode_DTW can be widely and efficiently, accurately applied on the ECG time series data set similarity queries.

4 References

[1] Psaila R A G, Wimmers Mohamed &It E L. Querying Shapes of Histories [J]. 1995, 5(3): 538-539

[2] Keogh E. A Decade of Progress in Indexing and Mining Large Time Series Databases[C]. Proceedings of the 32nd International Conference on Very Large Databases. VLDB Endowment, 2006: 1268-1268.

[3] Keogh E, Palpanas T, Zordan V B, et al. Indexing Large Human-motion Databases[C].Proceedings of the 30th International Conference on Very Large Databases. VLDB Endowment, 2004: 780-791.

[4] Bloomfield P. An Exponential Model for the Spectrum of a Scalar Time Series [J]. Biometrika, 1973, 60(2): 217-226.

[5] Chen L, Ng R. On the Marriage of LP-norms and Edit Distance[C]. Proceedings of the 30th International Conference on Very Large Databases. VLDB Endowment, 2004: 792-803.

[6] Morse M D, Patel J M. An Efficient and Accurate Method for Evaluating Time Series Similarity[C].Proceedings of the ACM SIGMOD International Conference on Management of Data. ACM, 2007: 569-580.

[7] Keogh E J, Pazzani M J. Derivative Dynamic Time Warping[C]. Proceedings of SIAM International Conference on Data Mining, 2001, 1: 5-7.

[8] Morse M D, Patel J M. An Efficient and Accurate Method for Evaluating Time Series Similarity[C].Proceedings of the 2007 ACM SIGMOD international conference on Management of data. ACM, 2007: 569-580.

[9] Chen Y, Nascimento M, Ooi B C, et al. Spade: On Shape-based Pattern Detection in Streaming Time Series[C].Proceedings of the 23rd IEEE International Conference on Data Engineering & ICDE. 2007: 786-795.

[10] Mustafa Gokce, Baydogan, George, Runger, Eugene, Tuv. A Bag-of-Features Framework to Classify Time Series [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(11):2796-2802.

[11] Batista G E, Keogh E J, Tataw O M, et al. CID: an efficient complexity-invariant distance for time series [J]. Data Mining and Knowledge Discovery, 2014, 28(3): 634-669.

[12] Rakthanmanon T, Keogh E. Fast shapelets: A scalable algorithm for discovering time series shapelets[C].Proceedings of the 13th SIAM conference on data mining (SDM). 2013: 12-15.

[13] Wang X, Mueen A, Ding H, et al. Experimental comparison of representation methods and distance measures for time series data [J]. Data Mining and Knowledge Discovery, 2013, 26(2): 275-309.

[14] Grabocka J, Schmidt-Thieme L. Invariant time-series factorization[J].Data Mining and Knowledge Discovery,2014,28(5-6):1455-147

Diagnostic tool development on embedded system for heart fitness measurement

M. Á. Goda[1], A. Tihanyi[2], and I. Osztheimer[3]
[1]Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest, Hungary
[2]Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest, Hungary
[3]The Heart and Vascular Center, Semmelweis University, Budapest, Hungary

Abstract— *The main objective of this study is the measurement of the activity of the autonomic nervous system through Heart Rate Variability (HRV). Determination of the ratio of parasympathetic and sympathetic tone has meaningful clinical benefit in the case of certain disease´s diagnosis, and a possible follow-up may be in the field of competitive sports.*

An embedded system was created to make measurements on a specific microcontroller. Instead of the previously studied Electrocardiogram (ECG) curves, we processed the more easily recorded Photoplethysmogram (PPG) signals by Pan-Tompkins peak detection algorithm. For this measurements we used the InnoCare Pico tool, which was provided by the Innomed Medical Ltd..

The algorithm enables the determination of the parameters of HRV, which is based on the obtained data from the PPG signals. We will cover the main parameters of the tool, resource requirements and the calculation capacity of the implemented algorithm.

Keywords: Heart Rate Variability (HRV), Photoplethysmogram (PPG), Embedded System, Pan-Tompkins algorithm

1. Introduction

Heart rate variability (HRV) analysis related research is increasing in the last half-decade, however the research has not been implemented widely in clinical usage. This is due to the often lengthy investigation and the requirement of expertise. On the one hand, this kind of equipment is primarily used during a 24-hour Holter monitoring, which is quite specific medically. On the other hand, the commercially available sports equipment is often very expensive and too sophisticated, and its diagnostic relevance is not necessarily reliable.

The tool presented here is designed to be relatively easy to be implemented not only for doctors, but also healthcare workers, perhaps even the public. Our aim was short-term monitoring. We did not want to validate the HRV analysis, because it has been a standard procedure for the last 20 years. Our goal was to develop a tool that makes real-time assessment of the HRV parameters without computer. Accordingly, it is an online embedded system, which is able to calculate the HRV parameters.

The tool development is especially for clinical purposes, which helps the doctors to follow the patient´s medical condition in the direct diagnosis. Even basics such as blood pressure, blood glucose level, the amount of the blood lactic acid or other diagnostic parameters are not necessary information, by themselves for the doctors. However, such data may have diagnostic force and aid doctors in becoming aware of the health status of the studied subjects.

Unfortunately, infarct is an especially Hungarian endemic, with many social implications. In this area, the biggest risk segment of the population is that of 40-year-old men, who are fathers in many cases and who accomplish useful tasks in terms of the society at work.

The number of heart attacks is growing continuously in the population thanks to the advanced medical and technical achievement. Therefore the patients do not die during the acute event like earlier, so the number of heart failures continues to increase. It is therefore becoming increasingly important to conduct better HRV analysis, which would help the work of doctors so that they would be able to predict cardiovascular disease, the occurrence of complications and these alterations.

2. Neurobiological mechanisms behind

If we would like to get to know any information about the state of health of the human body, we can use the following models. The main components of the exercise capacity are the heart and vascular conditions, pulmonary regulation and metabolic processes. These three components are like interlocking gears.

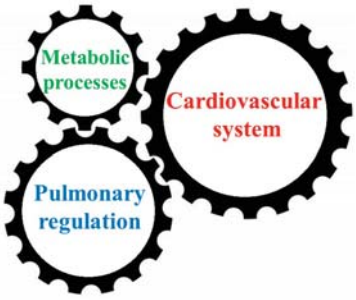


Fig. 1: The model of human organ

In simple terms, it could be said that the blood is the transport medium, which transmits the fresh oxygen to the cells metabolic processes.

2.1 Autonomic nervous system

The autonomic nervous system is responsible for maintaining homeostasis. It is closely related to the endocrine system. The autonomic nervous system is responsible for regulation of the central and the peripheral nervous system.

The autonomic nervous system has an important role in the regulation of organs and tissues, which include the cardiac muscle, the smooth muscle and the exocrine gland. The majority are not aware, however, of the autonomic nervous system, in which preganglionic neurons are located in the central nervous system centres[1].

The postganglionic cell bodies of neurons form autonomic ganglia, as ganglia outside the central nervous system. The postganglionic neurons are the paravertebral ganglia in the case of the sympathetic nervous system, and the ganglions are located in the wall of the organs in the case of parasympathetic nervous system, which regulation is fast. However, it is extending in the whole body, namely one postganglionic with several other neurons preganglionic neurons forms a synapse.

The sympathetic and parasympathetic nervous system has opposite effects in terms of most of the body. However, the two neural control functions are interconnected, and the balance of the activity can maintain the homeostasis.

2.2 The cardiac conduction and stimulus formation

The heart conduction system controls the generation and propagation of electrical signals or action potentials. They cause heart muscles to contract and the heart to pump blood.

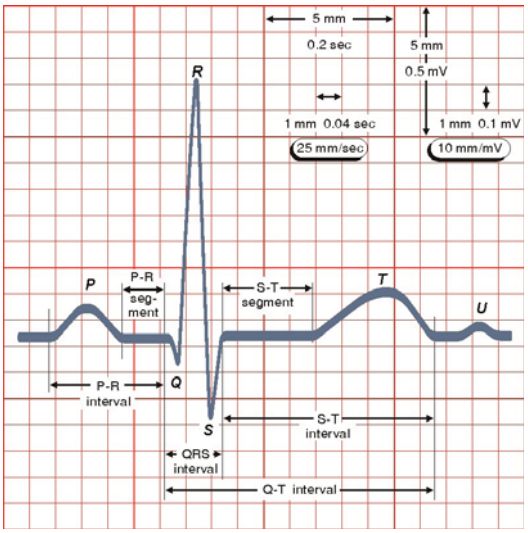


Fig. 2: The ECG QRS Complex represents this rapid ventricle depolarization. Atrial depolarization also occurs in this time. But any atrial activity is hidden on the ECG by the QRS complex [2].

This electrical activity can be measured by electrodes placed at specific points on the skin through the recording known as Electrocardiogram (ECG). A tracing of the overall electrical activity of the heart is possible, resulting from the propagation of many action potentials.

2.3 Photoplethysmograph

The blood oxygen level is in accordance with pulmonary processes in the cardiac cycle. The pulse is a peripheral output of the heart rate. Photoplethysmograph (PPG) measures the oxygen saturation (SpO2 level) of the peripheral capillaries and the volume of blood capillaries. SpO2 shows that the percentage of haemoglobin (Hb) binds oxygen molecules, i.e. how many percent saturated with oxygen. The PPG signal represents the blood volume, which corresponds to the pulse wave.

3. Heart Rate Variability

Cardiovascular autonomic nervous system activity results in different fluctuations in heart rate, which is called heart rate variability. The heart rate variability (Heart Rate Variability / HRV) [3] analysis is a non-invasive method that examines the autonomic nervous system. The studies and research go back to the 1980´s.

The heart rate frequency analysis is suitable to measure the activity of the autonomic nervous system. HRV reduction may indicate several pathological effects: sudden cardiac death, coronary artery diseases, cardiac failure, acute myocardial infarction, hypertension, renal failure, in varying degrees of damage to the brain etc.

3.1 Analysis of Heart Rate Variability

The HRV analysis is divided into two parts: the time and frequency domain analysis. As a general rule, it can say that the short-time (5 minutes) images are used for analysis of the frequency, and for the time domain a 24-hour analysis usually is required [4].

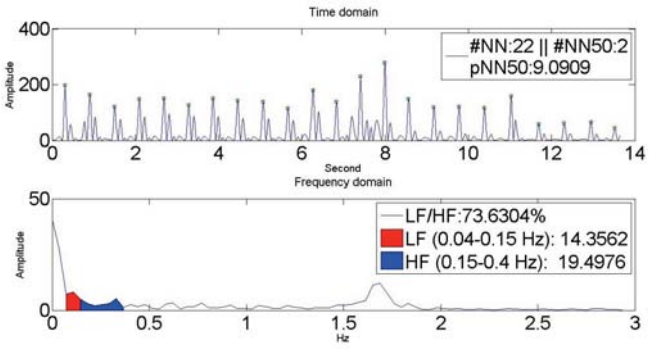


Fig. 3: Results of Time and Frequency Domain Analysis of HRV with the Pan-Tompkins algorithm. #NN is the number of peak to peak intervals. NN50 is the number of pairs of adjacent NN intervals differing by more than 50 ms in the entire recording. The pNN50 is the ratio between NN50 and the total number of NN intervals. The ration of the low and high frequency (LF/HF) integrals are depicted in the frequency domain

The problem of short-term analysis is that low frequency vibrations cause distractions. In contrast, the long-running analysis of the confounding factors may have environmental impacts.

3.2 Time Domain Analysis

This is a continuous ECG recording, which is suitable for examination of heart rhythm or measuring distance of the QRS complexes. The time range can be evaluated with the following parameters:

- Standard deviation of normal R-R intervals (SDNN).
- Standard deviation of the average normal R-R (SDANN).
- Root mean square of successive R-R interval differences (RMSSD).
- Ratio between NN50 and the total number of NN intervals (pNN50 or HRV index).In the *Fig. 2* 1.25 mm indicates 50 ms, i.e. 1 + $\frac{1}{4}$ small squares.

At the time-domain analysis it is necessary to understand the innervations of sympathetic and parasympathetic nerves, since the heart rate is controlled by the innervations. The parasympathetic nerve controls the sinus node by the vagus nerve. The sympathetic nerve is innervated at the lower neck 1-2, 5-6 and the upper thoracic segment. As a result, the expansion of the lungs during inhalation inhibits the parasympathetic path, which results in a sympathetic dominance, thus increasing the frequency of the heart. The effect disappears during exhalation: the parasympathetic nerve will dominate again, so as a result, there is a decrease of the heart frequency. This phenomenon is known as respiratory sinus arrhythmia. The pNN50 indicates that variability.

3.3 Frequency Domain Analysis

The spectrum can be divided into two components at the analysis of the short term HRV: a high frequency (HF) component (0.15 to 0.40 Hz) and a low-frequency (LF) component (0.04 to 0.15 Hz). These frequency components are involved in the parasympathetic and sympathetic nervous systems in the regulation . For the time being it is not yet known exactly what role the very low (VLF and ULF) frequency components play. The power of the frequency components is correlated with the following parameters [5]:

- HF: RMSSD and pNN50
- VLF and LF: SDNN
- ULF: SDNN and SDNN

The parasympathetic activity is characterized by pNN50 and RMSSD parameters, while SDNN parameter indicates the sympathetic activity. The high frequency components are responsible for the parasympathetic, while the low-frequency components for the sympathetic response. However it has been shown [6] that the SDNN and the SDANN, correlate with the ultra-low frequency domain (ULF); also these parameters indicates the control parasympathetic and sympathetic heart control.

Pulse spectrum changes in heart can be approximated by parametric and non-parametric methods:

- Autoregressive model:heart rate detection at abrupt changes, (whether in 1-5 minutes monitorin)
- Non-parametric models: Based on Fourier analysis, which is needed to transform the data of the R-R interval in frequency spectrum

3.4 The effects of abnormal HRV parameters

Before beginning to discuss pathological cases, it is important to understand the correlation between the autonomic nervous system and heart function. The parasympathetic nerve uses acetylcholine as a neurotransmitter, which breaks down and acts fast. However, it results in slow heartbeat. In contrast, the sympathetic nervous system uses noradrenaline (or adrenaline) as the neurotransmitter, which could slow down the process; and it breaks down slowly. This in turn accelerates the heart function. So it can be said that the sympathetic nervous system inhibits the secretion of acetylcholine and stimulates the secretion of adrenaline and noradrenaline in case of emergency.

An analogy of sympathetic regulation can be the gas pedal, which escalates the motor relatively slowly, and the parasympathetic control can be the brake, which is able to stop vehicle suddenly. It is important to see that the control of the sympathetic and parasympathetic nervous systems is different at the normal and pathological cases. HRV analysis can help to determine the effects of the cardiac autonomic nervous system.

HRV analysis is based on rapid fluctuations of the sympathetic and vagal activity. If the pNN50<RMSSD 3% and <25 ms, then the activity of the vagus nerve is decreasing. Similarly, if the SD<50 ms and SDNN<100 ms than abnormal activity of sympathetic regulation be observed [3]. The LF and HF ratio arises from to the sympathetic and vagal interaction. The balance of sympathetic reflex is disturbing with the reduction of LF/HF ratio and vagus nerve activity becomes dominant. So based on these it can be said that low HRV values imply sympathetic dominance in general, which could be the result of high sympathetic and/or low parasympathetic activity.

The clinical state of the patient must be known to be able to draw exact conclusions in clinical cases. The main factors are ethnic roots, the type of disease (diabetes, arteriosclerosis, heart failure etc.). In addition, it also plays a role when the test was made, for example, early in the morning, or early afternoon time. For the healthy middle-aged man, the heart rate has reduced variability in morning; then arrives to minimum in the early afternoon, followed by a continuous increase in the value until the next morning. Thus, the sympathetic modulation can be described with the LF / HF rate and pulse rate.

3.5 HRV analysis for clinical applications

After more than 30 years of HRV analysis, the first clinical review appeared only in 1996. Although the number of articles on this topic is constantly growing in recent years [7], the clinical usage is still lagging behind. In June 2015, the latest review came out for clinical purposes since 1996 [3], which draws attention to the implication of HRV analysis.

The relationship of autonomic nervous system and various cardiovascular diseases is highlighted in the study of the American Heart Association and the European Society of Cardiology in 1996. Since then, this study has produced several standard parameters in the whole world and more than 5,000 scientific studies cite this [7]. The statistical indication of sudden cardiac death was one of the aims of this study at the myocardial infarction patients. The death prediction of left ventricular ejection fraction (EF) [1] and HRV indexes are same. The HRV guarantees a more precise prediction at the arrhythmia (sudden cardiac death and ventricular tachycardia).

Based on this study it would be possible to say, the indication of the HRV is better at arrhythmic cardiac death, in contrast with the not arrhythmic case. However, the HRV indexes are not obviously different after acute myocardial infarct at the sudden and not sudden cardiac death. All the same, HRV could be a predictor of sudden cardiac death.

These standard methods are still not implemented on embedded systems. Firstly, our aims were to determine the HRV parameters on embedded systems. The implementation of the latest methods is our overarching plan. The latest published clinical study is about the myocardial infarct and congestive heart failure, which are connected to the HRV analysis.

In Hungary cardiovascular diseases are still the leading cause of death. With better care of organisation of the acute myocardial infarction cases, there is a proportional increase of heart failure. Later these patients will be in the health care system with heart failure. Heart failure can be an inherited condition or an acquired cardiac disease. In the developed countries 1-2% of the adult population has certain forms of heart failure, and this rate increases to 10% over 70 years of age [8].

3.6 Usage of PPG signals at HRV analysis

The photoplethysmography (PPG) signals can be easily measured with finger and ear pulse oximetry, while the ECG requires at least three body surface electrodes. The movement of skeletal muscle produces artefacts in the recorded signal. It is significant at the ECG; meanwhile, the PPG can provide low-noise signal even at movement. In clinical usage an important aspect is simplicity. The PPG is simply like the blood pressure measurement.

The PPG based HRV analysis is referred to by several studies [9], [10] and it has strong correlation with the ECG achievements., which was proved also my studies. It was a mayor point of my further studies. In a PPG based HRV analysis was shown [10] the R-R intervals (RRI) of ECG are better correlated with the valley to valley intervals (VVI) than with peak to peak intervals (PPI). Differences between heart rate and pulse can be the implication of extra systole, since they do not appear in the periphery, because the electrical impulses do not result in a volume of outflow from the heart in real time.

At the PPG signals two peaks can be observed ($Fig.6$): the first is the end-diastolic pressure wave of the right ventricle, the second peak resulted at the beginning of systole, when the semilunar valves are closing and generate a short pressure wave. At the peak, detection can be an artefact, when the second pressure wave is higher than the first one. In contrast, at the VVI detection, it cannot happen because of the well separated systolic pressure.

4. Embedded systems

Embedded systems are an area of electronics and nano-electronics systems which are suited for particular tasks, such as: MP3 players, digital cameras, washing machines, telephones etc.. The type of hardware and type of program required for these embedded systems are quite specific.

The embedded ECG signal processing has several motivations in the regulation of human disease e.g.: essential hypertension, obesity, chronic renal diseases, diabetes, orthostatic intolerance, and congestive heart failure.

In this study, a Tiva C Series TM4C123G microcontroller was used, which is supported by the Texas Instruments Tiva C Series development board ($Fig.5$). The processor supports a number of peripherals, but it was used primarily to debug port and two UART channels. In this microcontroller 8 UART channels are available. All of them must be configurable because of testing, in order to monitor each processing part. At the final usage it is enough to have only two channels.

5. Data processing

For the measurement an InnoCare Pico tool was used, which was provided by the Innomed Medical Ltd. Firstly data had to be converted into the suitable data format in the microprocessor. The InnoCare Pico is a multiparametric and telemetric tool. It can measure three ECG leading, haemoglobin saturation, breathing cure and capnography. The data processing is divided into several parts:

- Communication system
- Extraction and processing of raw data
- Time domain analysis

[1] Ejection fraction is the fraction of outbound blood pumped from the heart with each heartbeat

- Frequency domain analysis

The online data processing was directly connected to RX and TX pins of microcontroller the InnocCare Pico tool. The receiving data were processed firstly by a state-machine on the microprocessor. The state-machine provides the extraction of ECG and PPG signals, too. This was a critical point of real time data processing. The microprocessor is able to display the stored data, online data, and the processed data, which was necessary for the tool development.

5.1 Time domain analysis

The Pan-Tompkins algorithm is one of the most popular real-time QRS detection algorithms [11]. First of all, the Pan-Tompkins algorithm was implemented in MATLAB, and later on also implemented on a microcontroller. The Pan-Tompkins algorithm is mainly used for ECG peak detection. That is why the filter parameters had to change.

The Pan-Tompkins algorithm is divided into 5 parts: low-pass, high-pass filter and band-pass filter, differentiator, squaring operation, moving-window integrator and peak detection.

- **Low pass filter**: it helps to filter the low frequency noises. The transfer function can be described by the following difference equation.

$$H(z)_{lp} = \frac{1}{32}\frac{(1-z^{-6})^2}{(1-z^{-1})^2} \quad (1)$$

The output $y(n)$ is related to the input $x(n)$ as

$$y(n) = 2y(n-1) - y(n-2) + \frac{1}{32}[x(n) - 2x(n-6) + x(n-12)] \quad (2)$$

- **The transfer function of the high pass filter**:

$$H(z)_{hp} = z^{-16} - \frac{1}{32}H_{lp} \quad (3)$$

the input-output relationship is

$$y(n) = x(n-16) - \frac{1}{32}[y(n-1) + x(n) - x(n-32)] \quad (4)$$

which cut off frequency is 5 Hz

- **Derivative operator**: The differentiation of signal is coming after the filtering, which helps to emphasize QRS complex and suppress P and T wave components of the signal.

$$y(n) = \frac{1}{8}[2x(n) + x(n-1) - x(n-3) - 2x(n-4)] \quad (5)$$

- **Squaring**: It has two important roles; it makes the results positive and emphasizes larger differences resulting from QRS complexes.
- **Moving-window integrator/smoothing:** it helps to smooth the signals with a moving-window integrator. At noisy signal can be used two times. The transfer function can be described by the following differential equation:

$$y(n) = \frac{1}{N}[x(n-(N-1)) + x(n-(N-2)) + \ldots + x(n)] \quad (6)$$

The structure of Infinite Impulse Response (IIR) filters are difference between in MATLAB and CMSIS[2]. MATLAB calculates with the direct form structure, which uses the poles and zeros of transfer function. Nevertheless, by the CMSIS provided lattice form is faster. The filter structures are equivalent.

The diagnostic parameters can be defined by the Pan-Tompkins algorithm. The pNN50 time domain parameter characterize the HRV. The filtering provides the local maximum searching, which takes the 0.5-3 Hz heart frequency.

5.2 Frequency domain analysis

Fast Fourier Transform (FFT) can represent the dominant frequency component of the input signal. The frequency spectrum of ECG and PPG are same. The dominant frequency component is the heart rate (0.8-3 Hz), breathing rate (0.3-0.8 Hz) and the vegetative tone (0.04-0.4 Hz). This frequency component can modulate each other, which can be eliminated by Principal component analysis at many signals.

We observed first the vegetative tone (0.04-0.4 Hz) at the frequency domain analysis. The integral of the frequency domain spectra is necessary to determine the sympathetic and parasympathetic activity. The lowest observed frequency is 0.04 Hz; therefore, the frequency accuracy must be the same, or less. The spectral resolution depends on the sampling frequency and the FFT accuracy. According to the Nyquis-Shannon sampling theorem, the bandwidth of the signal is half of the sampling frequency. The PPG was sampled with 75 Hz in our measurement.

$$Spectral\ resolution = \frac{\text{Bandwidth}}{\text{FFT size}} = \frac{75/2}{2048} = 0,01831\text{Hz}$$

The FFT was also well defined on the microcontroller. The resource requirements of the tool are visible, the cornerstone of the 32 bit microprocessor is the size implemented FFT. A 2048 accuracy FFT was implemented, which has good enough spectral resolution to the signal processing. The sympathetic and parasympathetic activity can be determined by the spectra of the signal. The MATLAB calculates primly with the trapeze integral, and it was also implemented on the microcontroller for communicability.

6. The measuring tool

The measurements were carried out in a calm atmosphere. Raw data were delivered by InnoCare Pico tool from PPG sensor to the pins of the microcontroller. The InnoCare Pico can calculate the heart rate, pulse and oxygen saturation, they were not used for own calculation.

[2]The Cortex Microcontroller Software Interface Standard (CMSIS) is a vendor-independent hardware abstraction layer for the Cortex-M processor series and defines generic tool interfaces.

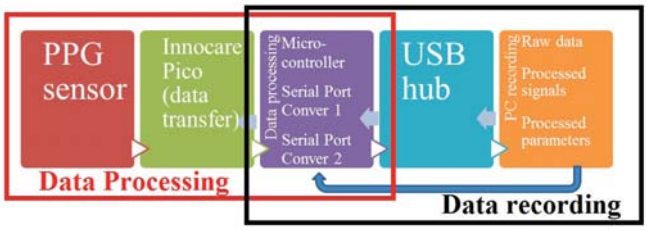


Fig. 4: The schematic figure of data processing and data capture.

The PPG signal was sampled at 75 Hz, and the ECG at 300 Hz by the InnoCare Pico. The input data were processed according to the given structure by a state-machine on the microcontroller.

The data were delivered to the PC in order to ensure reproducibility by a serial port and a debug port. The implemented state-machine and the data arriving were also tested physically by an oscilloscope. It was a critical point of the measurement to be really sure of the data arriving. In this construction the microcontroller is able to send data to InnoCare Pico, and receive from the PC.

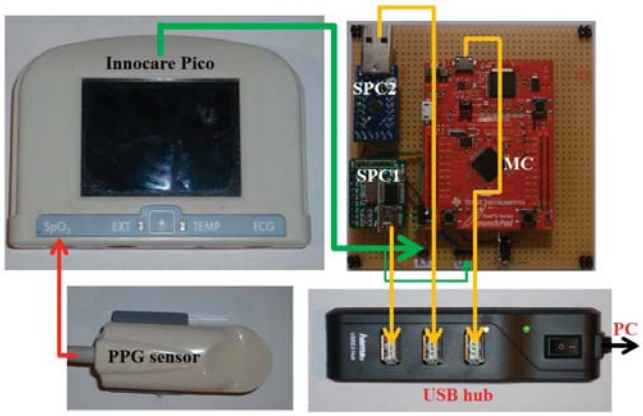


Fig. 5: The prototype of the tool. Data processing parts: PPG sensor, InnoCare Pico, serial port converters and microcontrollers. Date saving: microcontroller, serial port converters USB hub, PC.

For the measurements the main pillars were: simplicity and reproducibility. That is why the PPG signals used were inconstant with three leading ECG. Although the measuring tool has one input from the PPG sensor and one output to PC, it is possible to monitor simultaneously the raw data, the frequency and time domain curve and output processed parameters.

6.1 Data acquisition

The measurements were made on three different groups:

1) Infarcted peoples (15 subject)
2) Sportsmen (10 subject)
3) Normal test subject (20 subject)

The data acquisition was made at The Heart and Vascular Center of Semmelweis University (15 subject), at Faculty of Physical Education and Sport Sciences of Semmelweis University (10 subject) and at Faculty of Information Technology and Bionics of Pázmány Péter Catholic University (20 subject). Totally 45 test subjects were observed at the measurements.

According to the reported literature [3], the average HRV index (pNN50) would be around 35% for normal test subjects, it is higher for sportsmen, and definitely lower for infarcted people.

6.2 The tool resource requirements

Tiva ^{TM} C Series architecture offers a 80 MHz Cortex-M with FPU [3]. The power supply of the microcontroller system is 5 Volts, which is needed for the communication with the Innocare Pico. The microprocessor loads can be determined by the sampling frequency and size of the FFT.

The sampling frequency of the PPG sensor is 75 Hz. The constant load of the communication load is 3.8 milliseconds. The total load is a bit more because of the FFT usages, which is 28 milliseconds at every 2048th sample.

$$Processor\ utilization = \left(\frac{\text{Constant load of the communication load}}{\text{Sampling frequency}} + \frac{\text{Total FFT load}}{\text{FFT size}}\right) * 100 = \left(\frac{3.8 * 10^{-3}}{\frac{1}{75}} + \frac{28 * 10^{-3}}{2048}\right) * 100 = 28.501\%$$

It is visible that the total load of our microcontroller system is less than 30%, which can allow further additional development opportunities.

7. Results

The measurement results must be separated from the microcontroller implementation at the evaluation. Furthermore, it is necessary to review the resource requirements and the computing capacity.

7.1 Processing Capacity

The implemented algorithm is equal to the result of MATLAB, which is virtually the same. One fact is not evident, namely that various hardware issues may surface during data processing e.g.: memory misallocation, the usage not being an available variable, data mismatching etc.

[3] A floating-point unit is a part of a computer system specially designed to carry out operations on floating point numbers.

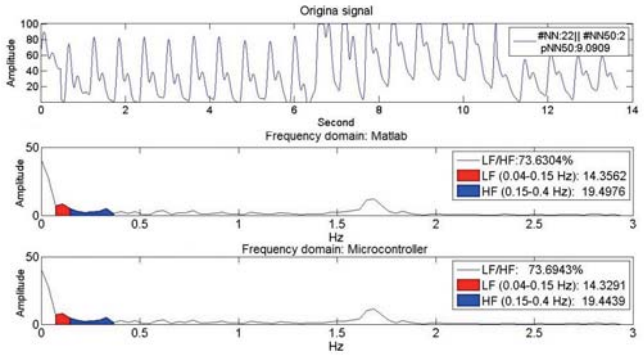


Fig. 6: Results of Pan-Tompkins algorithm in Matlab and microcontroller in the frequency domain.

The microcontrollers have a strong resource limitation that is why the optimization is important. Another source of error can be the soldering inaccuracy on the pin of the microcontroller, which can result in electrical background noises. These failures can be checked by an appropriate oscilloscope.

During the HRV analysis the data were recorded by PC. It was necessary for reproducibility. This signal can be reprocessed by the microcontroller. A five minute recording can be reprocessed by the microcontroller in 70 minutes, which is 10 times faster than the MATLAB data processing. The reproducibility allows further analysis. All of these requirements are indispensable for examining greater populations.

7.2 Measurements results

The reported literature and the prognosticate results are consistent with our measurement results. The HRV index was the lowest at the infarcted subjects, it was higher at the normal subject and the highest were sportsmen. Although neither the age, nor the sex ratio were the same for the observed subjects, even so this tool development reflects on the clinical significance of usage because of the simplicity and reproducibility.

It is a fact that the HRV index and the LF/HF ratio is decreasing by aging and health status deterioration. The low HRV index is not a primary indicator of sudden cardiac death, because it is subject specific. A good example for that is high blood pressure, which can be the result of cardiac vascular disease and heart failure. Although many people have healthy life with high blood pressure, but this has a diagnostic relevance at the whole population.

There were several prominent cases during the measurement. Low HRV indexes were expected at the infarcted subjects; nevertheless, in some cases it was higher than the variability of sportsman. It turned out that the observed patient has atrial fibrillation, which is a chaotic heart rate failure. However, it was not the goal to show that during the HRV analysis, this device is able to determine the atrial fibrillation.

Type	pNN50 (%)	LF/HF (%)	Avg. age	BMI (kg/m^2)	#Sub.
Infarcted	11.9±3.9	89.3±19.9	71	25.7±1.8	15
Sportsman	41.4±21.0	111.9±20.4	20	24.2±1.9	10
Normal	31.8±15.1	110.7±17.7	22	23.6±3.8	20

Table 1: HRV analysis results. The average HRV index (pNN50) would be around 35% [3] for the normal test subject, it is higher for sportsmen, and definitely lower for the infarcted peoples.

The low HRV index does not mean unambiguous heart failure, because it was observed for the sportsmen, too. However it was interesting fact that the HRV index of normal test subjects was definitely lower at smokers than others, who do regularly any kind of sport activity at least once a week.

The HRV analysis has relevant meaning for the health follow-up of subjects. A 24 hour HRV monitoring would be more precise, but a 5 minute analysis was able to show the main heart condition. As the hyperglycemia can be the result of diabetes or a current state after the meal, it is a prerequisite to measure it if somebody is diabetic. The continuous monitoring of HRV parameters can help doctors and sport trainers.

8. Discussion and Conclusions

The measurement results are consistent with the relevant literature. The main goal was a tool development, which makes the HRV analysis easier. Therefore the main achievement of this study is the algorithm implementation on an embedded system, which is able to carry out a real time heart condition assessment with PPG signals.

The results of the tool are reliable and produce accurate calculations, and the results of the measurements are reproducible. The tool has several development areas e.g.: internal memory upgrades and the development of a more compact measuring device.

The obtained results of the calculated diagnostic parameters form the basis of further algorithm implementations, which are reported in the recent standard methods [7]. These methods are capable of improvements. The techniques can be used for other nonlinear dynamic systems to describe more general heart conditions.

References

[1] F.Triposkiadis, G.Karayannis, G.Giamouzis, "The Sympathetic Nervous System in Heart Failure: Physiology, Pathophysiology, and Clinical Implications", *Elsevier Science Journal of the American College of Cardiology*, vol. 54, pp. 1747-1762, 2009.

[2] (2016) [Online]. Available: http://jonbarron.org/sites/default/files/images/ecg2.jpg

[3] M.Malik, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use", *John Wiley and Sons Annals of Noninvasive Electrocardiology*, vol. 1, pp. 151-181, 1996.

[4] B.Xhyheri, O.Manfrini, M.Mazzolini, et al. "Heart Rate Variability Today", *Elsevier Science Progress in Cardiovascular Diseases*, vol. 55, 2012.

[5] T.Bigger, J.Fleiss, R.Steinman, et al. "Correlations among time and frequency domain measures of heart period variability two weeks after acute myocardial infarction", *Elsevier ScienceThe American Journal of Cardiology*, vol. 69, pp. 151-181, 1996.
[6] J.Bigger; J.Fleiss, R.Steinman, "Correlations among time and frequency domain measures of heart period variability two weeks after acute myocardial infarction", *Elsevier Science The American Journal of Cardiology*, vol. 69, pp. 891-898, 1992.
[7] R.Sassi, S.Cerutti, F.Lombardi, et al. "Advances in heart rate variability signal analysis: joint position statement by the e-Cardiology ESC Working Group and the European Heart Rhythm Association co-endorsed by the Asia Pacific Heart Rhythm Society", *Oxford University Press Europace*, vol. 1, pp. euv015, 2015.
[8] J.Hradec, J.Vitovec, J.Spinar,"Summary of the ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012. Prepared by the Czech Society of Cardiology", *Cor et Vasa*, vol. 55, pp. e25-e40, 2013.
[9] G.Lu, F.Yang, J.A.Taylor, J.F.Stein,"A comparison of photoplethysmography and ECG recording to analyse heart rate variability in healthy subjects", *Informa plcJournal of Medical Engineering & Technology*, vol. 33, pp. 634-641, 2009.
[10] X.Chen, T.Chen, F.Luo, J.Li, "Comparison of valley-to-valley and peak-to-peak intervals from photoplethysmographic signals to obtain heart rate variability in the sitting position", *IEEE 2013 6th International Conference on Biomedical Engineering and Informatics*, vol. 33, pp. 214-218, 2013.
[11] R.M.Rangayyan, *Biomedical Signal Analysis: A Case Study Approach*, Univeristy of Calgary, Alberta, Canada, pp. 187-190, 2002.

THE RESEARCH OF HUMAN ACTIVITY STATE RECOGNITION BASE ON ACCELEROMETERS*

Chunkai Zhang[1], Jiayao Jiang[2], Guoquan Wang[3] and Zhiliang Hu[4]
[1234]Department of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, China
[1]ckzhang812@gmail.com
[2]616905919@qq.com

Abstract - *With the development of sensor technology and people pay more attention to the health, activity recognition has become a popular research direction, which is analyzing the acceleration sensor data and finding human activity pattern. So far, the research has mainly used the traditional classification techniques but the clustering method is seldom involved. Most of the current research methods only consider the prior knowledge, ignoring the characteristics of dynamic changes of data flow, resulting in classification model in the static data with high accuracy and bad practical experience. In addition, the current research methods do not take into account user differences, which cause serious individual problems. This paper has proposed using the K - Means clustering method to construct the human activity state identification model, and finally designed human movement activity system based on Android mobile phone. At the end of experiments, to prove the effectiveness of the clustering algorithm, three clustering algorithms and three classification algorithm of recognition results were compared to verify the model problem of individuality, and mature products were compared. The results show that human activity recognition model is feasible based on clustering methods, and the model has real-time, lightweight and easy adjustment features.*

Keywords: acceleration sensor, stream data mining, human activity recognition, K-Means clustering.

1 Introduction

In today's rapid development of science and technology environment, the internet technology permeates all aspects of life, such as the internet financial, medical internet, internet education, etc. As people living standard rises, people pay more and more attention to health. Many internet companies seize the business opportunities and develop various applications as an aid to improve people's health. Most of these applications have one thing in common: by monitoring the body's daily motion, calculating the amount of exercising in a day to provide advice and reminders. Due to the variation of people's daily movement, such as walking, running, sitting, etc., and energy consumption of each sports, it is important to design an exact activity state recognition algorithm.

The recognition research of human movement state is mainly analyzing human motion data, using data mining and machine learning method to mining data in the fixed model. There are two main aspects using the model: A. motion state recognition: according to the existing data to tell someone, which one is in the state of motion; B. identification: calculating the similarity between the data and model to distinguish the data belongs to whom, it is mainly used in forensic aspect [1].

At present, according to the study of different data types, motion recognition research is divided into the following three directions [2]:

1) Based on the movement of sound recognition: this method has obvious flaws: applicable scenario is very limited and the cost is expensive.

2) Activity state recognition based on video: this method is mainly based on the analysis of mining data from the camera to capture human body's movement range. The video data is influenced by weather, light, distance, azimuth, and other factors. Therefore, the scenario used is also very limited.

3) Activity state recognition based on wearable devices: this method is mainly through wearing sensors and analysis equipment to collect data. Relative to the above two methods, this method has the following advantages: a, low-cost and portable, the price of small equipment is lower and it is easy to wear; b, strong anti-interference, the external influence of collecting data process is small; c, collecting capacity data steadily: wearable devices can guarantee us to receive the data continuously.

Motion state recognition not only can help people to monitor the movement condition of the day, but also is a major research field of smart home. It may bring a new way of human-computer interaction, such as motion sensing games, so that the life of people is more intelligent [3]. Studying motion state recognition for improving the quality of human life has great significance.

The data that the three-axis acceleration sensor collected is a kind of typical time series data [4], which has infinite length, which makes it more challenging than the static data mining. Due to the characteristics of data streams, the typical traditional data mining technology cannot be directly applied to stream data. With the collected data, the mode may not be available to the coming data, it causes the phenomenon of

* This work is supported in part by National High Technology Research and Development Program of China (No. 2015AA016008)

"concept drift"[5], it is easy to cause inaccuracy of movement state recognition. Therefore, the classification model of this paper should be able to adjust itself as the data changing, but the attempt to design a classification algorithm which is applicable all the time is not desirable.

2 Basic Research

2.1 Data collection

At the beginning of the study, data collection was a very big problem, since at that time the sensor technology was not as advanced as presented. The sensor size was larger, users were not willing to carry them, so the data was only from the laboratory, and the data set and the value of the research were small. As the development of wearable equipment and the rise of cloud services, producing a surprising number of user data, which has huge potential research value. In addition, most of these motion data is acceleration sensor data, so the topic is based on the acceleration sensor for studying the human movement state recognition.

Acceleration sensor [6] is a kind of electrical equipment used to measure acceleration. By physical knowledge, we can know that acceleration can well reflect the movement of objects or people.

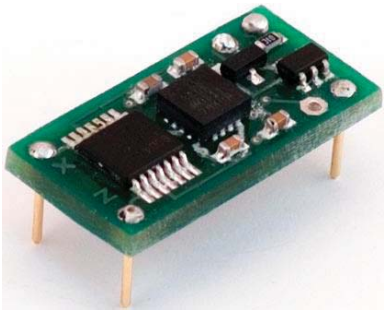


Figure 2-1 triaxial acceleration sensor

Figure 2-1 is a triaxial acceleration sensor device. Most smart phone on the market has a built-in triaxial acceleration sensor. The frequency of the acceleration sensor is fixed, about 20 Hz [7], that is collect data once every 50ms. According to the following format to record the sampling point (x_i, y_i, z_i), of which, the x, y, z are three axis acceleration values, collection of data will contain a series of data points, follow-up study is to analyze the training data points.

Carry ways of smart devices have a great influence on the collected data [8], some people like to take it in hand, some people like to hang it in the chest, and some like to put it in front of the pants pocket. The ways of carrying will cause a greater difference between the data collected, so it influences the experimental results.

Data collected by equipment is mainly cached to a local file.

2.2 Data preprocessing

The original data collected by accelerometer contain all kinds of noise. The cause of the noise has: acceleration of gravity instability of sensors, hardware, and the influence of the human body shake [9].

The acceleration signal data pretreatment methods including de-noising, smoothing, sliding window segmentation, etc. [10] General process is as follows: first use filter method, and then to smooth data with the average method, at the end do the data segmentation.

2.2.1 DE-noising and Smoothing

At this phase, the main method is to use filter. Most used methods in motion recognition research are round filter and Chebyshev type I digital low-pass filter.

After De-noising, we need more data processing, such as the data smoothing. Mainly use the averaging for curve smoothing.

2.2.2 Data Segmentation

After the de-noising and smoothing processing, stream data needs to be split processing. Mainly because the data is time series data, has a high sampling frequency. The data is generally long, feature extraction and classification for a single data does not make sense. Using the sliding window method [11] can segment the stream data into small pieces, each represents should to be a sample, and we can do feature extraction on these small pieces of data. The window size can be fixed and can be changed. When window size is fixed, there is no theory to prove how can achieve optimal, but the results show that as long as the guarantee of each window sample point number is about 120[11]. So to the acceleration sensor has the frequency of 20 HZ, 6 seconds is the most appropriate window size. The unfixed window's size is divided according to the event and each window represents from beginning to the end of an event.

2.2.3 Feature Extraction

After completing data segmentation process, need to extract a set of features to represent the window sample, the group features can be used to measure the similarity between two window data. Using these features, the data window can be classified or clustering analyzed.

The method used in this art for feature extraction is called the time domain analysis. Mathematical statistics feature are considered first in this method, for example, the mean, variance, the maximum and the minimum. And then using some statistical machine learning method to train models, such as J48, SVM, neural network and so on. The results depend directly on the feature selection policy.

There are two methods of feature extraction, which are widely used, and they are frequency domain analysis and time-frequency analysis respectively. A typical representative method of frequency domain analysis method is the fast Fourier transform. This method has the characteristics of high accuracy and high computational complexity. Because the mobile phone hardware resources are limited, this feature extraction method is not suitable for use on a mobile phone. Wavelet analysis is a classical time-frequency method, which combines the characteristics of both time and frequency.

2.3 Learning model building

After data preprocessing is completed, should use the data to construct the learning model. Aim to use the model on mobile phones, lightweight, flexible, easy to transplant, high accuracy are required.

Mainly used in the present study is almost the traditional algorithm of classification algorithms, such as support vector machines, decision tree, random forests, Bayesian and neural network, etc. The classification algorithm's problems are as follows: A. relying on prior knowledge, not pay much attention on the characteristics of stream data; B. some of the algorithms' construction phase is overweight, not easy to Deployment on the phone; C. some of the algorithms need the original training data to classify, makes little sense to practical use. Because of these problems, the classifier has a good performance on the training set but bad on using. The fundamental reason of this phenomenon is learning model does not considering the actual use complex situation.

To solve the existing problem of the above algorithm, this paper put forward constructing classification model based on the clustering algorithm, and finally constructs an adaptive learning model to deal with the real environment.

2.4 Human activity state real-time recognition system

After model constructed and tested by the test set, the next job is deploy model on mobile phones. Use mobile phone to build a real-time human motion recognition system, which is used in real environment to verify the effect, makes the research more practical.

3 Activity state recognition based on clustering algorithm

The research mainly includes two stages: offline phase and online phase. In the offline phase, the main task is to build a learning model. Through the study of labeled data clustering, finally get a learning model based on cluster of purity, this phase is completed on the PC. In the online phase, the principle task is to design movement recognition model on mobile phone, identify new data movement category and regulate the learning model by itself.

3.1 Prepare data

In this research, the data is collected by the triaxial accelerometer, and each sample point contains three float values, representing the acceleration in three directions. Because the model will be deployed on the mobile phones and users carry their mobile phones in different ways, it will cause a swap of the three directions of the acceleration value. Another existing problem of the acceleration sensor is that it is affected by gravity acceleration, so the data collected does not response the true acceleration of the body. These above two questions need to be solved before applying data. Solution is as follows:

(1) using the first-order low-pass filter to calculate the gravitational acceleration component

$$g_j(i) = \alpha g_j(i-1) + (1-\alpha)a_j(i) \qquad (3\text{-}1)$$

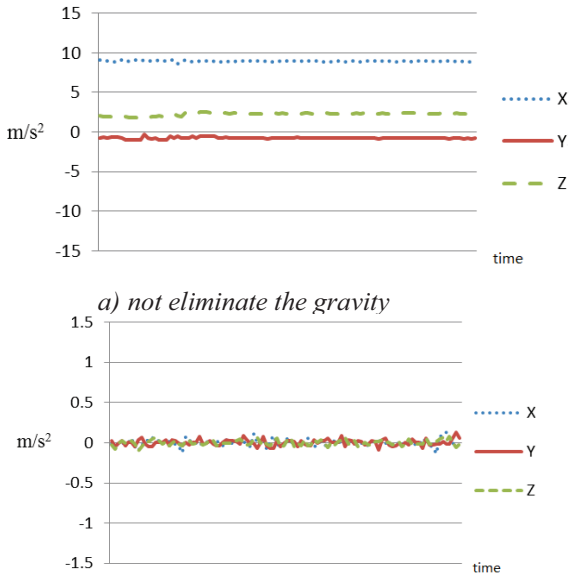
Which α is the filter parameters, g_j is the j-th acceleration component, $g_j(0) = \alpha_j(0)$. α is typically determined by the experimental results($0 < \alpha < 1$), it will be adjusted during the experiment, and the initial value is set 0.8 (from Android document Suggestions). In addition to the set value, it can also be calculated as the following formula Android documentation Suggestions. In addition to directly set can be controlled by the following formula:

$$\alpha = \frac{t}{t + dT} \qquad (3\text{-}2)$$

The t means the time constant of low pass filter; dT means frequency of sampling frequency.

After using the above formula to obtain the gravitational acceleration component on the X, Y, Z-axis, use the original value minus the gravitational acceleration component, and the result is the real motion acceleration value.

Formula (3-1), (3-2) from the Android official document sample code. In order to verify the effect of the formula, use a set of sitting data from a user to analyze. If the formula effectively, then before using the formula, the acceleration influenced by the acceleration of gravity does not equal to zero, and after using the formula processing, three values should be approximately equal to 0, and the result is as figure 1.



a) not eliminate the gravity

b) eliminate the gravity

Figure 1 the influence of gravity acceleration diagram

(2) increasing dimension to reduce the influence of the axes swaps.

There are three methods to reduce the influence of the axes swapping:

(a) By calculating the summation acceleration to reduce the influence of swapping. Transform the 3-d data points p (x, y, z)

into one dimension point p' by the formula $a=\sqrt{x^2+y^2+z^2}$. There is a serious problem that data loss serious by this way, because it is completely ignore the relationship between the three axes, makes the result not very satisfactory.

(b) Calculate the combined acceleration belongs X axis and Z axis, Y axis and Z axis respectively, so the data's dimension reduce one. This can reduce some influence of the axes swapping and have a good accuracy.

(c) Adding three gravity components and three true acceleration components of body, the data will increase to nine dimensions. By this way can reduce the influence of the axes swapping and do not lose any information.

3.2 Clustering stage

After data preparation is done, the next job is to use clustering algorithm to the training data. The clustering method of this paper is using the sample point without using windows partition method. Each point is a sample. Because using the original points to cluster, the scale of the problem will become very large, clustering process will take a long time. K-Means clustering method is chosen because its principle is easy, and running time is short, the number of clusters can be artificially set. K - Means cluster's goal is high cohesion and low coupling. Clustering results evaluated by the following function:

$$y=\frac{\sum_{i=1}^{k} error(C_i)}{n} \tag{3-3}$$

n represents the total number of samples, C_i represent the i-th cluster. Error function is as follows:

$$error(C_i)=\sum_{p\in C_i} dist(p,c_i) \tag{3-4}$$

c_i is the center of C_i, dist(p,c_i) is the distance between p and the center of the i-th cluster. Euclidean distance is used here.

The number of clusters should be determined before using K-Means clustering. This paper uses the inflection method. The relationship of error and k is shown as figure 3-2.

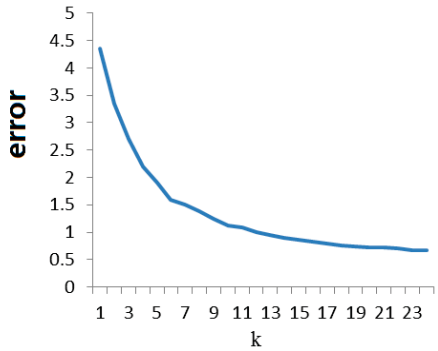


Figure 2 average error VS clusters number k curve

As the figure 2 shows that, the inflection appears when k is 5. So choosing 5 to be the number of the cluster at first is suggested. Then test other values around 5, finally choose the k which can get the best experiment result.

After the cluster number k was determined, use k - Means to cluster the data and the parameter settings are as follows: the maximum number of iterations is 2000, the error rate e is 10^{-6}. Because the initial center of the cluster influence the experiment seriously, so several experiment is necessary. In this paper cluster 20 times and finally choose the best one from the results. The evaluation criteria is the result of the recognition.

3.3 Extract learning model

After clustering is completed, the classification model should be built in accordance with the clustering result. And finally set classification strategy. The extracted model is shown below:

$$LM=\{C_1,C_2,\cdots,C_k\} \tag{3-9}$$

C_i is the i-th cluster, k is the number of clusters.

Classification strategy is as follows: firstly, training data is collected from the continuous windows. Window size should be bigger than a cycle time of an operation of an activity. In this paper, 2 seconds is chosen. In addition, the step length is 1 second so that the current window overlaps 50% with the last one. Secondly, for each sample point in the current window, calculate the dist(p,c_i) and i is range of 1 to k, and according to the distance from each cluster to find the nearest cluster. Then get the confidence rates of the point belong to every activities according to the distribution of the nearest cluster. Finally, calculate the sum of the confidence of every point, and get the largest confidence rate p. If p is larger than 0.6, believe the predictions are reliable. In order to identify the noise data, when a data point is coming, if the distance between the point and the nearest center of the clusters is 1.5 times larger than the radius of the cluster, then believe the data point is a noise point and discard it. Noise point determination conditions can be adjusted according to demand.

3.4 Design human activity real-time recognition system

This part belongs to the online phase, at this phase will use the result of the offline phase. The model will be used on android smartphone to complete activity recognition, in the actual dynamic environment using the model. Activity recognition system based on Android platform mainly includes the following modules: data collection module, data pretreatment module, motion recognition module, model-updating module and data display module.

3.4.1 Data collection:

There are two ways to collect data: mobile phones and hand ring. Because the Android phones and smart wristbands have accelerometer inside, so they can collect the motion data of human. Because people's hands perform complex actions every day and it cannot represent the true acceleration of the body. Besides, the mobile phone against the user's body generally all the day. In order to get a better effect, the

classification is designed for phones. The collected data contains four values, the first one is time stamp, and the others are the accelerations in x, y and z directions.

X, y and z directions in the real three-dimensional space are shown in figure 3:

Data collection process is summarized as follows: when the first time of the program starts, a service will automatic start, this service has two threads: one thread collects the acceleration data from the sensor and saves them into an array blocking queue. The second thread fetches these data from the queue at a frequency of 20 Hz. So there is no data overflow.

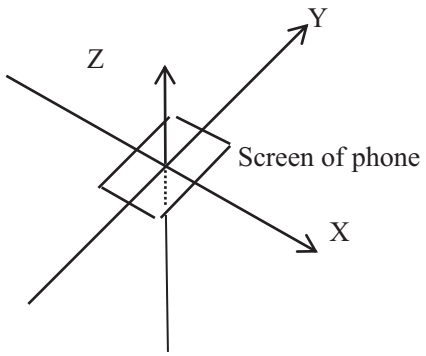


Figure3 triaxial acceleration in the space diagram

3.4.2 Data Pretreatment:

The main function of this module is the raw data processing and formatting. The method contains the elimination of gravity component, increase the data dimension, and de-noising.

3.4.3 Activity Recognition:

In the offline phase, the paper uses a clustering method to build classification model, in the online phase, use the model to build an activity recognition system.

3.4.4 Model Updating:

In the offline phase, each classification model can achieve a good result, and have a good performance on the training data. Because most of the data is detached, the model can fit the data very well. But when it is used in daily life, the result is not satisfactory. Reasons are mainly as follows: A. the actual data is continuous, and it change every time, as the time float, it will cause "concept drift" and the existing model may be not accurate; B. personalized problem, activity action of everyone have difference more or less, so the acceleration is different too, the model cannot fit every user; C. the training data is collected in an experiment environment, but the testing environment is much more complex, so the data is not so standard as the training data. The model must be adjusted by itself with the testing data.

In this paper proved the single sample updated is effective. The update formula is as follows:

$$centroid_{new} = \frac{(n-1)centroid_{old} + data_{new}}{n} \quad (3\text{-}10)$$

Among them, n is amount of the point in the cluster which will to be update, $centroid_{old}$ is of old center of the cluster and $centroid_{new}$ is center of cluster after the update, $data_{new}$ is the point to update the cluster.

Distribution of the clusters must be update as well; the method is similar as update the center.

4 results of experiments and analysis

4.1 Experiment environment and experiment data sets

4.1.1 Experiment Environment

This subject is mainly divided into two parts, the offline and online phases. Offline stage experiment is done on PC with the OS of win-x64. The online stage is design for android platform, and the testing environment is MI 4L TE-CMCC.

4.1.2 Experimental Data

In the offline stage, the model is built according to the data set from the United States Fordham university WISDM (Wireless Sensor Data Mining) laboratory. The data contain acceleration data and GPS data from 36 volunteers. The data set is used by many researches and has been proved to be effective. This topic is based on the acceleration data, so this environment uses the WISDM latest acceleration data set to build the classification model. The data set 's collection process is as follows: in the experimental environment, 36 volunteers carry the Android mobile phone with the same Android app and perform the given six activity include jogging, walking, up-stairs, down-stairs, sitting and standing to collect the acceleration data so that all data is labeled. The data set is collected in December 2014.

4.2 Experiment Plan

To demonstrate the effectiveness of the clustering method, first use part of the WISDM data to construction cluster model, and then use the other data and the real data collected by 5 volunteers in our laboratory to test the model. Results evaluated by two indicators precision and recall. Experiments final calculates the accuracy to measure the performance of the classification model.

One window of test data contains 40 points, so to the devices whose sampling rate is 20 Hz, the window size is 2 seconds. The moving step length is half of the window size, so that the current window overlaps 50% in the last one. Because the size of one test data is one window. So the window size will affect the performance of the system. So far, no one can prove how much the window size is can achieve the best results. The window size can only be selected by experimental results.

This paper mainly contains the following experiments: clustering method validation, comparison between several clustering method and several traditional method, personalized experimental accuracy and comparisons with mature products.

4.3 Result of the experiment and analysis

In order to verify the effectiveness of activity recognition model based on clustering algorithm, this paper tries to use K - Means algorithm to train data, and uses the cluster result to build the classification model.

Experiments using the k-fold cross validation to verify the validity of the classification model, and let k be 5. Firstly, cluster the raw data directly without any processing, the number of cluster is 10, and the result is shown in table 1.

Table 1 classification results with raw data

True / Predicted	Walking	Jogging	Sitting	Standing	Upstairs	Downstairs	Precision
Walking	1608	679	0	0	733	758	0.4256
Jogging	632	1009	0	0	103	113	0.5433
Sitting	0	0	2529	506	0	0	**0.8333**
Standing	20	0	511	559	0	0	0.5128
Upstairs	90	0	0	0	37	23	0.2467
Downstairs	71	0	0	0	2	14	0.1609
Recall	0.6642	0.6049	**0.8319**	0.5249	0.0418	0.0154	—

The results shows that the classification model based on the raw data has a poor performance. Only to "sitting" data can be distinguished out. The clustering results show that many different types of data are overlapped together. It cause most of the cluster has an evenly distribution, they are mixed with several acvitity data. So this paper use the low-pass filter to separate the acceleration of gravity, and increase the original 3-d data into 9-d, the form is shown as follow:

$$(x, y, z) \rightarrow (x, y, z, g_x, g_y, g_z, x - g_x, y - g_y, z - g_z)$$

After processing the original data, use the new cluser to build the classification model and test it. The reslut is shown in table 2.

Table 2 classification results after data processing

Tru / e Predicted	Walking	Jogging	Sitting	Standing	Upstairs	Downstairs	Precision
Walking	2301	109	0	0	805	845	0.5681
Jogging	0	1579	1	0	45	55	0.9398
Sitting	0	0	2911	121	0	0	0.9601
Standing	0	0	129	944	36	0	0.8512
Upstairs	0	0	0	0	0	0	0
Downstairs	0	0	0	0	0	20	1.0000
Recall	1.0000	0.9354	0.9573	0.8864	0.0000	0.0220	—

The result shows that the classification accuracy of "jogging", "sit" and "standing" is high, the precision rate and recall rate are above 85%, classifier recognition accuracy is 0.8195. But the classification can't distinguish the "upstairs" data and "downstairs" data, they are considered as "walk" data, and it causes the reducing of the accuracy.

Research shows that these considered to be "walking" data which not belong to "walking" have a highest confidence rate determined to be "walking", and also have a second and third highest confidence rate determined to be "upstaris" and "downstairs". The real indistinguishable data is "upstaris" and "downstairs". Since they overlap serious, single cluster classifier can't effectively distinguish them. So we merge them into one class called "stairs". The experimental results is shown in table 3.

Table 3 experimental results after mergeing the stairs data

True / Predicted	Walking	Jogging	Sitting	Standing	Stairs	Precision
Walking	1899	89	0	0	322	0.8221
Jogging	0	1583	3	0	35	0.9765
Sitting	0	0	2914	128	0	0.9579
Standing	201	0	126	937	36	0.8525
Stairs	321	0	0	0	1410	0.8145
Recall	0.7844	0.9468	0.9576	0.8798	0.7820	—

Table 3 shows that after merging the data, the classification accuracy is 0.8740, and for each kind of activity the model has a good classification effect.

In order to further verify the effectiveness of the clustering method, the experiment compare the K - Means, DBSCAN and gaussian mixture model (GMM) three kinds of clustering algorithm with SVM, random forest (RF), naive bayesian classification (NaiveBayes) three activity recognition classification algorithm. And the algorithm parameters are set as follows: A. K - Means algorithm: the number of cluster K is 10, The maximum number of iterations is 2000, the number of error threshold value is 10^{-6}, and run 20 times to get the optimal result; B. DBSCAN: the radius is 0.5 and the minimum number of a cluster is 50 samples; C.gaussian mixture model: the number of gaussian model is 10 and the number of iterations is 1000; D. the other three classification methods use the WeKa software[14] directly, and use the default parameters. Experimental results is shown in table 4-4 below.

Table 4 comparation between the clustering method and the traditional classification method

Algorithm	Walking	Jogging	Sitting	Standing	Stairs	All	Time
K-Means	0.8221	0.9765	0.9579	0.8525	0.8145	0.8740	7.03s
DBSCAN	0.5889	0.6043	0.7785	0.5561	0.4473	0.5229	21.21s
GMM	0.8007	0.9805	0.9488	0.8339	0.8037	0.8604	44.17s
SVM	0.9610	0.9743.	0.9640	0.9391	0.4869	0.8536	3.02s
RF	**0.9840**	**0.9872**	**0.9593**	**0.9448**	**0.8284**	**0.9443**	4.53s
NaiveBayes	0.8971	0.9340	0.8210	0.7572	0.3041	0.7426	0.12s

The table 4 shows that the result of the classification by using the six algorithms. We can find that the accuracy of them is very close, except the DBSCAN cluster model. But the cluster classifier is light weight, it only contains some cluster features and is not need to save the original data. And the biggest advantage of the cluster method is that it can adjust itself easily but the traditional classification method can't.

Although the RF have the best performance, but the cluster is the most suitable algorithm.

Ideally, the project hope to build a general classifier that everyone useing it will get a accurate results. But in fact everyone's sports movement is very different, so it is nearly impossible. The follow experiment is using one person's data to train the classifer and use it on others, and the result is shown in table 5.

Table 5 shows that the model can not have a good performance on other users. The third user gets a good result,

but other's is poor. Such as the user 9, the overall classification accuracy is only 0.4626. Personalization is in conflict with applicability, the only way is to let the model adjust itself by user's data. The next experiment is using the training set composed by all user. Each user's data is as long as 2 minutes,and the result is shown in table 6

Table 5 model of personalized classification accuracy results

User	*Walking*	*Jogging*	*Sitting*	*Standing*	*Stairs*	*All*
1	0.6335	0.7705	0.5576	0.5535	0.4133	0.5859
2	0.6773	0.5066	0.7443	0.6991	0.5773	0.6409
3	0.8223	0.8865	0.9078	0.8559	0.7768	0.8498
4	0.4633	0.5665.	0.4645	0.5388	0.4970	0.5060
5	0.5991	0.6977	0.7588	0.6459	0.7557	0.6914
6	0.6961	0.5359	0.6322	0.6568	0.5057	0.6053
7	0.7988	0.7447	0.7655	0.7445	0.6227	0.7358
8	0.8869	0.7366	0.8287	0.7572	0.7781	0.7975
9	0.4999	0.5339	0.5205	0.4546	0.3041	0.4626
10	0.3971	0.4340	0.5229	0.6554	0.4472	0.4913

Table 6 use multiple user data to construct the model

True / *Predicted*	*Walking*	*Jogging*	*Sitting*	*Standing*	*Stairs*	*Precision*
Walking	12488	2226	0	556	2066	0.7204
Jogging	3237	13043	0	8	1090	0.7505
Sitting	0	0	14447	2628	0	0.8461
Standing	701	0	2833	14088	3036	0.6820
Stairs	854	2011	0	0	11088	0.7947
Recall	0.72269	0.7548	0.8361	0.8152	0.6417	—

As the table 6 and table 3 shows, when use the multiple data, the result is worse than using one user's data. However, compared with the table 6 and 5, the previous is better.

So the model should be trained by the multiple data at first. Although the initial accuracy is not very high, but it will not be too low, so it can adjust itself when it is used. If the initial accuracy is too low, the adjustment results will be worse.

5 Conclusions

This paper explores how to use clustering methods to build a learning model and use it for human motion recognition, and eventually deploy the model on the android phone. The classifier uses the existing data to train the learning model, due to personalize reason, the accuracy of the system will be low at the start. But over time, the model is adjusting itself every time, so the accuracy will increase.

A large number of studies have shown that, through some data processing method, it can make the accuracy of simple human activity recognition reach 90%. But it can't be used in daily life. The reasons are following: A, the using environment is complex; B, personalized problem, the same kind of activity belongs to different users have a differect acceleration. C. Some actions are inseparable or boundaries are not clear.

The research results are as follows:

(1) Prove the effectiveness of using a cluster model to classify the human activity.
(2) Giving a suitable clustering algorithm model updating method on human activity recognition

The further research is to solve these several problems mentioned in this paper. One is how to process the data so that it can reflect the ture body acceleration when the phone has a slight shaking. The other is to find a better method to solve the personalized problem, and find a find a general model that can easy fit all ordinary person.

6 References

[1] Kim E, Helal S, Cook D. Human Activity Recognition and Pattern Discovery[J]. IEEE Pervasive Computing, 2010, 9(1): 48-53.

[2] Van Kasteren T, Noulas A, Englebienne G, et al. Accurate Activity Recognition in a Home Setting[C]//Proceedings of the 10th International Conference on Ubiquitous Computing. ACM, 2008: 1-9.

[3] Siirtola P, Juha Röning. Recognizing Human Activities User-independently on Smartphones Based on Accelerometer Data[J]. International Journal of Interactive Multimedia & Artificial Intelligence, 2012, 1(5): 38-45.

[4] Box G E P, Jenkins G M. Time Series Analysis Forecasting and Control[J]. Journal of Time, 1970, 3(2): 199-201.

[5] Widmer G, Kubat M. Learning in the Presence of Concept Drift and Hidden Contexts[J]. Machine Learning, 1996, 23(1): 69-101.

[6] Fang H, Long C, Srinivasan R. Influence of Time and Length Size Feature Selections for Human Activity Sequences Recognition[J]. ISA Transactions, 2014, 53(1): 134–140.

[7] Ordóñez F J, Iglesias J A, De Toledo P, et al. Online Activity Recognition Using Evolving Classifiers[J]. Expert Systems with Applications, 2013, 40(4): 1248-1255.

[8] Bhattacharya S, Nurmi P, Hammerla N, et al. Using Unlabeled Data In a Sparse-coding Framework For Human Activity Recognition[J]. Pervasive and Mobile Computing, 2014, 15: 242-262.

[9] Kwapisz J R, Weiss G M, Moore S A. Activity Recognition Using Cell Phone Accelerometers[J]. ACM SigKDD Explorations Newsletter, 2011, 12(2): 74-82.

[10] Lane N D, Xu Y, Lu H, et al. Enabling Large-scale Human Activity Inference on Smartphones Using Community Similarity Networks (CSN)[C]//Proceedings of the 13th International Conference on Ubiquitous Computing. ACM, 2011: 355-364.

[11] Park J G, Patel A, Curtis D, et al. Online Pose Classification and Walking Speed Estimation Using Handheld Devices[C]//Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM, 2012: 113-122.

SESSION

HEALTH INFORMATICS AND RELATED ISSUES

Chair(s)

TBA

ISBN: 1-60132-437-5, CSREA Press ©

ISBN: 1-60132-437-5, CSREA Press ©

A Case Study of Nurses Perceptions and Attitude of Electronic Medical Records in Riyadh and Jeddah's Hospitals

Afrah Almutairi [1], Professor Rachel McCrindlel[2]

[1] School of Systems Engineering, University of Reading, Whiteknights, Reading, Berkshire, RG6 6AY, United Kingdom

[2] School of Systems Engineering, University of Reading, Whiteknights, Reading, Berkshire, RG6 6AY, United Kingdom

Abstract - *The purpose of this study was to investigate the perceptions, level of knowledge, attitudes, and behaviour of nurses at the hospitals in Riyadh and Jeddah city towards using electronic medical records. A questionnaire comprised of closed and open questions was distributed online to all participated MOH's hospitals in Riyadh and Jeddah's city. This paper will present the results of the study highlighting key findings in relation to nurses' perceptions, knowledge, and attitudes to EMR with a view to identifying ways to help plan and organise better education and training programs for nurses in order to gain their support for enhanced use of EMR, thereby contributing to the success of the Ministry of Health's (MOH) e-health project in Saudi Arabia.*

Keywords: *E-Health, Electronic Medical Records, Nurses, Health Informatics, Hospitals, Saudi Arabia.*

1 Introduction

Health care in the Kingdom of Saudi Arabia has been developing since 1949, with one such development being the move from handwritten records as the method used to store medical information to the use of electronic health systems in many hospitals and organizations in Saudi Arabia. Health care and medical service in Saudi Arabia are divided into three major sectors (1) Ministry of Health, (2) Other Governmental agencies such as Teaching Hospitals, University Hospitals, Military Hospitals and (3) the private Health Care Sector [1]. The Saudi government and the private sector in Saudi Arabia have invested heavily to build the essential infrastructure required to ensure adequate health care is provides for all people [2, 3, 4, 5]. In doing so, they have established a high level of health care infrastructure and other resources analogous with health care levels in many developed countries [6, 7]. However, the use of EMR is often not centralized or standardized across hospitals and private health organizations, and harmonization between the different health care providers and other associated sectors is required [8]. Research by Gallagher highlights gaps that require further study such as improvements to learning systems and efforts made by hospitals to improve the readiness of staff to use EMR [5]. Al Sheifi suggests that in order to implement successful EMR, the Ministry of Health should give "strong support" to female nurses receiving computer training including search techniques and data entry skills, in preparation for their use of medical records [9].

Saudi government has given high priority to improve health care services at all levels: primary, secondary and tertiary. As a result, health care services in Saudi Arabia have improved but there still numeral of issues make challenges to the health care system, such as barrier, internal and external change, and technological, economical and social factors. This paper presents the case study survey involved 1428 nurses at MOH's hospitals in Riyadh and Jeddah city in Saudi Arabia to assess their view regarding the features of the current system, the benefits and barriers of more complete EMR. The findings show that there is a strong significant relation between years of Prior computer experience and the knowledge or attitude toward EMR P-value <0.000. The reason behind the significance relationship between level of qualification and knowledge or attitude toward EMR is because of the absence of proper teaching or modules that depend on EMR to perform nurses' tasks. This point highlights the need for improving the healthcare system to make it able to adopt with the development in medical recording.

1.1 Objective

Building on a previous pilot study undertaken with female nurses in single clinic at Jeddah, the aim of the PhD is to investigate more widely and in more depth the perceptions, knowledge, and attitudes of nurses at all levels of clinical practice toward the use of Electronic Medical Records in order to identify the main benefits of, and barriers that affect, adoption of EMRs within Riyadh and Jeddah City hospitals. The aim of this study to detect the perception, knowledge, and attitudes of nurses in Riyadh and Jeddah city toward using electronic medical records and how these can be used to build future adoption of EMR in Saudi Arabia.

2 Literature review

Nurses as a part of the medical teams in hospitals need to be targeted by investigating their perceptions, needs, and attitudes toward EMR and by using the information to conduct training programs and management initiatives to improve the commitment of nurses in the transition toward automated medical records. A study conducted at a large Magnet hospital in Southwest Florida used the five-item, Likert-type attitude

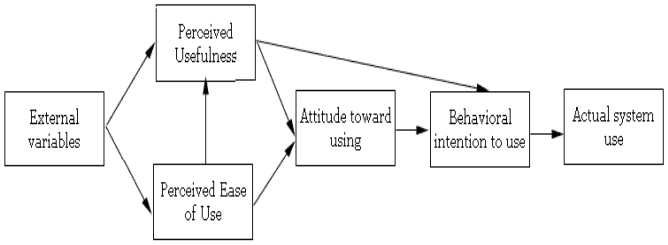
scale to assess 100 nursing personnel preferences, perceptions, and attitudes toward using Electronic Health Records; the study concluded that most of nurses believed that EHRs improved the quality and performance of patients care and nurses with expertise in computers, 80%, had positive attitudes toward EHRS than nurses with less expertise [10]. Another study conducted a survey among nurses to predict nurses willingness to participate and use a new electronic patient record system (EPRS); its results showed that there was an overall positive attitudes toward EPRS and age has a significant effect on determination toward the EPRS [11]. Whittaker et al., explain that the perceived of EHR usefulness was dependent on how the nurse acceptance it. The attitudes positive when a nurse perceived the advantages comparing with negative attitudes if a nurse had lacked on time management skills, poor training and technology[12]. Similarity with study suggested that to accept technology by healthcare professionals its need to be perceived as useful and has ease of use. Other study found that if a nurse had capability to use the IT system nervous tension and workload at work reduced. The study found that there was adiffrent of the size of the stress by gender female nurses reporting less stress and more satisfaction with work than the male nurses. The findings showed that nurses were more positive attitude and less nervous when not using IT or eaither internet which influnced nervous levels [13].

Other study was conducted in primary healthcare centres (PHC) in Al Ain, United Arab Emirates (UAE) by using qualitative study on three focus group interviews among physicians using open-ended questions. In this study Focus group contained of 7–9 physicians working in PHC as family medicine, a mix of males and females of different age groups and professional experience. The finding showed a positive perception of physicians who satisfied with EMR about the application of the system. Their participants stated that they were satisfied with EMR system because it was "fast, easy to use, well documented, more precise and provided patient engagement tools such as the patient education resources and patients' portal" [14].

In Saudi Arabia, however, there is a lack of information regarding the attitudes and perception of nurses toward EMR. This might have occurred due to what is described as "incomplete, inaccurate, unreliable and not timely" of data collection which lead to an ambiguous picture of the potential EMR systems in developing countries. One study in Saudi Arabia investigated the usefulness of EMR system implemented at a teaching hospital in the eastern province of Saudi Arabia [15]. The study surveyed 142 physicians in the hospital and considered confounding factors such as demographic data, physician computer experience. The study assessed the satisfaction of the physicians after the implementation of the EMR system. In Norway, more than 50% of the physicians were dissatisfied especially for those who do not have previous knowledge about computers and lack typing skills since going to different screens to review charts which take longer than reviewing paper records. Physicians preferred utilising paper records than utilising the system for less than half of the daily job [16].

3 Theoretical model of user acceptance

This study useed the Technology Acceptance Model (TAM) to assess the factors that particularly appropriate in the Health Information Technology field since it focuses on two particular variable assumed to effect the use of information technology. Perceived usefulness is the factor that signified the degree that the person trusts the IS which will evalute them in the performance of their job. Also, Perceived ease of use is the second factor that used to show how hard the person trusts the planned system would be to use. A review of the literature demonstrates few studies in the health information field which are used the TAM dealing with a large range of information technologies which found that person's behavioral intention is determined by the person's attitude. The Davis' model 1989 version (Figure 1), adopted to expect and clarify users' "acceptance and rejection" of computer technology [17].



Figuer.1 Technology acceptance model (Davis et al, 1989)

4 Research methods and Hypothesis

The researcher choose a mixed methods approach such as quantitative and qualitative methods to explore more about nurses perception, attitude and knowledge towards the use of EMR but this research primarily quantitative as it seek to evaluate the Health Information system between the nurses and EMR. It's let to test and identify the Hypotheses. To collect and analyse data the research design is used as a guideline to prepare the study [18]. The researcher used different technique when collecting the data, for example, distribute the questionnaire, interview with the most experience senior nurses and focus group with different group of nurses such as male female, Saudi and non Saudi, different ages to allow test different variable. In addition, after reviewed relevant studies, this research proposes three external variables: General Information, Professional Factors and Organizational Factors. The researcher trusts that the proposed external variables moderate the original TAM variables. Therefore, the following is null hypothises: 1) NH1: perceived ease of use affected negatively on perceived usefulness of the use of EMR.2) NH 2: perceived ease of use affected negatively on attitude towards the use of EMR.3) NH3: perceived usefulness affected negatively on Self Efficacy of the use of EMR.4) NH4: perceived ease of use affected negatively on Self-Efficacy of the use of EMR.5)

NH5: perceived ease of use affected negatively on Perceived Behavioural Control of the use of EMR.6) NH6: perceived usefulness affected negatively on Perceived Behavioural Control of the use of EMR.7) NH7: perceived usefulness affected negatively on Sufficient Training of the use of EMR.8) NH8: perceived ease of use affected negatively on Sufficient Training of the use of EMR. 9) NH9: general information affected negatively on attitude towards the use of EMR .

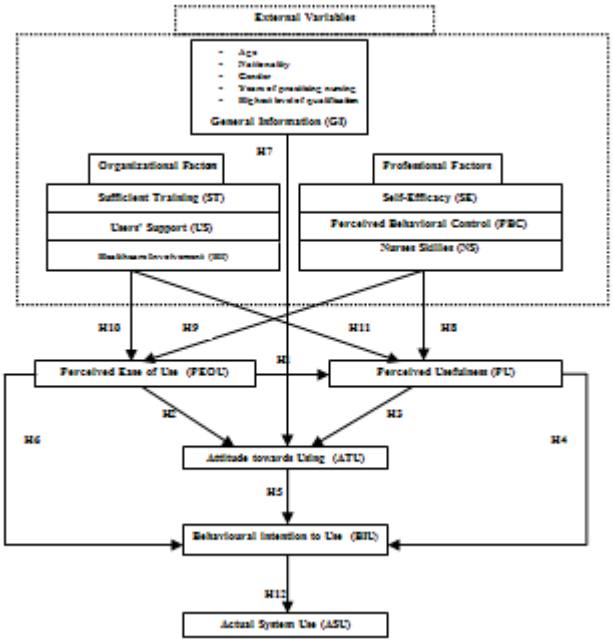


Figure 2 : Technology Acceptance Model for this study

5 Study methods

The study was conducted all nurses at the MOH's hospitals in Riyadh and Jeddah's City. Questionnaires were used as the primary research methodology in this study. The questionnaire used in this study was modified from the original David's measurement scales used in TAM and from other literatures by changing some wording and validation to fit the context of the use of EMR to make sure content validity. The questionnaire was written in both English and Arabic to take into account nurses' nationality and academic backgrounds. SPSS statically was applied to assess group differences across different variables. Participants – nurses at the MOH's hospitals in Riyadh and Jeddah's City – were asked closed questions about EMRs as well as being encouraged to write about their knowledge, perceptions and attitudes towards the use of EMR through three open-ended questions. The researcher designed a questionnaire 64-item. The questionnaire divided in to seven parts. The first section included questions about general information for nurse such as Age, Nationality, Gender, years of practicing nursing numbers of a years from "1 – 4" to "15 or more" *and* highest level of educational such as Licensed practical Nurse, Associates degree in Nursing, Bachelor of Science in Nursing, Master of Science in Nursing, Master of Science Non- Nursing, Doctorate Nursing and Doctorate Non-Nursing. The second section contained questions about computer literacy such as comfort with technology using a 5-point Likert scale ranging from "very uncomfortable" to "very comfortable", having personal computer or laptop, having computer in their office, spending time on using computer, having computer skills using "Yes" and "No" options and *years* of Prior computer *experience* numbers of a years from "1 – 4" to "15 or more". The third section included questions about nurses 'knowledge and perceptions of EMR contained a table of their perceptions regarding statements about EMR. These statements are: Paper-based are more credible than EMR, EMR require special training, EMR add a burden to nurses workloads, EMR are worth the time and effort required to use them, EMR will decrease productivity, EMR enable services such as access structured historic patient information to be efficiently provided, EMR mean that requested records are always available, EMR mean that requested records are delivered promptly, EMR improve communication between medical and nursing staff in hospitals, EMR enable medical staff to be cooperative and responsive to patients needs, There are concerns with the confidentiality of EMR, Staff is highly trained and knowledgeable about EMR, EMR documents/files are complete and well-organized, The format of EMR is highly acceptable, EMR documents are available in a timely manner to all authorized users, There are currently issues with EMR meeting international standards, EMR are Properly arranged, EMR works to reduce human errors, EMR improve patient safety and quality of care, EMR work well in practice, EMR works to facilitate the completion of the work, EMR assist patient data entry, EMR enable access to medical records from different places in the hospital, EMR assist medical staff to make the right decision in the care of patients, The use of EMR may lead to the loss of patient information because of technical errors. EMR will help in building a database of national health care using a 5-point Likert scale ranging from "strongly agree" to "strongly disagree".

The forth section contained questions about using EMR such as percentage of time spend when dealing with patient records at work using from "0%", "75-100%", recording and accessing clinical documentation using 100% as paper based records, 100% as Electronic Medical Records, Mostly paper based records, Mostly EMR, Approximately 50:50 paper based records and EMR and switch from paper=based records to EMR has been a positive experience overall using "Yes" and "No" options. The fifth section contained questions about nurses' satisfaction with EMR such as system provide the precise information, the information content meet their needs, provide sufficient information, the output is presented in a useful format, the information clear, easy to use, get the information on time and provide up-to-date information using a 4-point Likert scale ranging from "Never" to "Always" and their satisfaction with EMR in their department using a 5-point Likert scale ranging from "excellent" to "poor." EMR make it easier to review patients' problems, EMR make it easier to

find specific information from patient records and EMR can produce data reviews for specific patient groups, e.g. complication rate, diagnoses using "Yes" and "No" options. Select the problems that you have had with EMR use such as Downtime, Limits communication with other health care team members, Decrease amount of time spent with patient, Accessibility to computers, Accessibility of patient information, Speed of Log-in, No problems noted, Other. The sixth section included questions about global assessment about EMR in their department such as the performance of department, the performance of own tasks and the quality of the department using 7-point Likert scale ranging from "difficult" to "easier". The seventh section included questions about using EMR such as the ability to use EMR, EMR received any training in the use of EMR using "Yes" and "No" options, training received and for how long, ranking factors that might help to adapt with a new EMR in a hospital according to their importance to them. These factors are "training courses outside the hospital", "training courses inside the hospital", "and colleagues at the hospitals", "personal experience, other". The final items (62, 63, 64) was open-ended questions inviting a written response about the barriers and the benefits for applying effective EMR in hospitals and any issues the nurses felt were not adequately addressed by the questionnaire.

5.1 Demographics of Study Population

The reason to carried out this study to investigate nurses perceptions, knowledges, attitudes about the EMR system. The researcher contacted MOH's hospitals regarding to fill in the questionnaire and 21 Health care professionals responded (table 1). Some of participants have different backgrounds which are Religion, Geography and Art graduates with Licensed practical Nurse. The highst of them qualified Bachelor of Science in Nursing and only one Doctorate Non-Nursing.

Table :1 Data resourses used in this research

Items	Data resources	Number of return questionnaire	Location
Pilot study	Female Participants	230	Single clinic at Jeddah
	Questionnaire		
	Excel for analysis		
Case study	Different group of participants	1428	20 hospitals in Riyadh and Jeddah city
	Questionnaire		
	SPSS		
	Interview	3 Executive Director of Nursing	1 hospital in Riyadh and 2 hospitals at jeddah city
	Focus group	4 nurses at Riyadh hospital + 5 nurses at Jeddah hospital	2 hospitals in riyadh and jeddah city

5.2 Participants

The participants in this study were 1428 nurses from different hospitals and diffeent departments who willingly participated in the online survey +12 nurses willingly participated in the interviews and focus groups in Riyadh and Jeddah city. All participants in this study were member of private and public hospital working MOH, who suitable for the purpose and context of this study.

5.3 Instrumentation

The research instrument consists of seven main sections. The first section includes a nominal scale to classify particepants' demographic information. The second section containes computer Literacy such as skills and computer experience. The third and fourth section uses 5-point Likert scale where 5: Strongly disagree, 4: disagree, 3: don't know, 2: agree, 1: Strongly agree. The fifth section includes the use of EMR. The sixth section includes the overall Satisfaction of EMR. Finally, The seventh section includes training recived with EMR. The last sex sections includes TAM constructs.

5.4 Demographic characteristics

This section of the questioner defines particepantes' demographic characteristics. It includes 5 items such as age, gender, nationality, years of practicing nursing and highest level of qualification (table 2).

Table 2 : Questionnaire section1

Section 1: General information (GI)

1- Age :
□ 20-30 yrs □ 31-40 yrs □ 41-50 yrs □ > 50 yrs

2- Nationality: □ Saudi □ Non Saudi

3- Gender: □ Male □ Female

4- How many years do you practicing nursing?

☐ 1-4 yrs
☐ 5-9 yrs
☐ 10-14 yrs
☐ <15 yrs

5- Which of the following indicates your highest level of qualification preparation?

☐ Licensed practical Nurse.
☐ Associates degree in Nursing.
☐ Bachelor of Science in Nursing.
☐ Master of Science in Nursing.
☐ Master of Science Non- Nursing.
☐ Doctorate Nursing.
☐ Doctorate Non-Nursing.

5.5 Measuring TAM constructs

The second, third, fourth, fifth, fixth and feventh sections of the survey (Table3), as mentioned in the questionnaire design above, measures TAM constructs ulitised in this study. As shown in table 3, there are 59 items measured in accordance with the current study's research model. The measured items include Self-Efficacy (SE) (6 items), perceived ease of use (PEOU) (17items),perceived usefulness(PU) (16 items), attitude toward usage(ATU) (3 items), Perceived Behavioural Control (PBC) (14 items), and Sufficient Training as an external factor (ST) (3items).

Table 3 : Questionnaire sections 2, 3, 4, 5,6 and 7

Section 2: Self-Efficacy (SE)	
Computer Literacy:	
How would you rank your comfort with technology?	SE1
Do you have your own personal computer or laptop?	SE2
Do you have Computer in your office or word?	SE3
Do you spend time on using the computer and the internet every week?	SE4
Do you have computer skills?	SE5
How many years of Prior computer experience do you have?	SE6
Section 3: Perceived Ease of Use (PEOU)	
Knowledge and Perceptions of EMR:	
EMR require special training.	PEOU1
EMR enable services such as access structured historic patient information to be efficiently provided.	PEOU2
EMR mean that requested record are always available.	PEOU3
EMR mean that requested record are delivered promptly.	PEOU4
There are concerns with the confidentiality of EMR.	PEOU5
Staffs are highly trained and knowledgeable about EMR.	PEOU6
EMR documents files are complete and well-organized.	PEOU7
The format of EMR is highly acceptable.	PEOU8
EMR documents are available in a timely manner to all authorized users.	PEOU9
There are currently issues with EMR meeting international standards.	PEOU10
EMR are Properly arranged.	PEOU11
EMR enable access to medical records from different places in the hospital.	PEOU12
The use of EMR may lead to the loss of patient information because of technical errors.	PEOU13
Do you think you have the ability to use EMR effectively?	PEOU14
Please write below what you believe are the most significant benefits of adopting effective electronic medical records in hospitals	PEOU15
Write below what you believe are the most significant barriers or challenges for adopting effective electronic medical records in hospitals:	PEOU16
Are there any issues you feel they were not adequately addressed by the questionnaire?	PEOU17
Section 4: Perceived Usefulness (PU)	
Knowledge and Perceptions of EMR:	
Paper-based are more credible than EMR.	PU1
EMR add a burden to nurses' workloads.	PU2
EMR are worth the time and effort required to use them.	PU3
EMR will decrease productivity.	PU4
EMR improve communication between medical and nursing staff in hospitals.	PU5
EMR enable medical staff to be cooperative and responsive to patients needs.	PU6
EMR works to reduce human errors.	PU7
EMR improve patient safety and quality of care.	PU8
EMR work well in practice.	PU9
EMR works to facilitate the completion of the work.	PU10
EMR assist patient data entry.	PU11
EMR assist medical staff to make the right decision in the care of patients.	PU12
EMR will help in building a database of national health care.	PU13
The performance of our department's work has become	PU14
The performance of my own tasks has become	PU15
The quality of our department's work has become	PU16
Section 5: Attitude towards Using (ATU)	
The Use of EMR:	
Do you think whether the switch from paper=based records to EMR has been a positive experience overall?	ATU1
How do you currently record and access your clinical documentation?	ATU2
What percent of your time do you spend when you dealing with patient records at work?	ATU3
Section 6: Perceived Behavioural Control (PBC)	
The overall Satisfaction of EMR:	
How often does the EMR system provide the precise information you need?	PBC1
How often does the information contained in EMR meet your needs?	PBC2
How often does the EMR system provide reports that exactly meet your need?	PBC3
How often does the EMR system provide sufficient information?	PBC4
How often do you think the EMR output is presented in a useful format?	PBC5
How often is the EMR information clear?	PBC6
How often is the EMR system easy to learn how to use?	PBC7
How often does the EMR system deliver the information you need in a timely manner?	PBC8
How often does the EMR system provide up-to-date information?	PBC9
How would you rate your satisfaction with EMR in your department?	PBC10
Do you think EMR make it easier to review patients' problems?	PBC11
Do you think EMR make it easier to find specific information from patient record?	PBC12
Do you think EMR can produce data reviews for specific patient groups, e.g. complication rate, diagnoses?	PBC13
What are problems that you have had with EMR use?	PBC14
Section 7: Sufficient Training (ST)	
Have you received any training in the use of EMR?	ST1
Please write what training have you received and for how long?	ST2
Please tick, according to their importance to you, the following factors that might help you to adapt to the introduction of a new EMR system in a hospital?	ST3

6 Data analysis and results

6.1 Demographics

The majority of participants were between 20 and 30 years, with 25.56% from 31 to 40, 24.39% from 41 to 50, and 11.69% above 50. The female participants were almost more than male in term of gender, with 246 (17.23%) males and 1182 (82.77%) females. Saudi-nationality nurses the highest response rate, at 79.55%.The majority of participants practicing nursing were have 5-9 years, with 29.48% from 10-14 years, 16.39% from 1-4 years and with a low minority (13.80%) above 50. The majority of participants were have BS in nursing in term of highest level of qualification, with 36.83% have LP nurses, 11.55% have MS non nursing, 2.73% have Associates degree in nursing, 2.38% have MS in nursing, 1.12% have Doctorate nursing and with a low minority (0.07%) have Doctorate non nursing.(see table 4).

Table 4: Respondents' demographics information

Respondents	Frequency	Percent
Age		
20 -30 yrs	540	37.82%
31-40 yrs	365	25.56%
41-50 yrs	356	24.93%
>50 yrs	167	11.69%
Total	1428	100.00%
Gender		
Female	1182	82.77%
Male	246	17.23%
Total	1428	100.00%
Nationality		
Non Saudi	292	20.45%
Saudi	1136	79.55%
Total	1428	100.00%
Years of practicing nursing		
1-4 yrs	234	16.39%
5-9 yrs	576	40.34%
10-14 yrs	421	29.48%
>15 yrs	197	13.80%
Total	1428	100.00%
Highest level of qualification		
Associates degree in Nursing	39	2.73%
Bachelor of Science in Nursing	647	45.31%
Doctorate Non-Nursing	1	0.07%
Doctorate Nursing	16	1.12%
Licensed practical Nurse	526	36.83%
Master of Science in Nursing	34	2.38%
Master of Science Non-Nursing	165	11.55%
Total	1428	100.00%

6.2 Validity and reliability

The reliability is important for the researcher which protects his data and subject from repeat it by other study.

The reliability is an ability of the research creates consistent results (Sarantakos 1998).

Table 5: Instrument reliability Cronbach alpha

Scale	Number of Items	Cronbach Alpha
Self-Efficacy (SE)	6	-1.370 a
Perceived Ease of Use (PEOU)	17	0.978
Perceived Usefulness (PU)	16	0.750
Attitude towards Using (ATU)	3	0.094
Perceived Behavioural Control (PBC)	14	0.989
Sufficient Training (ST)	3	0.959
Overall reliability	59	0.754

The overall Cronbach's alpha reliability of the questionnaire items is 0. 754 (see table 5) and this value coefficient is considered as a high and acceptable. All measures for PU and PEOU in this study show a high level of reliability, ranging from 0.750 to 0.978. All Cronbach Alpha value > 0.70, and therefore the survey is considered reliable. Moreover, some of respondents were had plenty of knowledge and computer experience to respond to the entire questionnaire items. In this study, the reliability assessment was done using Statistical Package for Social Sciences (SPSS) version 21. As presented on table 6, there is a strong significant relationship between the perceived ease of use and perceived usefulness of the use of EMR .P-value = 0.000, NH1 is rejected.

Table: 6 PEOU and PU correlations

Correlations		
Factors		PU
PEOU	r-value	0.560**
	p-value	0.000
	N	1428
PEOU: Perceived ease of use; PU: Perceived usefulness		

As presented on table 7,8 there is a strong significant relationship between the perceived ease of use, perceived usefulness and attitude towards the use of EMR .P-value = 0.000, NH2 and NH3 are rejected.

Table: 7 PEOU and ATU correlations

Correlations		
Factors		ATU
PEOU	r-value	0.357**
	p-value	0.000
	N	1428
PEOU: Perceived ease of use; ATU: Attitude towards Using		

Table: 8 PU and ATU correlations

Correlations		
Factors		ATU
PU	r-value	-0.145**a
	p-value	0.000
	N	1428
PU: Perceived usefulness; ATU: Attitude towards Using		

a: The value is negative due to a negative average covariance among items. This violates reliability mel assumptions.

As presented on table 9,10 there is a strong significant relationship between the perceived usefulness, perceived ease of use and Self-Efficacy of the use of EMR .P-value = 0.000, NH4 and NH5are rejected.

Table: 9 PU and SE correlations

Correlations		
Factors		SE
PU	r-value	0.208**
	p-value	0.000
	N	1428
PU: Perceived usefulness; SE: Self-Efficacy		

Table: 10 PEOU and SE correlations

Correlations		
Factors		SE
PEOU	r-value	12.404**
	p-value	0.000
	N	1428
PEOU: Perceived ease of use; SE: Self-Efficacy		

As presented on table 11, 12 there is a strong significant relationship between the perceived ease of use, perceived usefulness and Perceived Behavioural Control of the use of EMR .P-value = 0.000, NH6 and NH7are rejected.

Table: 11 PEOU and PBC correlations

Correlations		
Factors		PBC
PEOU	r-value	0.721**
	p-value	0.000
	N	1428
PEOU: Perceived ease of use; PBC :Perceived Behavioural Control		

Table: 12 PU and PBC correlations

Correlations		
Factors		PBC
PU	r-value	0.694**
	p-value	0.000
	N	1428
PU: Perceived usefulness; PBC :Perceived Behavioural Control		

As presented on table 13,14 there is a strong significant relationship between perceived usefulness, perceived ease of use and Sufficient Training of the use of EMR .P-value = 0.000, NH8 is rejected.

Table: 13 PU and ST correlations

Correlations		
Factors		ST
PU	r-value	0.549**
	p-value	0.000
	N	1428
PU: Perceived usefulness; ST : Sufficient Training		

Table: 14 PEOU and ST correlations

Correlations		
Factors		ST
PEOU	r-value	0.670**
	p-value	0.000
	N	1428
PEOU: Perceived ease of use; ST : Sufficient Training		

As presented on table 15, there is a strong significant relationship between general information and attitude towards the use of EMR .P-value = 0.000, NH9 is rejected.

Table: 15 GI and ATU correlations

Correlations		
Factors		ATU
GI	r-value	0.083**
	p-value	0.000
	N	1428
GI: General information ; ATU: Attitude towards Using		

7 Conclusions

The findings show that there is a significant relation between ages, gender, nationality and the level of the qualification, and the knowledge or attitude towards EMR, although there is a significant relationship between gender, nationality, the level of the qualification and the nurses' work experience to use EMR. Most of the views asked for better qualifications in the area of EMR and health information systems in general. These statements need to be considered in the national processes of change towards e-health. Other barriers or concerns in relation to EMRs were linked to technical, financial and workload issues. The results from this study will helped other study investigating the perceptions of EMR and barrier to its uptake in both male and female nurses at all stages in their careers.

8 References

[1] R. Malakar, "Electronic medical records". *Indian Journal of Dermatology*, 51(2), pp. 140-141,2006.

[2] Ministry of Health Kingdom of Saudi Arabia.*"World Health Statistics Report." 2013, Available at http://www.moh.gov.sa/en/Ministry/Statistics/book/Documents/Statistics-Book-1434.pdf Accessed on [8 December 2015]*

[3] Ministry of Foreign Affairs. Kingdom of Saudi Arabia, available at: http://www.mofa.gov.sa/sites/mofaen/EServ/VisitingSaudiArabia/aboutKingDom/Pages/KingdomGeography46466.aspx]. Accessed on: [8 December 2015].

[4] A.Al-Shekh, "Perceptions of hospital experiences in Riyadh City, Saudi Arabia: a comparison of service quality in public and private hospitals", *University of Wales*, Swansea, 2003.

[5] M.Al-Yousuf, T. M. Akerelel, and Y.Al-Mazrou, "Organization of the Saudi Health System", *Eastern Mediterranean Health Journal*, 8(4-5):645-53, 2002.

[6] Saudi Association for Health Informatics."*National e-Health Strategy*" 2015. [Online] Available from: http://www.moh.gov.sa/en/Ministry/nehs/Pages/default.aspx [Accessed 8 December 2015]

[7] G A.A. Alshammasi, "The Influence of Economic, Political and Socio-cultural Factors on the Development of Health Services in Saudi Arabia"**.** University of Hull,1986.

[8] M. Almalki, G.Fitzgerald, , M.Clark, . "Health care system in Saudi Arabia: an overview". Eastern Mediterranean Health Journal. 17(10), 2011.

[9] M.M.Altuwaijri, "Electronic-health in Saudi Arabia. Just around the corner?" Saudi medical journal, 29(2), pp. 171-178, 2008.

[10] L.E. Moody, E. Slocumb, B. Berg, and D. Jackson, "Electronic health records documentation in nursing: nurses' perceptions, attitudes, and preferences". *Computers, Informatics, Nursing : CIN,* 22(6), pp. 337-344, 2004.

[11] T.W. Dillon, R. Blankenship, and T. JR. Crews, "Nursing attitudes and images of electronic patient record systems". *Computers, informatics, nursing : CIN,* 23(3), pp. 139-145, 2005.

[12] A.Whitaker, M. Aufdenkamp, S. Tinley, "Barriers and facilitators to electronic documentation in a rural hospital". J Nurs Scholarsh.41:293-300, 2009.

[13] M. Koivunen, R. Kontio, A. Pitkänen, J. Katajisto, and M. Välimäki, "Occupational Stress and Implementation of Information Technology Among Nurses Working on Acute Psychiatric Wards". Perspectives in Psychiatric Care. 49(1).P.41-49,2013.

[14] S. Al Alawi, A. Al Dhaheri, D. Al Baloushi, M. Al Dhaheri, and E.A.M. Prinsloo, "Physician user satisfaction with an electronic medical records system in primary healthcare centres in Al Ain: a qualitative study". BMJ Open, 2014.

[15] M.M. Nour el din, "Physicians' use of and attitudes toward electronic medical record system implemented at a teaching hospital in Saudi Arabia," The Journal of the Egyptian Public Health Association, 2007, 82(5-6), pp. 347-364.

[16] H.Lærum, T. Karlsen, A. Faxvaag, "Doctors' use of electronic medical records systems in hospitals: cross sectional survey". BMJ 323:1344-8, 2003.

[17] J.Handy, R.Whiddett, and I.Hunter, "A TECHNOLOGY ACCEPTANCE MODEL FOR INTER-ORGANISATIONAL ELECTRONIC MEDICAL RECORDS SYSTEMS". AJIS 9(1). 2001'School of Psychology and 'Department of Information Systems, Massey University, Palmerston North, New Zealand.

[18] A.Churchill, and C.Gilbert, "Marketing research: methodological foundations" (6th ed) New York: Dryden. 1995.

Bipolar Depression Druid: Wireless Technology Framework to Predict Bipolar Depression

Arshia A. Khan[1], Rushmeet Bahra[2]

[#]*Department of Computer Science, University Minnesota Duluth*
320 Heller Hall, 1114 Kirby Drive
Duluth MN, 55812-3016, USA
[1]akhan@d.umn.edu
[2]bahrad010@d.umn.edu

Abstract— *Bipolar Disorder affects approx. 5.7 million adult Americans(NIHM). It is the 6th leading cause of disability in the world(WHO) driving a need of constant monitoring. Bipolar Depression, one of the expressions of Bipolar Disorder is a chronic mental illness that can prove harmful if not monitored and treated early. Early detection is the key to the prevention of adverse consequences of a bipolar episode that can in an extreme situation lead to suicidal attempts. With the wide use of mobile devices, there is a great potential to harness the power of size, mobility, convenience, cost efficiency and easy access to efficiently augment and complement the management of chronic illnesses such as Bipolar Depression [3].*

This study proposes a wireless solution to monitor bipolar depression. The solution involves the use of wearable wireless sensor to track the patient vitals – such as heart rate, in addition to sleep, mood patterns and medication to identify the prodrome to predict a bipolar episode.

Keywords **–** bipolar depression, health informatics, wireless solution to bipolar depression

1 Introduction

Bipolar Disorder affects approximately 5.7 million adult Americans every year or about 2.6% of the U.S. population age 18 and older every year. The median age of onset for bipolar disorder is 25 years, although the illness can start in early childhood or as late as the 40's and 50's. An equal number of men and women develop bipolar illness and it is found in all ages, races, ethnic groups and social classes

More than two-thirds of people with bipolar disorder have at least one close relative with the illness or with unipolar major depression, indicating that the disease has a heritable component [4].

Although only 1% of the population in the United States reported Bipolar Disorder (BPD) in a study that replicated the National Comorbidity Survey, this illness can lead to grim outcomes such as suicide and other consequences such as heart disease and diabetes. The threshold criteria to diagnose BPD are necessary and in it's absence can lead to hospitalizations, increased healthcare costs and associated medical conditions that can result due to untreated BPD. The most premature deaths from suicide are caused due to BPD. There is an observed increase in healthcare costs due to manic and depressive episodes. Bipolar disorder can present itself in many forms including mania, which can last for as long as a week or more; hypomania, which can last as long as four or more weeks; and major depressive episode (MDE). Some consistent socio demographic correlations can be observed across the BPD spectrum such as – an inverse correlation to age, education level, previous marital status, and employment status; and no correlation to age, ethnicity, race and family income [1].

1.1 Bipolar Disorder

Depression is a chronic mental illness experienced by numerous individuals, irrespective of age, sex, caste, and creed. It is one of the most common prevalent mental illnesses where an individual experiences sadness, dullness, anger and frustration for a longer period of time.

Various types of Depression are listed below:

- Bipolar Depression - a type of bipolar disorder is a mental condition where an individual experience episodes of extreme melancholy even leading to a heightened desire to commit suicide.
- Postpartum Depression is most common among pregnant women characterized by a state of shock after childbirth. Women under extreme stress are more prone to this type of depression.
- Seasonal Depression can be triggered by changes in weather, especially during long winters where the lack of sunlight can activate a depression episode.
- Premenstrual Dysphoric Disorder is experienced by women during their menstruation cycles, which is characterized by frustration and edginess before the onset of their menstruation, also known as premenstrual syndrome. Prolonged symptoms can be indicative of premenstrual dysphoric disorder.

Bipolar Depression patients may get rush of feelings, known as the episode where they can either harm themselves or people around them.

1.2 Need for Treatment That Involves Technology

Treatment is common among 80% of the patients suffering from lifetime BPD. Of these some are treated by psychiatrists while others are treated by medical professionals that are not psychiatrists. Approximately 45% of the patients who are receiving treatment from psychiatrists receive the appropriate treatment while only 9% of the patients treated by non-psychiatrist professionals receive appropriate treatment. There is a clear lack in treatment by psychiatrists. To reduce healthcare costs health institutes are moving towards the employment of non specialized processionals[1]. Mobile interventions have been used to provide psychoeducation [12] and adscititious support to treatment by augmenting the reach of therapy [2].

1.3 Mobile Interventions to Enhance the Care of Patients Suffering from BPD

The Mobile Health Worker Project study revealed five key benefits of the application of mobile technology to improve the healthcare delivery – i) the healthcare professionals were found to benefit from mobile technology; ii) the patient data could be shared and viewed by clinicians at multiple clinical services; iii) the time spent with patients could be improved by much as 104%; iv) there was a reduction in the duplication of data by 92%; and v) the number of no access visits were reduced by 50%. This and other studies provide evidence that the mobile interventions have a potential to improve healthcare delivery and practice [5]. In addition the management of behavioral risk factors and education of chronic health conditions can potentially improve healthcare management. 70% of deaths are caused due to some chronic health condition, while 75% of healthcare costs can be attributed to the management of chronic illnesses. Mobile technology can not only extend the reach of the management of chronic diseases such as heart, cancer and lung but also BPD and the result of self monitoring and management can bring more awareness and education of symptom management and prevention [6]. The efficient use of mobile technology can help manage prevent and extend the reach of therapy in patients with mental illnesses. Although the mobile technology cannot replace the face-to-face therapy, it can certainly augment cognitive behavioral therapy and improve self- management of mental illness. Mobile phones not only increase access to care but increase supervision and surveillance and the monitoring of health outcomes [7].

1.4 Behavioral Science and it's Role in Enhancing BPD Management

Behavioral science research has emphasized the understanding of human health behavior and how technology can be leveraged to improve health. Mobile technology plays an important role due to its ease of accessibility and availability in understanding health behavior and implementing mobile interventions to alter this behavior to impact public health [8]. The integration of behavioral science and technologies such as mobile devices, wearable sensors, and the ability to bring access to therapeutic support has the potential to better manage mental health by developing systematic structures and effective evidence based algorithms [11].

1.5 Challenges of Mobile Interventions in Mental Care Management

Although mobile technology has the potential to aid in the management, acre and prevention of mental illnesses, it also poses many challenges such as ethical use of the technology, the implementation of proper therapy algorithms, and the advise and guidance of qualified professionals from the field of psychology [7].

Another challenge is that the mobile interventions have not been evaluated by means clinical trials. There is a lack of research on the effectiveness of the mobile interventions. There is clearly a need to understand health behavior and it's impact on health, especially mental health concerns including BPD [8][9]. In the process of mental health reform, mobile solutions for the assessment, prevention and treatment are emerging. The challenge is to ensure that the population served by the mobile technologies have sufficient access, are engaged and most importantly benefit from these technologies [9].

2 Framework with Emphasis on Structure and Function in BPD

The advocates, promoters and supporters of mental health reform encourage collaborative investigative practices that would make mobile mental health solutions a reality. In order to accomplish this the mobile solutions should be grounded in interdisciplinary evidence based research from psychology, computer science, public health and engineering. This practical and accessible solution should be founded on a conceptual framework that emphasis structure and function of the mobile intervention. The structure component would involve the delivery mechanism with the ability to reach various age groups, from children through adulthood and geriatrics. In this sense the mobile solution appears to be one of the most viable and sustainable solution due to its ability to extend and make therapy accessible. In terms of functionality various tools were explored that proposed mental health assessment and treatment and the problems from a psychological perspective were studied to propose a codependency on sleep patterns, mood, heart rate and BPD. The function component of this proposed solution is based on research that identifies lack of sleep, mood and heart rate as variables that are affected by each other and can potentially lead to a BPD episode [9] [10][11]. Lack of sleep can lead to changes in mood and vice versa, changes in mood can lead to lack of sleep, with both sleep and mood playing a major role

in the onset of BPD [10][12]. The proposed conceptual framework can be seen in Fig 1.

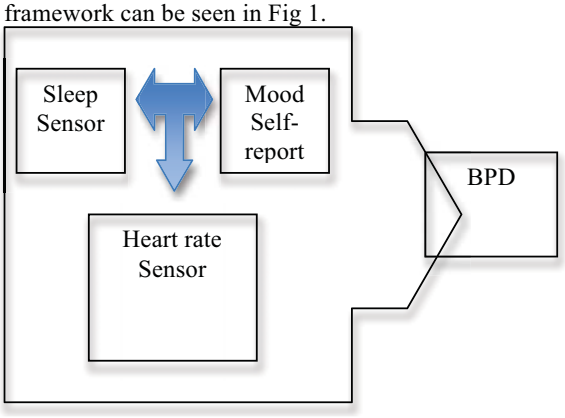


Fig 1. Conceptual Framework for BPD [10]

2.1 Mood Spectrum

The mood swing element is derived from the mood spectrum as seen in Fig 2. The state of depression increasing from left to right in this figure, with Unipolar being the least complex form of depression and the Bipolar I being the most complex with manic episodes. Depression can be either genetic/chemical in nature or situational, for example death of a loved one [12].

Fig 2. Mood Spectrum [12]

3 Methods

A study conducted by [2] evaluated the quality and applicability of existing mobile solutions to the management of BPD. Of the apps evaluated 36% were informational apps that provided information on BPD, 15% provided guidelines to best practices, the rest offered tools for symptom monitoring, assessment, screening and community support. This study identified some features and capabilities that the existing apps were lacking. Some of the shortfalls of the existing apps are as follows:

- Privacy and security was not addresses appropriately
- Important factors such as medication management was not taken into consideration in the symptom management as medication compliance can play an important role in BPD
- Another important factor that was ignored was the sleep patterns
- The lack of citation of the resources
- And the insufficient use of the validated screening measures [2].

The solution proposed in this paper takes into consideration all of the above concerns and addresses them individually and incorporates them in the proposed framework.

Patients who suffer from bipolar depression lack sleep and this in turn can lead into mood swings leading to bipolar depression. Medication compliance can positively or negatively impact the occurrence of a BPD episode. Non-compliance of prescribed medication and therapy in a timely and accurate manner is a major cause for the triggering of BPD.

3.1 Technological Structure and Platform Based on the Conceptual Model

The app proposed to predict the prodrome for BPD is built on the iOS platform with Objective-C as the langue for programming with Xcode as the integrated development environment. The Basis Peak sensors for the heart rate and the sleep are employed to gather the sleep and heart rate via Bluetooth. The app is designed to read the data from the sensors and store it in real time with a time stamp using the Bluetooth manager framework in the iOS platform.

3.2 Technological Functions Based on the Conceptual Model

An app is proposed as a solution for tracking and monitoring the bipolar depression. A wearable wrist sensor that tracks sleep and heart rate is employed to monitor sleep patters. The mood can be tracked by the app as a self-reporting tool. This app has the following components

- Mood tracker – the mood tracker prompts the user for the mood. The user will enter the mood by clicking on the face emoji that represents their mood. This algorithm is based on the mood spectrum described in the previous section. The self reported mood data gathered by the app will be saved and analyzed along with numerous other patients mood data to better understand the role of mood in BPD and other mental illnesses.
- Sleep tracker – The sleep data is acquired from a wearable sleep-tracking sensor. The real time data is decrypted and stored to be processed and analyzed independently to understand sleep and make evidence based recommendations based on the sleep data collected from multiple patients. The same data is also processed in conjunction with the mood data to not only better understand BPD but also predict the onset of a BPD episode.
- Heart rate tracker – There is evidence that the heart rate is a physiological measure of stress and illness severity. The heart rate was more constricted among patients with increased illness severity. Hence there can be a correlation between illness severity, heart rate and BPD. The heart rate of a person suffering from bipolar

depression varies considerably. The wearable sensor that can track the heart rate is employed to acquire real time heart rate and processed independently and in correlation with the mood and sleep data with respect to BPD. This data can be further used to make evidence-based recommendations to patients suffering from BPD [14][17][18].

- Medication manager – This part of the app verifies medication compliance to the prescription in addition to providing notifications and reminders to take medications. One of the major problems with chronic illnesses is medication non-compliance. Medication compliance is critical in the prevention of BPD episodes [15] [16].
- HIPAA compliance of privacy and security – Management of the data according to the HIPAA law is an essential component of this app.
- Psycheducation component – This is another componenet built into the app to provide information on BPD. It is critical for patients to understand the prodrome for the BPD and psychological interventions such as the psycheducation can help the patient identify their prodrome before the onset of the episode and take precautionary preventive measures to avoid an episode[12][16]
- Data Collection for evidence based practice. The data from the app will consist of self-reported mood data and the medication compliance data. This data along with the data from the sensor that tracks sleep and the heart rate is collected and stored in the cloud to be analyzed to generate future recommendations to patients based o the evidence gathered from this data.
- Health behavior management – Involves monitoring of sleep patterns and making recommendations to realize and management it to develop healthy sleep habits is essential for the mental well being. Clinicians can make recommendations to change sleep behavior based on the recorded sleep data.

The sensors data is acquired through Bluetooth technology and corroborated with the mood to predict a bipolar episode.

3.3 Research Question and Hypothesis

Research Question: There exists an algorithm that combines sleep, mood and heart rate that can be detected using mobile technology to predict BPD.

The hypothesis are:

- H1: sleep does have a direct impact on the onset of BPD
- H2: mood changes have a direct impact on the onset of BPD
- H3: sudden changes in heart rate can help determine the onset of BPD
- H4: The data analysis of the sleep, mood and heart rate can produce significant recommendations evidence based practice.

The algorithm for the bipolar depression app is shown in fig 3.

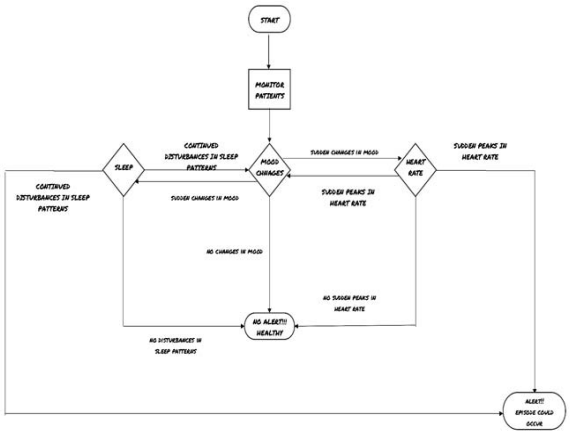


Fig 3 – Algorithm for the detection of BPD episode

The screen shots for the app are shown in figs 4,5,6,7.

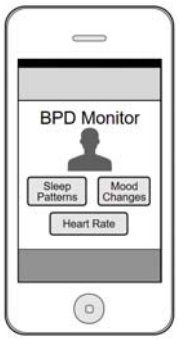


Fig 4

Fig 5

Fig 6

Fig 7

4 Discussions

A proposal was developed to monitor bipolar depression episodes using wireless mobile technology. Sleep patterns, mood swings and physiological elements such as heart rate can be used to predict the onset of an episode of bipolar depression. Early intervention can alleviate the episode of BPD. Use of wireless technology can help the patient from harming themselves or people around them. Mobile devices has become the device of communication in most households, with 25-29% of households relying on mobile devices solely as a means of communication instead of the land lines [21]. The mobile device with the wearable sensors utilize an app that is designed to take the input from human on the mood and read the sleep pattern from the wearable sensor and finally take into consideration a physiological vital such as the heart rate. This data will be stored in the cloud and analyzed to find patterns that would help develop an algorithm to predict the prodrome for a bipolar episode. The algorithm in figure 3 depicts the path of the app in predicting the prodrome for BPD. The purpose of this app is prevent the onset of new episodes by warning the person and suggesting behavior changes to prevent the episode from occurring. Inaddition data collected from numerous patients can be corroborated to predict patterns based on which the clinicians can make recommendations to patients- practicing evidence based decision-making.

The app has the following features:

1. Educational component
2. Self-reporting mood tracker – Fig 7
3. Wireless wearable sensor that monitors and tracks sleep –Fig 5
4. Wireless wearable sensor that monitors heart rate tracker – Fig 6
5. Medication manager

4.1 Outcomes

- Proposal to develop mobile intervention to provide psych education and offer adscititious .
- A framework based on sleep patterns, mood swings, heart rate, and medication compliance is proposed.
- The proposed solution attempts to predict a bipolar episode which can be treated immediately
- Immediate treatment has a potential to reduction in hospitalizations, healthcare costs, and most importantly premature death due to suicide.

5 Conclusions

The framework for a wireless solution that uses sensors to monitor and identify a prodrome to predict bipolar depression is proposed. The iOS platform is employed with Objective-C as the primary programming language with Xcode as the integrated development environment. Based on the evaluation of existing mobile solutions for BPD a solution that uses sensors and Bluetooth technology is proposed. This evidence based conceptual framework resulted from the integration of mobile devices, wearable sensors, the need to bring access to therapeutic support and behavioral science to better manage mental health by developing an effective evidence based algorithm [10][11]. The heart rate and the sleep sensor are employed to track the sleep and heart rate. This sensor data is acquired via Bluetooth and stored in the cloud along with the self reported data from the mood tracker app.

There are many benefits to the use of smart mobile devices to manage mental illnesses such as BPD. Of the many benefits some can be identified as an improvement in the physician-patient communication, increased therapeutic reach, accessibility, symptom assessment, psycheducation, resource locator, a formalized system to track treatment progression and the ability to practice evidence based health recommendations [20]. Studies have found mobile intervention to improve the delivery of psychotherapy and behavioral interventions. Hence this app can augment the current therapy for BPD, overcome barriers and enhance the clinical practices [13][19].

Some of the challenges associated with the use of technological solutions to enhance care are cost, usability, network bandwidth, accessibility, battery power issues, limited resources due to the mobile platform and privacy issues. The mobile solution we propose has been developed taking all of these considerations into account and to maximize it's use and efficiency while maintaining security and privacy not sacrificing any computation power due to the low resources of the mobile app. Some of these hurdles have been overcome by the use of cloud as a platform for storage and data analysis[21].

6 References

[1] Merikangas, Kathleen R., et al. "Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication." Archives of general psychiatry 64.5 (2007): 543-552.

[2] Nicholas, Jennifer, et al. "Mobile apps for bipolar disorder: a systematic review of features and content quality." Journal of medical Internet research17.8 (2015): e198.

[3] Belmaker, R. H. "Bipolar disorder." New England Journal of Medicine 351.5 (2004): 476-486.

[4] Nimh.nih.gov,. 'NIMH » Bipolar Disorder'. N. p., 2014. Web. 11 Dec. 2015.

[5] Drayton, K. "How mobile technology can improve healthcare." Nursing times109.11 (2012): 16-18.

[6] Stellefson, Michael, et al. "Use of Health Information and Communication Technologies to Promote Health and Manage Behavioral Risk Factors Associated With Chronic Disease: Applications in the Field of Health Education." American Journal of Health Education 46.4 (2015): 185-191.

[7] Norris, Lexi, Leslie Swartz, and Mark Tomlinson. "Mobile phone technology for improved mental health care in South Africa: possibilities and challenges."South African Journal of Psychology 43.3 (2013): 379-388.

[8] Pagoto, Sherry, and Gary G. Bennett. "How behavioral science can advance digital health." Translational behavioral medicine 3.3 (2013): 271-276.

[9] Jones, Deborah J. "Future directions in the design, development, and investigation of technology as a service delivery vehicle." Journal of Clinical Child & Adolescent Psychology 43.1 (2014): 128-142.

[10] Talbot, L. S., Stone, S., Gruber, J., Hairston, I. S., Eidelman, P., & Harvey, A. G. (2011, August 15). A Test of the Bidirectional Association Between Sleep and Mood in Bipolar Disorder and Insomnia. Journal of Abnormal Psychology. Advance online publication. doi:10.1037/a0024946
[11] Dallery, Jesse, Allison Kurti, and Philip Erb. "A New Frontier: Integrating Behavioral and Digital Technology to Promote Health Behavior." The Behavior Analyst 38.1 (2014): 19-49.
[12] Psycheducation.org,. "Psycheducation | Treating The Mood Spectrum". N. p., 2015. Web. 21 Dec. 2015.
[13] Kratzke, Cynthia, and Carolyn Cox. "Smartphone technology and apps: Rapidly changing health promotion." International Electronic Journal of Health Education 15 (2012): 72.
[14] Levy, Boaz. "Illness severity, trait anxiety, cognitive impairment and heart rate variability in bipolar disorder." Psychiatry research 220.3 (2014): 890-895.
[15] Keck Jr, Paul E., et al. "Compliance with maintenance treatment in bipolar disorder." Psychopharmacology bulletin 33.1 (1997): 87.
[16] Gastó, Cristòbal. "Psychoeducation efficacy in bipolar disorders: beyond compliance enhancement." J Clin Psychiatry 64.9 (2003): 1101-1105.
[17] Chang, Hsin- An, et al. "Heart rate variability in unmedicated patients with bipolar disorder in the manic phase." Psychiatry and clinical neurosciences68.9 (2014): 674-682.
[18] Migliorini, Matteo, Martin O. Mendez, and Anna M. Bianchi. "Study of heart rate variability in bipolar disorder: linear and non-linear parameters during sleep." Frontiers in neuroengineering 4 (2011).
[19] Lindhiem, Oliver, et al. "Mobile Technology Boosts the Effectiveness of Psychotherapy and Behavioral Interventions A Meta-Analysis." Behavior modification 39.6 (2015): 785-804.
[20] Luxton, David D., et al. "mHealth for mental health: Integrating smartphone technology in behavioral healthcare." Professional Psychology: Research and Practice 42.6 (2011): 505.
[21] Boulos, Maged NK, et al. "How smartphones are changing the face of mobile and participatory healthcare: an overview, with example from eCAALYX."Biomedical engineering online 10.1 (2011): 24.

Active-Workspaces: A Dynamic Collaborative Business Process Model for Disease Surveillance Systems

Nsaibirni Robert Fondze Jr[1,5], Eric Badouel[2], Gaëtan Texier[3,5], and Georges-Edouard Kouamou[4]
[1]LIRIMA, University of Yaounde 1, PO Box 812, Yaounde, Cameroon
[2]Inria and LIRIMA, Campus de Beaulieu, 35042 Rennes, France
[3]Centre d'épidémiologie et de santé publique des armées (CESPA),
UMR 912 - SESSTIM - INSERM/IRD/Aix-Marseille Université
[4]LIRIMA, ENSP, PO Box 8390, Yaounde, Cameroon
[5]Centre Pasteur du Cameroun, Yaoundé, Cameroun.

Abstract— *Flexibility and change at both design- and run-time are fast becoming the Rule rather than the Exception in disease surveillance processes. This is attributed to the diversity in public health threats, to continuous advances in domain knowledge, the increase in expert knowledge, and the diverse and heterogeneous nature of contextual variables. Disease surveillance is one such processes and it is characterized by collaborative work and decision making between users with heterogeneous profiles on processes designed on-the-fly. A model for disease surveillance processes should thus natively support flexible workflow design and enactment as well as human interactions. We show in this paper how the Active Workspaces model proposed by Badouel et al. for distributed collaborative systems provides this support.*

Keywords: Disease Surveillance, Business Process Modeling, Collaborative Systems, Active-Workspaces

1. Introduction

For over twenty years, public health information systems have prospered in all medical areas and activities, in line with the advances in health informatics and related technologies. These systems are identified by the American Medical Informatics Association (AMIA) as belonging to Public Health Informatics (PHI), a specific subdomain of Health Informatics, defined as "the systematic application of information and computer science and technology to public health practice, research, and learning [23]. The scope of PHI was described as "the conceptualization, design, development, deployment, refinement, maintenance, and evaluation of communication, surveillance, information, and learning systems relevant to public health." A recent article in the AMIA yearbook of medical informatics [27] introduces a review of English-language PHI publications in Medline (2012-2014), in which authors propose main essential services such as monitoring health, supporting diagnosis, investigating outbreaks, and evaluating systems. The systems providing these services could be considered as decision support systems since it uses data, documents, knowledge and/or models to identify and solve problems and make decisions.

Concerning syndromic surveillance, defined as the continuous monitoring of public health-related information sources for early detection of adverse disease events, numerous early warning systems are currently used by experts belonging to international, national or local public health institutions. This decreases the response delays, improves effectiveness, and reduces the health impact of the outbreak. According to Chaudet et al. [26][17], outbreak identification and confirmation are managed by epidemiologists during "situation diagnosis," which consists in validating (or revoking) an alarm (signal identified as aberrant or abnormal) and transforming it into an alert (real characterized outbreak), then proposing initial countermeasures.

In health domains, known for their complexity and uncertainty, carrying out situation diagnosis implies complex decision-making processes and involves a wide range of interrelated human, biological and/or environmental activities. A disease surveillance network is thus a socio-technical system which associates geographically distant medical stakeholders (up to a few thousand people in different specialties) with dedicated systems and technical tools (telephone, satellite, digital documentation, ...) collaborating to detect and manage outbreaks [28]. More so, disease surveillance is a semi-structured [16] process which entails that only high level tasks can be clearly defined prior to process execution since most of the activities are discovered at runtime as data becomes available. This increases the complexity as users have to design and run the process-workflow on-the-fly.

Such a system in which users collaborate and share information intensively over a process model defined on-the-fly is termed a dynamic knowledge intensive system [2][11]. The modelling objective in such systems is not to completely automate the processes and their orchestrations but to provide users with expressive tools to permit them flexibly and efficiently create and run processes while making optimal use of the resources at their disposal. These tools can be grouped into four main categories:

1) **Tools for Real-time Iterative Workflow Construction and Orchestration:** As mentioned above, situation diagnoses for instance which is a major phase

in the syndromic surveillance process is an expert activity[17][26]. This means that the decisions and actions to be taken are determined by the expert usually based on incomplete non-pathogenic data. Thus the activity though standardised but remains highly unpredictable.

2) **Tools for User-Interactions:** Disease surveillance is a distributed collaborative activity (spatial and temporal) involving several stakeholders with diverse profiles[29][28]. These stakeholders interact (asynchronously) in myriad ways to find solutions to questions raised during disease surveillance[28].
3) **Tools for managing Exceptions and Uncertainty:** Disease surveillance data is usually described as being incomplete, non-pathogenic, and biased [17]. These are sources of uncertainty and inconclusive decision making. This uncertainty is even accentuated when attempts are made to predict future disease incidences. [30] presents uncertainty as one of the cross-cutting issues that all disease surveillance systems need to address.
4) **Tools to support Decision Making:** The main objective of monitoring diseases is to facilitate decision making and take timely action against public health threats[8][31][32]. In [31], PHI decision support is defined as the process of bringing relevant knowledge to bear to aid decisions involving the health and wellbeing of a population through the use of electronic information. Providing decision support is thus mandatory in all PHI information systems.

In this paper, we present an informal description of the Active-Workspaces model [1], a distributed, user-centric, and data-driven business process model built on guarded attribute grammars. Though the Active Workspaces model can be easily extended to address all of the four tools above, we limit this paper to showing how it provides support for Tools 1 and 2.

The rest of the paper is organized thus: section 2, presents related works in disease surveillance process modeling and business process modeling tools; section 3 presents an illustrative scenario; Sections 4 and 5 respectively elucidate the Active-Workspaces model with its user-centered collaborative constructs, and how the workspace can evolve. Conclusions and future works are stated in section 6.

2. Related Work

Research in public health informatics and disease surveillance in particular has focused on identifying trends/patterns in diseases, potentially viable data variables and sources, and developing novel methods of collecting, aggregating, analysing, and interpreting surveillance information. Little has been done to capture the activities, data, decision, and collaboration schemes that are involved in disease surveillance. In [6], [5], [19] and [8] high-level steps are presented with sample activities that can be carried out at each of them. They go further to characterise the environments (pre-conditions) that favour the application of each of these activities. These pre-conditions only become satisfied at run-time thus supporting our argument for iterative process design and execution.

Futhermore, business process modelling use cases have evolved so far from models that stress on the control and coordination of tasks using state-based formalisms like automata and petri-nets [18][21][14][20][13][10], through data-centric approaches [22][25][4] that use data to dictate the orchestration of activities in a business process, to artifact centric workflows [7][3][24][15] that combine data and activities in one whole (artifacts) and use state-based [24] or declarative [3] [7][15] constructs to guide the evolution of these artifacts in a business process. These techniques however are adapted for structured-domains since they lack the required flexibility needed in disease surveillance processes and place users in the external environment.

3. Illustrative Scenario

We describe below scenarios in syndromic surveillance to better motivate the work presented in this paper.

Several users participate in this scenario: clinicians, biologists, epidemiologists, and pharmacists. We suppose that an Influenza outbreak alarm has just been raised and an epidemiologist assigned to investigate the alarm. We recall that the investigation process aims at confirming the alarm into an alert or revoking it.

The epidemiologist knows of the existence of the different actors listed above but cannot say a priori when or how he will need them during the investigation. Suppose for example that the indicator variable that produced the alarm was $pharmarcy\_sales$. He will start by contacting pharmacists in the epidemic zone to ensure that the sales hike is genuine, that is, it is not caused by some commercial campaign or a similar activity. The alarm is immediately revoked if the latter is true. Otherwise, he has to investigate more. Given the high sensitivity associated with using $pharmacy\_sales$ as an indicator, he decides to pursue tasks that use data tightly correlated with the outbreak. In this case clinical and laboratory diagnostic data. He contacts clinicians in health districts around the epidemic zone for consultation data and runs additional analysis. He requests that patients with Influenza symptoms be contacted and samples obtained if possible and that this be carried out systematically for all new patients presenting symptoms of Influenza. He can even go ahead to request that each sample be multiplied and sent to different biologists for laboratory analysis. This especially if he possesses the required resources or if several tests need to be carried out and he wants to maximize time by spreading the tests across several laboratories.

In parallel to the activities above, he also has to manage a number of support activities such as organizing the transportation of samples from health centers to laboratories, ensuring that the laboratories possess the required reagents and equipment to run the requested tests, etc. He also has to report regularly to public health officials to help them prepare the resources to contain the potential outbreak. He continuous to initiate and run activities collaboratively with other actors until he reaches a conclusion.

If on the other hand the indicator variable was different, say $school\_absenteeism$ or $triade\_calls$ or $consultation\_data$, a completely different set of activities will probably be executed. Futhermore, if this task was assigned to a different epidemiologist, it is not certain that he will run the activities in the same order, or even use the same set of activities. This is because the latter and their ordering highly depend on the experience and expertise of the user and on how much he knows of his environment. Hence the Knowledge-Intensive character of surveillance systems.

This scenario shows how complex resolving a simple task might become when new data becomes available and how unpredictable the surveillance process can be. A model for such a process should therefore provide flexible constructs for building and executing process workflows on-the-fly. The fact that the process model changes is the rule and not the exception.

We also note different forms of interactions between the users and their working environments and equipment (phones, computers, etc.), and among users. For example, the epidemiologist has to interact with his work environment to accept and complete the alarm investigation request and at some point he needs to communicate with other users by sending new requests. Suppose that for some reason in the middle of the investigation, the acting epidemiologist becomes unavailable, the activities he has carried out as well as the information he has gathered will have to be transfered to the new epidemiologist. This is another form of interaction between users: synchronizing expert data.

4. Active-Workspace : User-centered Flexible and Collaborative constructs

In this section we present a succinct informal definition of the Active-Workspaces model. We lay emphasis on the properties that are required to address the two preoccupations treated in this paper. A more formal and complete description of the model is found in [1].

4.1 Active-Workspace

The Active-Workspaces (AW) model is an asynchronous cooperation model in which each participating user is assigned a workspace. A user's workspace is an arborescent (mindmap-like) structure that holds all tasks in which the user is involved as well as the data required to resolve these tasks. The arborescent structure is reminiscent of the hierarchical organisation of tasks in which large complex tasks are broken down to small less complex ones. Each node of the mindmap has a *sort* s indicating the name of the task assigned to it. Task s can be further decomposed into subtasks $s_1, \ldots, s_n$ by applying *production* $P : s \rightarrow s_1, \ldots, s_n$. A node is said to be *closed* when one such production P has been applied to it, otherwise, it is an *open node*. In the former case, the node has successors corresponding to subtasks in the right hand side of P. If the right hand side of P is empty, then node s is a *leave* of the tree. Open nodes, also called *buds*, have no successor nodes. A bud represents a *pending task* that requires the attention of the user: the bud grows when the user decides to apply a production to it. When this happens, the bud becomes a closed node associated with the production and it has n successor nodes that are newly created buds given by the subtasks $s_1, \ldots, s_n$ in the right-hand side of the production.

The hierarchical decomposition of tasks is thus not predefined but depends on decisions made by the user at each step. In disease surveillance, this is particularly useful especially during situation diagnosis. For example, faced with an Influenza outbreak alarm, an expert has to decide whether to use an approach that integrates clinical information, laboratory diagnostic information, spatial data, more profound data analysis, etc. or to just stick with an approach that combines a few of these activities. These approaches can be captured in different productions with the same sort from which the expert can choose when necessary.

Also, Active Workspaces have two main structurally independent layers: an underlying guarded attribute grammar (GAG) model and a GAG execution engine. Any changes made to the underlying grammar are directly visible to the execution engine. This means that new production rules can be added to the grammar at any time and they are immediately available for subsequent task resolution. In the example above, if the expert wants to use an approach for which no defining production exists, he can instantly create one and use it.

4.2 Collaboration and User Interactions

Each workspace is associated with at least one service rendered by the user. A service is represented by a unique sort called the *axiom* of the grammar. The particularity of this sort is that it does not appear in the right hand side of any production of the grammar. Nodes whose sorts are axioms (service nodes) are directly attached to the root node of the workspace tree. The resulting sub-tree rooted at such a node is called an *artifact*. A service call therefore instantiates a new artifact, reduced to a single bud at the root of the workspace. This artifact then develops by the application of productions until it contains no open nodes, that is, the service has been completely rendered. In a multi-user context, we model collaboration between the different

workspaces. Each workspace is associated with at least one grammar identified by its axiom and a set of productions. The sorts of a grammar are either local to the grammar (that is, they appear at the left hand side of at least one production of the grammar), or external (that is, they make reference to axioms of other grammars). Applying a production is just like in a single user scenario with the difference that a sort at the right hand side of a production which references a different grammar will be interpreted as a call to an external service. Resolving this kind of open node provokes the creation of a new artifact in the workspace of the user to whom the grammar is attached. The behaviour of the workspace remains the same as in the single user scenario but for the fact that parts of an artifact will be developed at distant sites when service calls are made.

For example, in the syndromic surveillance scenario above, the epidemiologist requests the expertise of clinicians and pharmacists to investigate the alarm. The clinician in turn requests the services of biologists to run a series of tests on extracted samples. All these interactions between the users are materialised through service calls in the Active-Workspaces model.

4.2.1 Roles

Usually several users play the same role in a system. For example in disease surveillance, there exist several clinicians, several biologists, several epidemiologists, etc. This means that these users (in the same role) are attached to the same grammars after a local renaming of the local sorts. Technically, a role is defined by a generic grammar G and we obtain the disjoint union of these grammars as follows $\oplus(r :: R)G = \biguplus_{r::R} G[r]$ where r is a user who plays role R and $G[r]$ is the grammar obtained from G by replacing each sort (including the axiom s_0) by $s[r]$. Hence $s_0[r]$ represents service s_0 offered by r.

We note $G'\{G[r]\ where\ r :: R\}$ the grammar made up of $\oplus(r :: R)G$ and of a grammar G' that calls this role. That is, G' will at some point need to request a service from a user in this role. In G', we will find productions with parameters such as $P[r] : s \to s_0[r]$ expressing that when the user chooses production P to apply at an open node, he inputs a user r playing role R. The effective production is thus an instance of this generic production. We can also find in G productions of the form $P : s \to s_0[R]$ expressing that a service call is made to all users of the role R. In this case, the production has no parameters since the request will not be made to a particular user.

When a grammar needs to call several roles, we note $G\{G_1[r_1]\ where\ r_1 :: R_1; G_2[r_2]\ where\ r_2 :: R_2; \ldots\}$ and this construction can be applied hierarchically to model chained calls as follows: $G_1\{G_2[r_2]\ where\ r_2 :: R_2\ and\ G_2 = G\{G_3[r_3]\ where\ r_3 :: R_3\ and\ G_3 = G\{\ldots\}\}\}$. This constitution of roles is dynamic as new users can subscribe and/or un-subscribe from one or more roles at any moment. Adding a new user to a role poses no particular difficulty since it does not modify existing workspace specifications but only modifies productions which will be called subsequently. However, removing a user from a role might become problematic if there exist in his workspace artifacts with buds. We can in such a situation either forbid the user from unsubscribing from the corresponding role, or transfer the pending artifacts to the workspace of some user of the same role. Also, as we will see later on in this paper, it is possible for a user to define new productions and extend his local grammar. This means that two users with the same role and thus with identical grammars initially might later possess different grammars. In this case, a synchronization of the two grammars is necessary before the transfer operation.

4.3 Attributes and Guards

Productions are used to structure a user's workspace. They are however not sufficient to model the interactions and data exchanges between the various tasks associated with open nodes (buds). For that purpose, we attach additional information to open nodes using *attributes*. Each sort $s \in S$ comes equipped with a set of *inherited* attributes and a set of *synthesized* attributes. Values of attributes are given by terms over a ranked alphabet. Calling a *task* is written as $(y_1, \ldots, y_m) \leftarrow s(t_1, \ldots, t_n)$ where the $t_i's$ are terms denoting the values of the inherited attributes of task and $y_1, \ldots, y_m$ are (distinct) variables subscribing to the values of its synthesized attributes. The rationale is that we invoke a task by filling in the inherited positions of the form –the inputs– and by indicating the variables that expect to receive the results returned during task execution –the subscriptions– . A (business) rule R with underlying production $s_0 \to s_1 \ldots s_k$, which we note as $P[r] :: s_0 \to s_1[r] \ldots s_k$, is expressed using the following notation:

$$
\begin{array}{l}
s_0(p_1, \ldots, p_n) \ = \\
\textbf{input}(r, z_1, \ldots, z_l) \\
\ \textbf{do} \\
\qquad (y_1^{(1)}, \ldots, y_{m_1}^{(1)}) \leftarrow \mathbf{s_1}[r](t_1^{(1)}, \ldots, t_{n_1}^{(1)}) \\
\qquad \ldots \\
\qquad (y_1^{(k)}, \ldots, y_{m_k}^{(k)}) \leftarrow s_k(t_1^{(k)}, \ldots, t_{n_k}^{(k)}) \\
\qquad \textbf{return}\ (u_1, \ldots, u_m)
\end{array}
$$

This functional presentation stresses out the operational purpose of business rules: Each task has an input –inherited attributes– seen as parameters and an output –synthesized attributes– seen as returned values.

- The p_i's are patterns serving as *guards* for the rule.
- Variables z_l inside the **input** directive represent values not directly inherited from parent tasks (including users, r) and which will have to be provided by the user when the rule is chosen. This directive is omitted if no such variables exist in a rule.

- The u_j's describe the synthesized values produced when applying the rule.
- The expressions in the right-hand side are the subtasks that will be associated with the newly created open nodes.

The variables $y_i^{(j)}$ and the variables occurring in patterns are the input variables, they are pairwise disjoint and denote respectively the information synthesized by the subtasks and the information stemming from the context of the node. The $t_i^{(j)}$ and the u_j are terms over the input variables called the *semantic rules*. They provide respectively the values of the inherited attributes of the subtasks and the values of the synthesized attributes of the main task. In this way, the values of attributes determine the rules that are applicable to resolve a task. That is, rules that are applicable at a bud.

Below is a sample grammar that models the beginning of the alarm investigation process described in Section 3. The grammar depicts a service offered by an epidemiologist. We have written in **bold** names of external sorts that make reference to other grammars in distant sites. These sorts therefore have no defining rules in this grammar. The rules are labeled **R1** to **R3** with **R1** and **R3** having parameters which will have to be filled in by the epidemiologist during execution. These parameters indicate the effective users in whose workspaces the external service requests will be made. In **R2** and **R3**, we introduced guards $FALSE$ and $TRUE$. These guards automatically filter which of the two rules with sort *continue* to apply when the first task terminates. If the alarm is seen to be genuine, $TRUE$, the epidemiologist contacts a clinician sending the alarm information and a set of requests $Todos$. The result returned by this external service request is used to run additional analysis *runAnalysis* to confirm or revoke the alarm. Note that when **R3** is applied for instance, all its subtasks become buds (ready for execution). However, the *runAnalysis* bud will have to wait for the other task to complete due to variable dependences. This shows that though no predefined ordering exists between subtasks, an ordering can be introduced using variables synthesized within the subtree.

R1:

$\mathbf{investigateAlarm}(Alarm) =$
$\mathbf{input}(pha)$
$\mathbf{do}$
$(real) \leftarrow \mathbf{contactPharm}[pha](Alarm)$
$(results) \leftarrow continue(real, Alarm)$
$\mathbf{return}\ (results)$

R2:

$continue(FALSE, Alarm) =$
$\mathbf{return}\ (False\_Alarm)$

R3:

$continue(TRUE, Alarm) =$
$\mathbf{input}(cli)$
$\mathbf{do}$
$(lab\_res, Patient\_data) \leftarrow$
$\mathbf{contactClinician}[cli](Alarm, Todos)$
$(analysis\_res) \leftarrow$
$runAnalysis(lab\_res, Patient\_data)$
$\mathbf{return}\ (analysis\_res)$

In some situations it is necessary that semantic rules are not given by plain terms but by more general functional expressions. This is in particular the case when one invokes a service to all individuals playing some particular role. For instance assume that the right-hand side of a rule contains a call of the form $(y) \leftarrow s[r :: R](x, y[r :: R])$. Then each individual playing role R must resolve task s to produce a synthesized result y using an inherited attribute x as well as the values y synthesized by other users of the same role. This means that to produce his results, each user uses the results produced by other users. Now, a variable y synthesized by a sort $s[r :: R]$ can only be used elsewhere in the form $y[r :: R]$, that is, a vector indexed by elements of R. Such vectors of variables cannot be used directly within terms. One might add projections to extract the variable associated with a given individual $r :: R$, which we would write $y[r]$. But in general we are not aware of a particular individual in a given role (and moreover as noted before this set of individuals can vary in time) and one is rather interested in stating conditions such as "there exist $r :: R$ such that $y[r]$" or "for all $r :: R$, $y[r]$", or even "there exist at least 3 individuals $r :: R$ such that $y[r]$", "at least 50% of $r :: R$ verify $y[r]$" etc. when variable y holds a boolean value. More generally one can express the semantic rules using any kind of functional expressions as long as the values of inherited attributes evaluate to terms so that they can be matched against patterns.

For example in scenario described in section 3, when a laboratory test is sought from several biologists (call them pete, bob, john, ...), and they need to each return a lab test result, $labTR$, the value returned by each of them is accessed as follows: $labtTR[pete :: biologist]$, $labtTR[bob :: biologist]$, ... It is also possible to use these in conditions like "there exist $labTR[r :: biologist]$" which checks if there exist any biologist who as already provided a result, "at least 3 $labTR[r :: biologist]$" which asserts that at least three biologist have provided results for the lab test, etc. These conditions coupled with terms are useful to drive the application of other rules at buds.

4.4 Temporal Dependencies and Constraints

Time is a critical and determining factor in user-satisfaction, cost reduction and increased productivity in business processes. In disease surveillance in particular,

timeliness is a major metric used to assert and/or evaluate the effectiveness and relevance of the process. Due to space constraints and given the extensiveness of this topic, we only present high level temporal constraints and dependencies.

We add a time-dimension to the Active-Workspaces model using the concepts defined in [34] and [33] based on Allen's Intervel Algebra[35]. These works identify the following intuitive temporal constraints: *Must Start On (MSO)*, *Must Finish On (MFO)*, *As Soon As Possible (ASAP)*, *As Late As Possible (ALAP)*, *No Earlier Than (NET)* and *No Later Than (NLT)*. These constraints are attached to tasks at specification time and are used by a scheduler to control task start and end times. The specifications of the scheduler are beyond the scope of this paper. All constraints for subtasks are defined and interpreted relative to some reference point, usually the start and end times of the parent task or of sibling tasks. For instance, if data collection and data analysis are subtasks of the disease surveillance task, both subtasks can be defined to start ASAP, but the constraint on the collection task interpreted relative to the parent task and that on the analysis task interpreted relative to the collection task. The MSO and MFO constraints are strict and force the task to start or stop at exactly some time-point from the reference time. If no constraint is specified for a task, it is assumed that the task starts ASAP and finishes ASAP. Such a task is immediately executed when all necessary inputs become available and finishes as soon as all computations complete.

Also, based on the temporal constraints that exist between tasks and their data dependencies, we deduce temporal dependencies between tasks within a business rule. By temporal dependency, we mean any relationship between two tasks in which the start or end of one depends on the start or end of the other. The following four temporal dependency relationships are possible: *Start to Finish (SF)*, *Start to Start (SS)*, *Finish to Start (FS)*, and *Finish to Finish (FF)*. For instance, the data collection and data analysis tasks described in the previous paragraph have an SS relationship written $SS(data\ collection, data\ analysis)$ meaning that data analysis cannot start until data collection has started. In like manner, an SF, FS, or FF relationship between two tasks S_1 and S_2 respectively means that S_2 cannot finish until S_1 starts, S_2 cannot start until S_1 finishes, and S_2 cannot finish until S_1 finishes.

Lastly, additional temporal components, Lag-Time and Lead-Time can be added to temporal dependencies to respectively account for waiting times between tasks and for overlapping tasks.

5. Workspace Evolution

The Active-Workspaces model is adapted for "Open Systems" in which the actions of users are not explicitly specified at design time. These systems are distributed and evolve dynamically with users playing a primordial role. They need to continuously design and run parts of a business process and collaborate with each other. Even when task specifications exist, the effective actions a user undertakes (deciding which task to run, providing input data, etc.) are not specified in advance.

In section 4 we presented two ways in which a workspace can evolve. A user can either explicitly add new productions to an existing grammar or obtain productions defined by another user when their workspaces are synchronized. Also, as noted in [1], the process always interacts with external tools such as databases, email systems, time servers, etc. the so-called side-effects. These external systems complement the active-workspaces model. They allow that real world activities like extracting samples from patients, sending messages, etc. be associated to a rules. Also, these external tools can be used to extract information from enacted artifacts to build dashboards or to feed some local database that are later used to guide the user on her choice of the rule to apply for a pending task. They may, in a more coercive fashion, suggest a specific rule to apply or even inhibit some of the rules. Some information from dashboards or contained in a local database can also be used to populate some input parameters of a rule in place of the user.

6. Conclusion and Future work

In this paper, we characterized the process of monitoring diseases and health conditions of interest for unwanted events as being user-driven and data-centric. The unpredictable nature of the process further justified its knowledge-intensive characteristic. We identified four major modeling use cases that should be fulfilled by disease surveillance process modelers. The active-workspaces model can be extended to offer all four use cases. In this paper, we explicitly present how the active-workspaces model can address the first two use cases namely: dynamic workflow construction and execution, and user interactions. A prototype for the Active Workspaces model which is currently under construction will further demonstrate its pertinence and applicability.

Due to space constraints, we left out certain aspects of the Active Workspaces model which of course we will gladly add in an extended version if this paper is considered for publication. These aspects include:

- The architecture of an Active Workspace: this comprises, the underlying Guarded Attribute Grammar (GAG) engine; the Active Workspaces server that manages users and roles (adding/removing users and/or roles, subscribing/un-subscribing a user from a role), and managing communication between workspaces.
- The user interface: with visualizations of artifact trees, interfacing with external tools, enacting workflows, etc.
- GAG specifications of the key steps in the scenarios described in this paper.
- The extensive formal specification of temporal constraints and dependencies.

We are currently extending the Active Workspaces model to integrate external support for workspace construction using data mining techniques, process mining techniques, and connecting the model with a disease surveillance knowledge base. These will be necessary to demonstrate how the Active Workspaces model can be used to manage uncertainty and effective decision making, two of the use cases identified at the beginning of this paper.

Acknowledgement:
Research by the author NSAIBIRNI Robert FONDZE Jr leading to this result has been partially funded by the US Department of Health and Human Services (DHHS) through the ASIDE PROJECT (www.asideproject.org) pioneered by the Institut Pasteur of Paris and the Centre Pasteur du Cameroun.

References

[1] Eric Badouel, Loïc Hélouët, Georges-Edouard Kouamou, Christophe Morvan, and Robert Fondze Jr Nsaibirni. Active Workspaces : Distributed Collaborative Systems based on Guarded Attribute Grammars. *ACM SIGAPP Applied Computing Review* **15**(3): 6–34, 2015.

[2] Claudio Di Ciccio, Andrea Marrella, and Alessandro Russo. Knowledge-Intensive Processes: Characteristics, Requirements and Analysis of Contemporary Approaches. *Journal on Data Semantics*, pages 29–57, 2014.

[3] R. Hull, E. Damaggio, F. Fournier, M. Gupta, Fenno Terry Heath, S. Hobson, M. H Linehan, S. Maradugu, A. Nigam, P. Sukaviriya, and R. Vaculín. Introducing the Guard-Stage-Milestone Approach for Specifying Business Entity Lifecycles. In *Web Services and Formal Methods - 7th International Workshop, WS-FM 2010, Hoboken, NJ, USA*, volume 6551 of *Lecture Notes in Computer Science*, pages 1–24. Springer, 2011.

[4] Kunzle V, Reichert M PHILharmonicFlows: towards a framework for object-aware process management *Journal of Software Maintenance and Evolution: Research and Practice*,2011

[5] M.M. Wagner, L.S. Gresham, and V. Dato. Chapter 3 - case detection, outbreak detection, and outbreak characterization. In M.M. Wagner, A.W. Moore, and R.M. Aryel, editors, *Handbook of Biosurveillance*, pages 27 – 50. Academic Press, Burlington, 2006.

[6] International Society for Disease Surveillance. Final Recommendation: Core Processes and EHR Requirements for Public Health Syndromic Surveillance. Technical report, ISDS, 2011.

[7] R. Hull, E. Damaggio, and R. De Masellis. Business artifacts with guard-stage-milestone lifecycles: managing artifact interactions with conditions and events. ... *on Distributed event-* ..., pages 51–62, 2011.

[8] Centers For Disease Control World Health Organization. Technical Guidelines for Intergrated Disease Surveillance and Response in the African Region. *Technical report, WHO/CDC, Georgia, USA* 2001.

[9] Andrea Marrella, Massimo Mecella, Sebastian Sardina SmartPM: An Adaptive Process Management System through Situation Calculus, IndiGolog, and Classical Planning *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference, {KR}* 2014, Vienna, Austria, July 20-24, 2014

[10] Roger Atsa Etoundi, Marcel Fouda Ndjodo, and Ghislain Abessolo Aloo. A Formal Framework for Business Process Modeling. *International Journal of Computer Applications*, 13(6):27–32, 2011.

[11] Claudio Di Ciccio, Andrea Marrella, and Alessandro Russo. Knowledge-intensive Processes: An overview of contemporary approaches. *CEUR Workshop Proceedings*, 861:33–47, 2012.

[12] Reichert M, Rinderle S, Kreher U, Dadam P Adaptive Process Management with ADEPT2 *ICDE*, 2005

[13] ter Hofstede AHM, van der Aalst WMP, Adams M, Russell N Modern Business Process Automation: YAWL and its Support Environment. *Springer*, 2009

[14] Object Management Group Omg. Business Process Model and Notation V2.0. Technical Report December, 2010.

[15] Roman Vaculín, Richard Hull, Terry Heath, Craig Cochran, Anil Nigam, and Piyawadee Sukaviriya. Declarative business artifact centric modeling of decision and knowledge intensive business processes. In *Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOC*, number Edoc, pages 151–160, 2011.

[16] Wil M. P. van der Aalst. Business Process Management: A Comprehensive Survey. *ISRN Software Engineering*, 2013:1–37, 2013.

[17] G Texier and Y Buisson. From epidemic outbreak detection to anticipation. *Revue d'Epidemiologie et de Sante Publique*, 2010.

[18] W. M. P. Van Der Aalst. the Application of Petri Nets To Workflow Management. *Journal of Circuits, Systems and Computers*, 08(01):21–66, 1998.

[19] Clementine Calba, Flavie L Goutard, Linda Hoinville, Pascal Hendrikx, Ann Lindberg, Claude Saegerman, and Marisa Peyre. Surveillance systems evaluation: a systematic review of the existing approaches. *BMC public health*, 15:448, 2015.

[20] W. M P Van Der Aalst and a. H M Ter Hofstede. YAWL: Yet another workflow language. *Information Systems*, 30(4):245–275, 2005.

[21] W.M.P. van der Aalst. Formalization and verification of event-driven process chains. *Information and Software Technology*, 41(10):639–650, 1999.

[22] W.M.P. van der Aalst, R.S. Mans, and N.C. Russell. Workflow Support Using Proclets: Divide, Interact, and Conquer. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 2009.

[23] W Yasnoff, P W O'Carroll, D Koo, R W Linkins, and E M Kilbourne. Public health informatics: improving and transforming public health in the information age. *Journal of public health management and practice : JPHMP*, 6(6):67–75, 2000.

[24] Kamal Bhattacharya, Cagdas Gerede, Richard Hull, Rong Liu, and Jianwen Su. Towards Formal Analysis of Artifact-Centric Business Process Models. *Alonso, G., Dadam, P., Rosemann, M. (eds.) BPM2007. LNCS*, 4714:288–304, 2007.

[25] David Cohn and Richard Hull. Business Artifacts : A Data-centric Approach to Modeling Business Operations and Processes. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 32(3):1–7, 2009.

[26] Hervé Chaudet, Pellegrin Liliane, Jean-Baptiste Meynard, Gaëtan Texier, Olivier Tournebize, Benjamin Queyriaux, and Jean-Paul Boutin. Web Services Based Syndromic Surveillance for Early Warning within French Forces. *Studies in health technology and informatics (2006)*, 124:666–671, 2006.

[27] H P Lehmann, B E Dixon, and H Kharrazi. Public Health and Epidemiology Informatics: Recent Research and Trends in the United States. *Yearbook of medical informatics*, 10(1):199–206, 2015.

[28] Liliane Pellegrin, Charlotte Gaudin, Nathalie Bonnardel, and Hervé Chaudet. Collaborative activities during an outbreak early warning assisted by a decision-supported system (ASTER). *Int. J. Hum. Comput. Interaction*, 26(2&3):262–277, 2010.

[29] Daniel Zeng, Hsinchun Chen, Carlos Castillo-Chavez, and William B. Lober. *Infectious Disease Informatics and Biosurveillance*, volume 28. 2012.

[30] Nkuchia M. M'ikanatha and John K. Iskander. Concepts and Methods in Infectious Disease Surveillance. *John Wiley & Sons, Ltd, 2014*.

[31] Brian E Dixon, Roland E Gamache, and Shaun J Grannis. Towards public health decision support: a systematic review of bidirectional communication approaches. *Journal of the American Medical Informatics Association : JAMIA*, 20(3):577–83, 2013.

[32] Zaruhi R Mnatsakanyan and Joseph S Lombardo. Decision Support Models for Public Health Informatics. *John Hopkins APL Technican Digest*, 27(4):332–339, 2008.

[33] Denis Gagne and André Trudel. *Fisher, L. (Ed), 2008 BPM and Workflow Handbook*, The Temporal Perspective: Expressing Temporal Constraints and Dependencies in Process Models, pages 247-260.

[34] Denis Gagne and André Trudel. Time-bpmn. In *Proceedings of the 2009 IEEE Conference on Commerce and Enterprise Computing*, CEC '09, pages 361–367, Washington, DC, USA, 2009. IEEE Computer Society.

[35] James F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, November 1983.

What Motivates High School Students to Take Precautions against the Spread of Influenza? A Data Science Approach to Latent Modeling of Compliance with Preventative Practice

*[1a] William L. Romine, [2b]Tanvi Banerjee, [3c]William R. Folk, [4d]Lloyd H. Barrow

[a]Dept. of Biological Sciences, [b]Dept. of Computer Science and Engineering, Wright State University, 3640 Colonel Glenn Hwy, Dayton OH 45435

[c]Dept. of Biochemistry, [d]Dept. of Learning, Teaching, and Curriculum, University of Missouri, Columbia, MO 65211

[1]romine.william@gmail.com, [2]tanvi@knoesis.org, [3]folkw@missouri.edu, [4]barrowl@missouri.edu

Abstract – *This study focuses on a central question: What key behavioral factors influence high school students' compliance with preventative measures against the transmission of influenza? We use multi-level logistic regression to equate logit measures for eight precautions to students' latent compliance levels on a common scale. Using linear regression, we explore the efficacy of knowledge of influenza, affective perceptions about influenza and its prevention, prior illness, and gender in predicting compliance. Hand washing and respiratory etiquette are the easiest precautions for students, and hand sanitizer use and keeping the hands away from the face are the most difficult. Perceptions of barriers against taking precautions and sense of social responsibility had the greatest influence on compliance.*

Keywords: influenza mitigation, multilevel logistic regression, health informatics, quantitative analysis, decision support system

1 Introduction

In the United States, influenza imposes a heavy cost to our health and financial wellbeing, accounting for over 100,000 deaths and 1.7 million hospital stays over 10 influenza seasons (1999-2009) [1]. Influenza can result in medical costs of approximately $10 billion, lost earnings of $16 billion annually, and a total economic burden of $87 billion [2]. Given that approximately 10% of our school children contract influenza each year [3], influenza impedes education. Students missing school due to illness [4] results in reduced learning [5], free or reduced lunch benefit [4], parents missing work for childcare [6], and delinquency when children go unsupervised [7].

Both pharmaceutical (e.g. stockpiling vaccines and antivirals) and non-pharmaceutical options (e.g. quarantine and school closure) have been considered for managing severe influenza epidemics and pandemics [8, 9]. While these can be effective in reducing the spread of influenza, they can be socially intrusive and economically expensive [8, 9].

The motivation for this study comes from the hypothesis that educational or behavioral interventions focused on increasing compliance with preventative measures are an economical and effective way to reduce the spread of influenza. The central question of this study is: *What key cognitive and behavioral factors influence high school students' compliance with preventative measures against influenza transmission?* In addressing this question, we focus on two sub-questions: (1) what hierarchy exists in students' compliance with recommended precautions for preventing the spread of influenza, and (2) what is the efficacy of four variables: (1) students' knowledge of influenza, (2) affective perceptions of influenza and its prevention, (3) prior illness, and (4) gender in predicting students' compliance? We explore the relationship between these variables and compliance in a data driven approach that can improve targeted interventions supporting influenza management in schools.

2 Related Work

Multiple studies have suggested that cognitive, affective, and demographic factors may lead to compliance with measures to prevent the spread of influenza. However, a majority of these studies have targeted compliance with vaccination [10-14] and hand washing [14, 15] exclusively. Knowledge of influenza was found to increase vaccination rates in nurses [10, 11] and parents of school children [12]. A positive increase in vaccination rates in relation to perceived risk of influenza [13] and perceived complications of influenza [14] was found in university students and employees, and nurses, respectively. Ethnicity [14] and gender [13] were also found to impact compliance with vaccination. Barriers against compliance with vaccination include concerns over contracting the flu from vaccination, belief that vaccination is not effective, aversion to needles, and belief that influenza does not pose a significant health risk [10].

Findings regarding compliance with handwashing bear similarity to those for vaccination. Improved compliance with hand washing among hospital nurses was promoted by posters describing how infection is transmitted by the hands [15]. A positive relationship between knowledge of influenza and compliance with hand washing was also found in high school students [16]. Perceived barriers such as skin irritation, inconvenience, wearing gloves, and absent-mindedness were shown to impede compliance with hand washing across multiple populations [15-17]. Females were found to exhibit higher compliance with hand washing than males [17].

Studies addressing precautions against flu transmission as a holistic construct targeted high school students [16, 18] and the general public [19]. Using separate logistic regression models, these studies found that, along with vaccination and hand washing, perceived severity of influenza was a predictor for social distancing, and perceived efficacy was a positive predictor for all precautions. Other elements of hygiene such as respiratory etiquette and keeping hands away from the face were positively related to knowledge of influenza in high school students [17]. Perceived complications from influenza also played a positive role in students' decisions to stay home when sick and stay away from peers who were visibly sick [17].

While survey methods have been used in prior research of compliance with measures to prevent spread of influenza [17-19], we know of no research on measuring compliance with preventative behavior as a latent variable. Looking at multiple preventative measures hierarchically using a common scale, as opposed to looking at one precaution at a time, is essential in obtaining a more holistic understanding of how knowledge of influenza and affective perceptions influence compliance with preventative measures. Such analyses can help health education specialists target specific factors to improve student understanding and help reduce the transmission of influenza. In the next section, we describe our data-driven approach to develop a decision support system aimed at understanding and improving student compliance with influenza mitigation behaviors.

3 Experiments and Analysis

3.1 Conceptual Framework

Cognitive, affective, and demographic factors were related to compliance with behaviors to prevent influenza transmission using data described in [16, 18]. These variables were measured using the Student Influenza Survey described in our earlier work [18]. Cognition was qualified in terms of knowledge of influenza [18, 20]. Affective variables were derived from the Health Belief Model [21], which suggests that health behaviors are guided by perceptions of risk of contracting influenza, perceived severity of complications from influenza, barriers against taking preventative measures, and sense of social responsibility [21]. These variables were normalized to a scale of standard deviations centered at zero. Gender was the only demographic factor explored in this study given its documented importance [13]. Despite some literature pointing to the importance of ethnicity [14], the ethnicity distribution in our sample (described in Section 3.2) was not sufficiently diverse to warrant statistical exploration. Eight precautions (right side of Figure 1) were measured on an ordinal 1-5 scale, where a value of 1 indicates complete neglect of the precaution, and a value of 5 indicates frequent, accepted practice.

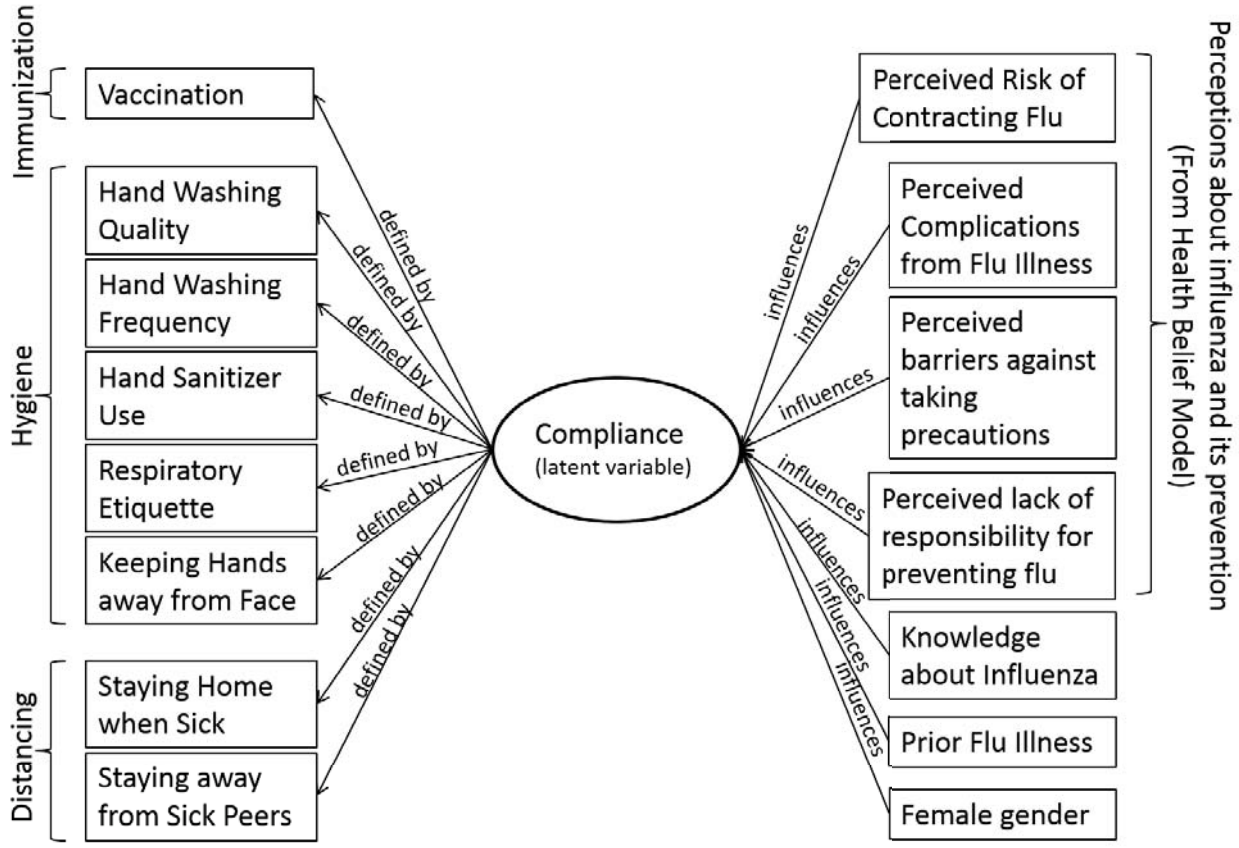


Figure 1. Conceptual framework of our decision system which models compliance with influenza prevention as a latent variable. Our latent variable of compliance is defined by immunization, five hygiene behaviors, and two distancing behaviors. We hypothesize that four perceptions of influenza derived from the Health Belief Model [21], knowledge about influenza [16, 18], gender [13], and prior flu illness influence high school students' compliance with measures to prevent the spread of influenza.

Middle (2, 3, and 4) levels represent a monotonic gradation between the lowest and highest levels. As an example, when asked about vaccination, the value 1 in the scale aligns with the students' report that they never get vaccinated for influenza. The value 5 indicates that the students get vaccinated against influenza every year. In the next subsection, we discuss our data collection process and the statistical methods used to understand the data.

3.2 Data Collection and Statistical Models

The Student Influenza Survey was administered to a sample of 410 students enrolled in grades 9-12 (median age of 16 years) from five school districts. Of the 375 students reporting their gender, 169 were male and 206 were female. A majority reported White ethnicity (n = 266). However, Black (n = 50), Asian (n = 27), and Hispanic (n = 22) ethnicities were also reported in the sample.

Multi-level logistic regression modeling using BIGSTEPS [23] was used to equate students' compliance with precautions and the difficulty of individual precautions on a common logit (log-odds) scale. This was specified as an ordinal random intercept model where students were modeled as the random factor and the eight precautions were treated as fixed factors. Important advantages of using multilevel logistic regression include: (1) ability to equate student compliance measures and precaution difficulty measures on a common logit scale; and (2) ability to obtain student compliance measures that are survey-independent and precaution difficulty measures that are student-independent [22]. Using linear regression, students' logit measures for compliance along the latent scale were equated to (1) knowledge of influenza, (2) perceived risk of contracting influenza, (3) perceived complications from the flu illness, (4) students' perceived barriers getting in the way of complying with precautions, (5) lack of perceived social responsibility (or inefficacy),

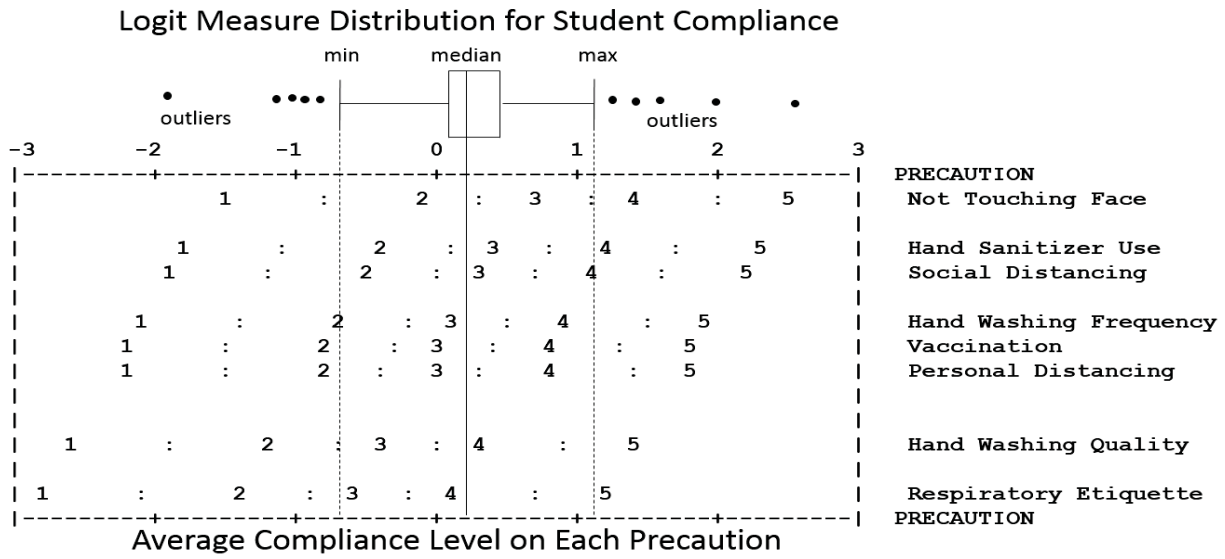


Figure 2. Expected average ordinal compliance level (1=lowest; 5=highest) on the eight precautions plotted against students' logit scale measures for the latent variable. On average, students at the median are expected to exhibit a compliance level between a 2 and a 3 for "not touching the face." The same students are expected to have an average compliance level just above a 4 for "respiratory etiquette." The Student Influenza Survey and detailed descriptions of response levels is available in [18].

(6) flu illness the prior year, and (7) gender. Statistical significance of these effects on compliance with precautions was evaluated at the 95% confidence level.

3.3 Formulating a Compliance Scale

The pattern of student responses for each precaution fit well with the multi-level ordinal random intercept logistic regression model, with normed chi-square values between 0.66 and 1.42. This illustrates the efficacy of the eight precautions in positioning students along a common latent compliance metric (Figures 1 and 2) [24]. Principal components analysis was implemented on the model residuals, revealing a first eigenvalue of 1.37 items of variance. This indicated that the systematic variance in the model was sufficient to explain the data [25], and that the remaining variance was random noise. Figure 2 displays the student measure distribution along the logit scale (the box-whisker plot at the top of Figure 2), and displays the average compliance level on each precaution that we would expect from students of a particular logit measure along the hierarchy of preventative behaviors. The behaviors are ranked in an increasing order such that those at the top of Figure 2 are the most difficult for students to comply with, and those at the bottom are easiest. As an example, respiratory etiquette and quality hand washing appear to garner the greatest compliance levels from students. Students at the median (the middle vertical line labeled "median" in Figure 2) are expected to exhibit behavior levels around 4, indicating that these students wash their hands for a duration with soap and water, and that they use a fabric or tissue in which to cough or sneeze (we encourage the reader to consult the Student Influenza Survey in [18] for additional qualitative details on the 5-level ordinal scale for the eight precautions). The students at the top of the scale (the right vertical line labeled "max" in Figure 2) approach a 5 level, indicating that they wash with soap and water for at least 30 seconds and use their sleeve or a tissue (which is then thrown away) when they cough or sneeze. As indicated by Figure 2, the following precautions at the top of the scale: keeping hands away from the face, and hand sanitizer use are the most difficult precautions for students to practice. Students at the median exhibit compliance levels below 3 on these behaviors indicating that they touch their eyes, nose, and mouth with their hands multiple times per day and use alcohol-based hand sanitizers (when available) around once per day. Students at the top of the scale have an expected level of 4 for these behaviors indicating that they touch their eyes, nose, and mouth only a few times within a 1-week time frame, and use hand sanitizers more than once daily. It is interesting that vaccination, which has perhaps received the greatest attention as a measure to prevent

seasonal epidemics and pandemics in schools [26, 1], sits at the middle of the scale. Students in the middle of the latent compliance scale comply with vaccination at a 3 level, meaning that he/she has taken the vaccination before, but does not plan on taking it in the current flu season. Students at the top of the scale exceed a 4 level on average, indicating that they plan on receiving the vaccination during the current flu season. While it is encouraging that vaccination is not among the most difficult precautions for students to take, this finding shows that efforts are needed to make compliance with vaccination easier for high school students.

Other behaviors of moderate difficulty include frequent hand washing and distancing behaviors such as staying home when sick (personal distancing) and staying away from peers who are visibly sick (social distancing). Students in the middle of the compliance scale are expected to behave at a 3 level, indicating willingness to wash hands 3-4 times a day and keep distance from sick peers. These students, however, generally attend school if they consider their symptoms to be minor. Students at the top of the scale, however, behave at a 4 level on these precautions, indicating that they wash their hands 5-6 times per day and will request to the teacher that they be moved if a visibly sick student is sitting near them. These students indicate willingness to stay home when they are sick as long as they do not have an important school engagement such as an exam.

3.4 Factors Influencing Compliance

We now explore factors which influence students' latent compliance levels in our decision support model (right side of Figure 1). These factors include student knowledge of influenza, perceptions of influenza and its prevention, gender, and prior flu illness. The linear regression decision model (shown in Table 1) for the latent outcome of compliance was statistically significant ($F_{7,343} = 13.6$, $p << 0.001$, $r^2_{adj} = 0.20$) indicating that this models the data significantly better than simply calculating students' mean latent compliance level. The variables in the model collectively explain 20% of the variation in the latent variable indicating potential efficacy as a decision support model. Overwhelmingly the most significant predictors of compliance are students' perceptions of barriers against taking effective preventative behaviors and inefficacy (lack of perceived social responsibility for taking appropriate precautions against the spread of influenza).

Table 1. Predictors for high school students' positions along the latent scale for compliance with preventative measures against the spread of influenza.

Predictor	Coef.	SE	T	partial r^2
Risk	0.088	0.051	1.730	0.009
Complications	0.056	0.050	1.110	0.004
Barriers	-0.404	0.061	-6.640*	0.114
Inefficacy	-0.228	0.042	-5.390*	0.078
Flu Illness	0.001	0.060	0.020	0.000
Female	0.117	0.044	2.650*	0.020
Flu Knowledge	-0.041	0.037	-1.110	0.004
Intercept	0.076	0.043	1.780	0.009

*Significant at 95% confidence level

An example of perceived barriers includes the belief that the flu vaccination is harmful, ineffective, or difficult to obtain. Other examples include difficulty or unavailability of access to hand sanitizers, or belief that it is socially difficult to distance oneself from peers who are sick.

Indicators of inefficacy include the beliefs that: (1) one has no control over contracting influenza, (2) becoming sick will not affect schoolwork, and (3) taking precautions does not affect social relationships with friends and teachers. A one logit increase in perceived barriers and inefficacy leads to a decrease of 0.4 and 0.23 logits along the compliance scale (Figure 2), respectively. Taken together, a one logit increase in these factors could lead to over one half of a logit decrease in a student's latent compliance level. Further, a less responsible student who finds taking precautions difficult may be over one logit lower on the compliance scale than a more responsible student who takes precautions as part of his/her routine. This could make a difference of over one level (for example, a 3 level instead of a 4 level) along the ordinal scale for individual preventative behaviors (Figure 2).

The model also shows that females have an average level of compliance approximately 0.12 logits higher than males. While this difference is statistically significant, Figure 2 shows that gender is unlikely to make a major difference in students' individual preventative behaviors except in borderline cases. Although gender does not have the practical significance of perceived barriers or inefficacy, it nonetheless suggests that targeting of males could be one element of a successful behavioral intervention aiming to improve flu prevention in high school students.

4 Conclusion

We develop and implement a decision system that: (1) explains compliance with behaviors which mitigate spread of influenza as a continuous latent variable composed of eight accepted preventative practices, and (2) uses gender and two affective variables to predict students' levels of compliance. While measures like quarantine, mass vaccination campaigns, and school closure have been invoked in the context of severe seasonal epidemics and pandemics such as the relatively recent H1N1 Swine Flu pandemic [27], these measures can be expensive and difficult to implement for relatively frequently occurring diseases like influenza—behavioral interventions could serve as a supplement.

Our analyses suggest that it is feasible to develop behavioral interventions which encourage students to take precautionary measures such as immunization, hygiene, and distancing. From our model, we find that the most effective interventions should address the students' barriers in taking preventative measures. An effective program will educate them about the availability of the flu vaccine and resources for hygiene such as tissues and hand washing facilities. Students also should be made aware of the significance of taking preventative measures on a civic level—while people are generally aware that these behaviors can protect themselves against the flu, they are often unaware that efforts to take preventative measures also protect others [28].

Finally, our analysis demonstrates the importance of school policies in preventing the spread of influenza, and perhaps other viruses such as measles and zika, among students and school staff. Implementation of school vaccination programs would eliminate students' perception that the vaccine is expensive or unavailable. Installation of alcohol-based hand sanitizers would have a similar effect, making hand hygiene more accessible for students. Policies encouraging social and personal distancing may serve to mitigate outbreaks by reducing influenza transmission if infected students choose to stay home when they are sick. These may involve the decisions to: (1) not reward perfect attendance, especially during flu season; (2) encourage teachers to develop alternative assignments for students who miss class due to illness, and (3) not penalize students for missing class examinations, or providing opportunities to make up for missed examinations during the flu season.

Using a data driven approach to understand high school students' compliance with precautions against influenza transmission, our analysis suggests that schools have many options for improving practices to prevent the spread of infectious diseases like influenza. Schools which deliver effective educational programs and prevention-friendly policies will give students a sense of control over their own health and the health of their peers, and will likely reap the reward of reduced illness and absenteeism during the flu season. These benefits are likely to be transferable to other diseases which spread in a similar manner as influenza.

5 References

[1] Reed, C., Meltzer, M. I., Finelli, L., & Fiore, A. (2012). Public health impact of including two lineages of influenza B in a quadrivalent seasonal influenza vaccine. *Vaccine*, *30*(11), 1993-1998.

[2] Molinari, N., Ortega-Sanchez, I., Messionnier, M., Thompson, W., Wortley, P., Weintraub, E., & Bridges, C. (2007). The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine, 25*(27), 5086-5096.

[3] Principi, N., Esposito, S., Marchisio, P., Gasparini, R., & Crovari, P. (2003). Socioeconomic impact of influenza on healthy children and their families. *Pediatric Infectious Disease Journal, 22*(10), S207-S210.

[4] Wong K. K., Shi J., Gao H., et al. (2014) Why Is School Closed Today? Unplanned K-12 School Closures in the United States, 2011–2013. *PLoS ONE 9(12),* e113755.doi:10.1371/journal.pone.0113755

[5] Marcotte, D. E., & Hansen, B. (2010). Time for school. *Education Next*, *10*(1), 52-59.

[6] Sadique, M. Z., Adams, E. J., & Edmunds, W. J. (2008). Estimating the costs of school closure for mitigating an influenza pandemic. *BMC Public Health*, *8*(1), 135.

[7] Cauchemez, S., Ferguson, N., Wachtel, C., Tegnell, A., Saour, G., Duncan, B. and Nicoll, A. (2009). Closure of schools during an influenza pandemic. *Lancet Infect. Dis., 9*, 473-481.

[8] Cauchemez, S., Valleron, A. J., Boelle, P. Y., Flahault, A., & Ferguson, N. M. (2008). Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature*, *452*(7188), 750-754.

[9] Ferguson, N. M., Cummings, D. A., Fraser, C., Cajka, J. C., Cooley, P. C., & Burke, D. S. (2006). Strategies for mitigating an influenza pandemic. *Nature*, *442*(7101), 448-452.

[10] Martinello, R. A., Jones, L., Topal, J. E. (2003). Correlation between healthcare workers' knowledge of influenza vaccine and vaccine receipt. *Infection Control and Hospital Epidemiology, 24,* 845-847.

[11] Falomir-Pichastor, J., Toscani, L., & Despointes, S. (2009). Determinants of flu vaccination among nurses: The effects of group identification and personal responsibility. *Applied Psychology: An International Review, 58*(1), 42–58.

[12] Joshi, A., Lichenstein, L., King, J., Arora, M., & Khan, S. (2009). Evaluation of a computer-based patient education and motivation tool on knowledge, attitudes, and practice towards influenza vaccination. *International Electronic Journal of Health Education, 12*, 1-15.

[13] Weinstein, N. D., McCaul, K., Gerrard, M., Gibbons, F. X. Kwitel, A., & Magnan, R. (2004). Risk perceptions: Assessment and relationship to influenza vaccination. *Health Psychology, 26*(2), 146-151.

[14] Chen, J. Fox, S., Cantrell, C., Stockdale, S., & Kagawa-Singer, M. (2006). Health disparities and prevention: Racial/ethnic barriers to flu vaccinations. *Journal of Community Health, 32*(1), 5-20.

[15] Pittet, D. (2000). Improving compliance with hand hygiene in hospitals. *Infection Control and Hospital Epidemiology, 21*, 381–386.

[16] Romine, W. L., Banerjee, T., Barrow, L. H., & Folk, W. R. (2012). Exploring the impact of knowledge and social environment on influenza prevention and transmission in Midwestern United States high school students. *European Journal of Health and Biology Education,1*(1), 75-115.

[17] Kretzer, E. K., & Larson, E. L. (1998). Behavioural interventions to improve infection control practices. *Am J Infect Control, 26,* 245–253.

[18] Romine, W. (2011). *Development and validation of two influenza assessments: Exploring the impact of knowledge and social environment on health behaviors* (Doctoral dissertation, University of Missouri--Columbia).

[19] Kiviniemi, M., Ram, P., Kozlowski, L., & Smith, K. (2011). Perceptions of and willingness to engage in public health precautions to prevent 2009 H1N1 transmission. *BMC Public Health, 11,* 152.

[20] Romine, W. L., Barrow, L. H., & Folk, W. R. (2013). Exploring secondary students' knowledge and misconceptions about influenza: Development, validation, and implementation of a multiple-choice influenza knowledge scale. *International Journal of Science Education*, *35*(11), 1874-1901.

[21] Rosenstock, I. (1974). Historical origins of the Health Belief Model. *Health Education Monographs, 2*, 328-335.

[22] De Ayala, R. J. (2013). *The Theory and Practice of Item Response Theory*. Guilford Publications: New York.

[23] Linacre, J. M., & Wright, B. D. (1993). *A user's guide to BIGSTEPS: Rasch-model computer Program*. Mesa Press: Chicago.

[24] Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

[25] Linacre, J. M., & Tennant, A. (2009). More about critical eigenvalue sizes in standardized-residual principal components analysis (PCA). *Rasch Measurement Transactions*, *23*(3), 1228.

[26] Carpenter, L. R., Lott, J., Lawson, B. M., Hall, S., Craig, A. S., Schaffner, W., & Jones, T. F. (2007). Mass distribution of free, intranasally administered influenza vaccine in a public school system. *Pediatrics*, *120*(1), e172-e178.

[27] Coburn, B. J., Wagner, B. G., & Blower, S. (2009). Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1). *BMC Medicine*, *7*(1), 30.

[28] Stinchfield, P. (2008). Practice-proven interventions to increase vaccination rates and broaden the immunization season. *The American Journal of Medicine, 121,* S11-S21.

Probabilistic Analysis of Contracting Ebola Virus Using Contextual Intelligence

Arjun Gopalakrishnan and Krishna Kavi
Department of Computer Science, University of North Texas, USA

Abstract

The West African countries witnessed an "extraordinary" outbreak of the Ebola virus in August 2014. It was declared to be a Public Health Emergency of International Concern (PHEIC) by the World Health Organization (WHO). Due to the complex nature of the outbreak, Centers for Disease Control and Prevention (CDC) has created interim guidance for monitoring people potentially exposed to Ebola and for evaluating their intended travel and restricting the movements of carriers when needed. Tools to evaluate the risk of individuals and groups of individuals contracting the disease could mitigate the fear and anxiety. Our goal is to understand and analyze the nature of risk an individual would posses when he/she comes in contact with a carrier. This paper presents a tool that makes use of contextual data intelligence to predict the risk factor of individuals who come in contact with the carrier.

Keywords: Ebola, iDid app, Contextual intelligence, Susceptibility, Risk factor, Bayes theorem

1 Introduction

When the efforts to prevent a disease fails and an outbreak occurs, the resulting distribution of cases may take various forms that are called epidemic curves. These epidemic curves project the nature of a disease outbreak within a population that are potentially at risk of contracting the disease [12]. Although they indicate the nature of an outbreak, they do not provide sufficient data to understand the chances that a particular individual gets affected by a disease outbreak. To minimize the spread of an epidemic such as the Ebola virus, we need effective contact tracing. The problem is complex as the contact tracing must be done retroactively after a patient is diagnosed with the disease. Any lapse in the tracing could fail to track the citizens at risk. Ad hoc tracing, relying on the infected carrier's recollection of places visited and people met, may lead to inaccurate findings. We need a contextually intelligent application that can keep track of both the individuals' movement and the carrier's movements to identify if the individual is at low risk or at high risk of contracting the Ebola virus. This paper provides a means of tracing and analyzing the risk of contracting the Ebola virus using contextual intelligence and other contributing factors which are also discussed in detail in the following sections.

2 Related Work

There have been several attempts to create right mathematical models that can predict the nature of a disease spread and also monitor individuals health on a daily basis. Achrekar et.al [2] presented a method for gathering twitter data to curb large scale spread of epidemic diseases. This paper presented the Social Network Enabled Flu Trends(SNEFT) architecture as a continuous data collection engine which combines the detection and prediction capability on social networks to discover real world flu trends. Johnson et.al [9] provided a means to calculate a susceptibility ratio using mathematical models. The SIR Model (the number Susceptible, Infectious, or Recovered (immune)) is used in epidemiology to compute the degree of susceptibility for an infected/recovered group of people in a population. This model is an appropriate one to use under the following assumptions.

(1) The population is fixed.
(2) The only way a person can leave the susceptible group is to become infected.The only way a person can leave the infected group is to recover from the disease. Once a person has recovered, the person received immunity
(3) Age, sex, social status, and race do not affect the probability of being infected.
(4) Members of the population mix homogeneously (have the same interactions with one another to the same degree).

IBM has pressed Data Analytics, Mobility and Cloud Computing Technology into service to bring the spread of Ebola in Sierra Leone under control. IBM has deployed SoftLayer cloud technology to set up an Ebola Open Data Repository, to provide governments, aid agencies and researchers with free and open access to the data. [10].

HealthMap, a team of researchers, epidemiologists and software developers at Boston Children's Hospital founded in 2006, is an established global leader in utilizing online informal sources for disease outbreak monitoring and real-time surveillance of emerging public health threats. The freely available Web site 'healthmap.org' and mobile app 'Outbreaks Near Me' deliver real-time intelligence on a broad range of emerging infectious diseases for a diverse audience including libraries, local health departments, governments, and international travelers. It achieves a unified and comprehensive view of the current global state of infectious diseases and their effect on human health [8]. These works can help in preventing the virus from becoming an outbreak, and even provide means to avert the disease spread but at an individual level, it still remains a mystery as to how far they are exposed to the virus.

This paper focuses on monitoring each individual and their exposure to the disease to predict their chances of contracting the disease thereby allowing the user to understand their chances of contracting any virus that they maybe prone to such as the Ebola virus.

3 Factors of disease susceptibility

When studying disease outbreaks and their nature, any infectious disease involves four important factors that cause an individual to be susceptible to the disease. They are:

(1) Time of Exposure
(2) Proximity of Carrier
(3) Carrier Status
(4) Individual Medical History

3.1 Time of exposure

The time of exposure provides us information about the amount of time spent by an individual with the carrier. Certain communicable diseases have a higher rate of susceptibility even if the exposure is for a short duration. Based on research, we can say the exposure rate for contracting the Ebola virus (or any other infectious disease) is dependent on the duration of contacts made by the individual with a carrier.

3.2 Proximity of carrier

The proximity or the nature of contact with the infected individual also plays a vital role in the probability of contracting any disease. The distance of the carrier from an individual can indicate the risk of contracting the disease.

When an individual comes in contact with the carrier on just one occasion, their probability of contracting Ebola will be less when compared to the individual who comes in contact on multiple occasions. In cases like the Ebola virus, actual physical contact is needed to contract a disease while in other cases (such as flu) no physical contact is needed.

3.3 Carrier Status

The Ebola virus is a disease which has an incubation period of 21 days during which time the infected carrier can spread the disease. Thus an individual's probability of contracting from the carrier varies depending on the day within the infectious period of the carrier. The carrier status refers to the day on which an individual comes in contact with the carrier.

3.4 Individual Medical History

An individual's medical history is important to see how the immune levels of the person could either strengthen or weaken the prospect of contracting a disease. Medical records can be used to gather information to predict how prone an individual is to contracting the Ebola virus.

4 Other factors

Besides these primary factors, there are a few incidental factors that contribute to an individual's susceptibility.

4.1 Environment of Individual and Facilities

The environment or the locality play a vital role in analyzing the risk: if the individual is located in a place with very good health facilities and has generally good hygiene, then their risk of contracting or spreading the disease is less when compared to an individual living in an environment with few medical facilities and having generally poor hygiene.

4.2 Population Demography

The implications of demographic changes for the spread and control of infectious diseases are not fully understood. But an individual's susceptibility can be studied based on the population structure and the marked effect it can have on any disease transmission. A population with more carriers can indicate higher risk for any individual located in that area. Suppose an individual is located in Sierra Leone, then he/she has a higher risk of contracting Ebola than an individual living in New York.

Another factor relates to cultural habits including touching and treating infected persons. This can vary from one environment to another.

4.3 Heat Zones

The Global Surveillance Network developed by the CDC in 1995 was based on the concept of a data collection network for the surveillance of travel related morbidity. The goal was to direct clinics to be ideally situated so as to effectively detect geographic and tem-

poral trends in morbidity among travelers, immigrants and refugees [7]. Such a concept would be useful for tracking carriers of the Ebola virus who can be monitored and thus provide a heat zone to indicate how a particular area is geo-fenced based on the nature of the affected population. Using data collected by

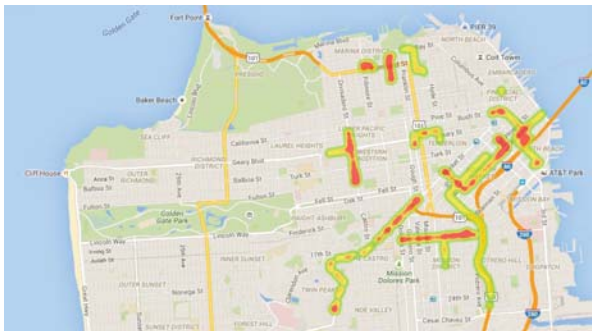


Figure 1: Heat zones

the CDC based on the survellance network established, heat zones can be constructed. Figure 1 shows a heat zoned area in California, where the traces of a carrier's movement is shown on the map. A zone's opacity indicates the area's degree of risk relative to other area. This could be useful for an individual travelling to an address that is close to the hot zones.

5 Implementation of our Application

5.1 Contextual data intelligence and iDid Inc app

As stated previously, to accurately track diseases, it is necessary to track the daily activities of both the carrier and the individual in question over the previous days and weeks. For example, since an Ebola infected carrier is contagious for 21 days, it would be necessary to discover all the locations the carrier visited and the people the carrier met. This information can then be used by individuals to correlate their own movements over the past 21 days to assess their risks. In many cases it is possible to predict past activities and behaviors of individuals based on current or future activities. However, the process of collecting data on an individual's activity and maintaining a repository is a tedious task: it must be collected in real-time to have an active monitoring scheme. It is also necessary to collect geographical information (GPS data) as well as the nature of the activity.

For example, GPS can be used to track the locations a person visits and perhaps provide contextual information to determine the duration of each visit.

Case 1: If a person visits a restaurant, it is reasonable to assume he/she will spend considerable time there to order food and consume it on the premises, which increases his/her chances of coming in contact with a carrier.

Case 2: If a person drives between locations, there is very little possibility of direct or very close contact with a carrier (unless the carrier is inside the vehicle).

Similar contextual information can be used to assess the risk of contracting or spreading infectious diseases. We use an application developed by iDid Inc. [1] to track activities carried out by individuals on a daily basis and generate reports of those activities at the end of the day.

5.2 Integration of iDid Inc and risk factor generation app

Data collected on individuals using the iDid app can be used to track their movements on a routine basis and log this information in a database. The database contains information such as duration spent at a place, time of visit, time of travel to a new location, number of contacts made with the individuals at the new location, etc. Data is stored in a backend repository and a users future schedule is predicted by analyzing his/her currently available data.

The data thus collected can be correlated with the data collected for the carrier (assuming such data exists) to generate discrete susceptibility ratio graphs (indicating probability of contracting the disease).

5.3 Creation of a web application

The purpose of the application is to provide an individual with an estimate of their risk or how susceptible they are to contracting the disease. The application collects information from the back-end which contains the individuals daily activities as described above. In addition, the individual's medical records can be used to improve the accuracy of the predictions.

A composite risk factor is generated based on the data collected and on the various factors described previously.

5.4 Data Privacy and Security

The information required to calculate the probability of contracting the disease is personal and often covered by HIPAA. Therefore, data privacy is considered to be of paramount importance. To protect a user's privacy in our system, the data for a given user will be maintained in their own smartphone or their personal storage spaces, often protected with passwords and data encryption. The iDid app uses calendar selected by the user when he/she installs the app and a private database is maintained to log the completed activities such as drives, flights, places visited, sleep, etc. Our application does not save user contextual information or medical history, but uses the data only to compute risk factors. The iDid app also uses the Google Maps

API(s) [11].

With the users permission, the risk probabilities are used to track disease spread in a population, without any information that can be used to identify the individual.

6 Bayesian Analysis of Data

We use Bayesian probabilities to assess the risk of contracting a disease utilizing contextual information and medical history [13].

6.1 Quick Bayesian Risk Calculator

$$P(B_1|A) = \frac{P(B_1 \cap A)}{P(A)} = \frac{P(B_1)P(A|B_1)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)}$$

Figure 2: Bayesian Analysis Formula

The Bayesian network for determining risk describes the probability of an event, based on the conditions that might relate to that event. If an individual travels to the same places as the carrier and engages in activities that require them to spend time in proximity to the carrier, then their risk factor depends on the total time spent wit the carrier doing the activities.

Instance: Suppose a carrier is sitting at Starbucks at 10 am and another patron enters the store to have a cup of coffee. One could assume the patron might spend 15-20 minutes in the store and so their risk factor is likely higher than it is for an individual who drives past the same Starbucks and does not come in close quarters with the carrier. This Bayesian network involves knowing the type of activity, place, time spent by an individual, by which we can calculate the risk factor instantly at that given time.

6.2 Bayesian Analysis of Medical record

An individual's medical history contributes to a fair share of the person's risk to a disease. The information they provide could lead to an accurate analysis of understanding the individual's probability of conracting the Ebola virus.

6.3 Contextual Intelligence Probability Calculator (CIPC)

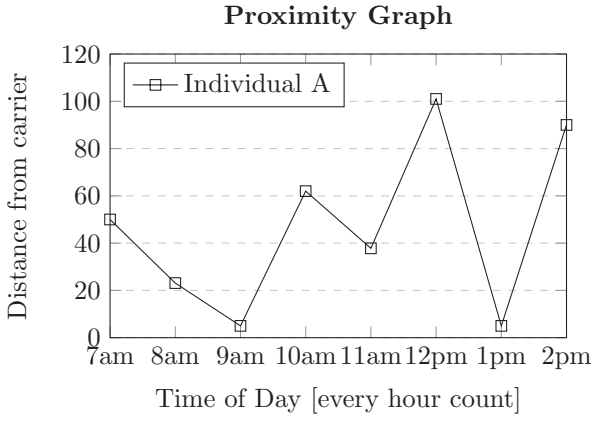
The goal is to provide the individual with objective data to assess their risk. The CIPC will yield a precise risk value which maybe high or low depending on the individual's susceptibility. It is calculated by the integration of probability values calculated from the Quick Bayesian calculator and the medical history Bayesian value, along with the probability values obtained from the factors that contribute to the risk, such as time of exposure, number of contacts made and the environment.

7 Graphical Representation of Individual Monitoring

7.1 Proximity Graph-

An individual's proximity is monitored to understand their closeness to the carrier. By monitoring the carrier and individual contextually, the data gets stored on an hourly basis.

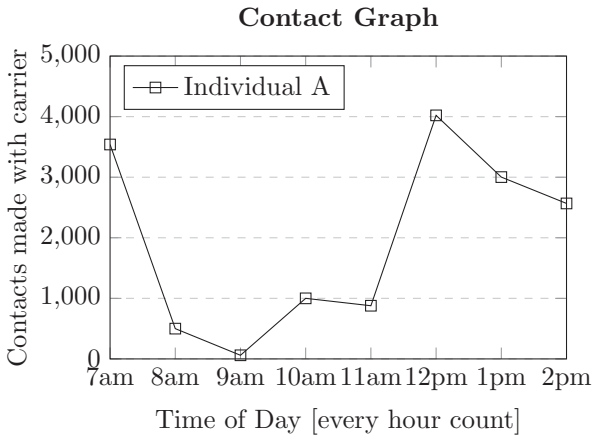
The proximity graph could indicate an individual A, whose distance from the carrier is being tracked from 7am to 12pm. The tracking yields a line graph that shows the individual's proximity to the carrier. Using the data collected, we can provide a probabilistic estimate of how prone the individual is to infection in terms of their distance from the carrier. This value is integrated along with the CIPC to calculate the overall risk factor.



7.2 Contact Graph-

The Ebola virus in particular is a communicable disease, as we could infer from the earlier discussion that the virus spreads swiftly through physical contacts, transfusion of blood, etc. Thus, examining the number of contacts an individual makes with the carrier would be highly beneficial to our objectives.

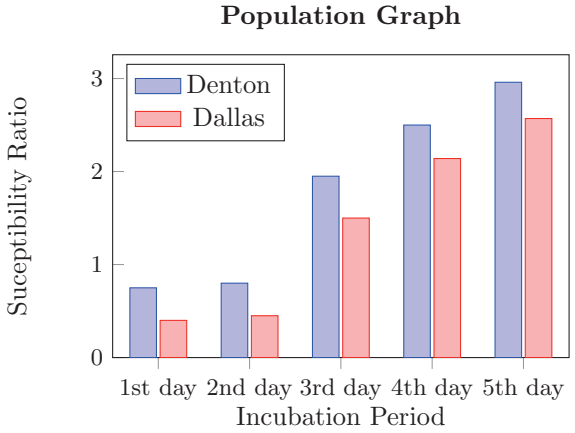
The Contact graph shows an individual A who has made contact with a carrier at various times during a day. The day of exposure, along with the proximity of the individual to the carrier, aids in recognizing the individual's risk factor. Realistically, the time spent with the carrier could be monitored and its significance can be judged on the basis of how close the individual was to the carrier during that time. We can then track the number of contacts and provide a likelihood value to help us derive an overall risk factor.



7.3 Population graph-

The demographic structure of a population is a key determinant of patterns of contact and hence of infectious disease spread, with implications for the design of effective control measures [6]. The population distribution in a given place can alter the risk for an individual. The increased risk maybe associated with a household, the impact of health facilities in the locality, the age of the population, etc.

The sample graph shows the population of Dallas and Denton, where the contrasting demography between the two cities indicates how an individual is at a higher risk if he/she comes in contact with a carrier in Denton as opposed to Dallas. This variation is an implication of how the faclities at Dallas are better than Denton. The awareness and treatment in Dallas is expeditious compared to Denton.



8 Working of the model

(1) **Individual/Carrier Movement:** The individual/carrier movement is tracked in real time by actively monitoring them using the iDid app. The data is stored and the tracking data on previous movements of individuals/carriers is used to calculate the risk factor of contracting the Ebola Virus.

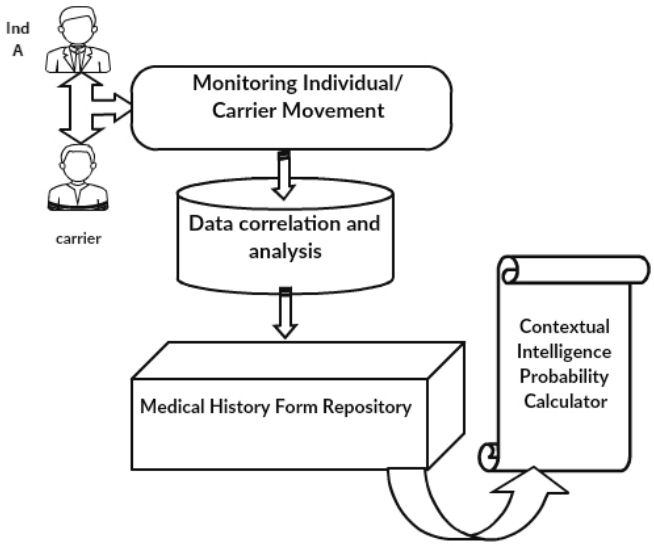


Figure 3: Framework of Application

(2) **Data Correlation:** The data is collected from the individuals and stored in the repository. They are correlated based on the factors that help to analyze the risk of contracting the Ebola virus.

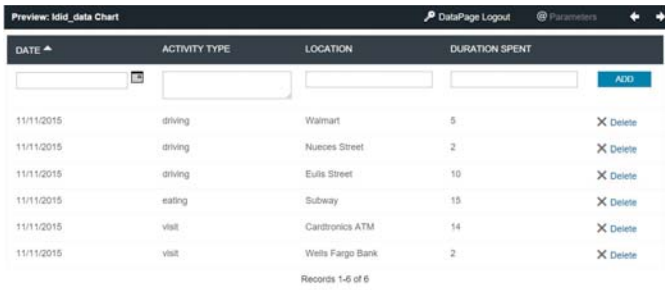
(3) **Medical History Form Repository:** An individual's medical history will support a more accurate analysis of their risk of contracting Ebola from a carrier. Using Bayesian Data analysis, we can predict a more accurate value of their risk.

(4) **CIPC:** The Contextual Intelligence Probability Calculator computes a composite risk by convolving the value generated from the medical history of the individual and the conditional probability calculated from the individual's movement with respect to the carrier at a given moment in time. The type of activity, place of visit and time spent at a place while the carrier is also in motion are correlated along with the main contributing factors such as day of exposure and number of contacts made. The conditional probability obtained is used to provide an accurate value indicating whether the individual is at high or low risk.

9 Experimental Results

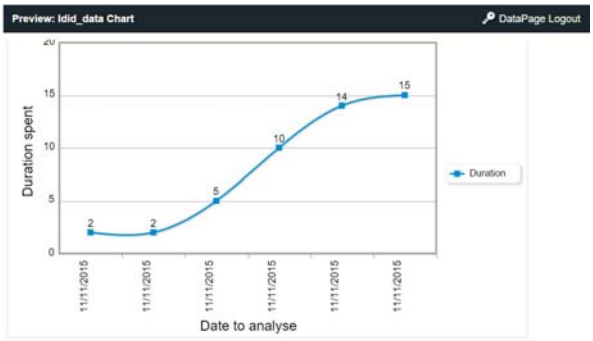
To evaluate the application, a series of tests were conducted and the results were analyzed. Due to the unavailability of real epidemic data, the tests were based on statistical and experimental data sets. The database containing the sample test data is created by tracking the movement of individuals and a mock-up carrier using the iDid application. The following is a series of data collected using the iDid app and stored in the repository.

The sample data shows six rows corresponding to



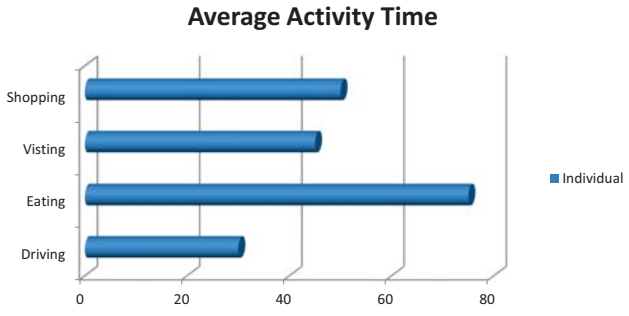
the movement of an individual with respect to the carrier. The user can add/delete rows depending on the correctness of the location. The location can be modified and the values will be updated accordingly.

The application will render a graph showing their proximity to the carrier and the risk probability value.



9.1 Average Activity Time

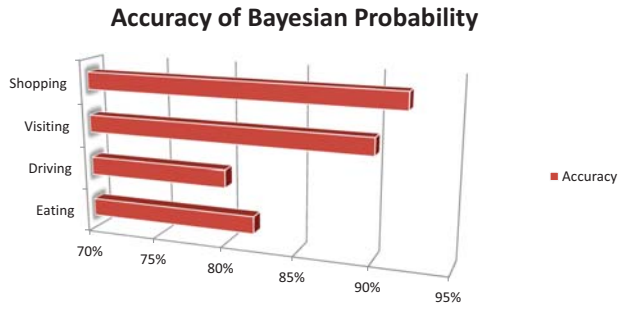
Based on analysis of time spent on various activities, we compute an average time for each activity.



This chart describes the average time an individual spends on a particluar activity. This may be helpful for estimating the duration of untracked activities for which empirical data is not otherwise available. This could reduce the number of false positives when predicting a risk value.

9.2 Accuracy of CIPC(Contextual Intelligence Probability Calculator)

The application monitored a series of individuals located at different proximities to the carrier, then probability values were assigned based on the contributing factors such as time of exposure, number of contacts made and medical history. The Bayesian Conditional probability was measured based on movements with respect to their activities.



This chart describes the accuracy of Contextual intelligence Bayesian analysis based on the different factors monitored for an individual with respect to the carrier's movement.

10 Future improvements

Although the thrust of this paper has been on epidemiological research relating to the factors that contribute to contracting the Ebola virus, there are several other communicable diseases that could affect individuals. One of the main reasons that the Ebola virus spread across Africa was due to the lack of awareness among the population about the severity of the epidemic. There are similar deadly diseases for which people fail to understand the risk factors and unwittingly become carriers. An epidemic of cholera infections was documented in Haiti for the first time in more than 100 years in October 2010. Cases have continued to occur, raising the question of whether the microorganism has established environmental reservoirs in Haiti [3].The patterns of cholera transmission and the seasonality of cholera in an environment is largely based on water contamination, poor sanitation facilities and inadequate hygiene.

Our application described in this paper can be extended to predict contaminated water locations and the probability of individual's susceptibility to Cholera based on location and climatic conditions. In May 2015, the Pan American Health Organization (PAHO) issued an alert regarding the first confirmed Zika virus infection in Brazil and on Feb 1, 2016, the World Health Organization (WHO) declared the Zika virus a public health emergency of international concern (PHEIC). Local transmission has been reported in many other

countries and territories [5]. The Zika virus will likely continue to spread to new areas. Our approach to Ebola risk analysis can be used to analyze risk posed by other diseases such as Zika, by changing the factors that play a role in the spread of the disease, probability curves and other external factors described in this paper.

11 Conclusion

In this paper we described a framework that can be used to assess the risk posed to individual by relating contextual information that tracks the activities of the individual and correlates this data with that of a carrier. We relied on iDid app and demonstrated how our system works. Since actual contextual information on any specific carriers is unavailable, we used made up data and provided users with privacy of data. We used Bayesian models to combine the risks emanating from several factors into a single risk value. We plan to extend our study to model other infectuous diseases

In [4], the report states that in 2014, a team of researchers from Virginia Tech Institute tried to create a model and characterize the nature of the disease outbreak in West Africa. But the research yielded results that did not prove to be accurate. Nevertheless, it can be a stepping stone to understand and analyze an individual's behavior and their movement around the carrier to statistically predict the nature of a disease outbreak. If each individual is able to understand their risk factor and susceptibility to the disease, it could mitigate the possibility of a disease outbreak.

This application reported the ways in which mobile and wireless technologies can be used to implement the vision of pervasive healthcare.

12 Acknowledgements

This research is supported in part by the NSF grant 1513369. The authors also acknowledge the help of Davis Struble for his editorial comments.

References

[1] *Idid Inc.* http://idid-inc.com/businesses/, 2013.

[2] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. *Predicting Flu Trends using Twitter Data*. 2011.

[3] Meer T Alam, Thomas A. Weppelmann, and Chad D. Weber. *Monitoring Water Sources for Environmental Reservoirs of Toxigenic Vibrio cholerae O1, Haiti*. Center for Disease Control and Prevention, 2014.

[4] David Brown. *How Computer Modelers Took On the Ebola Outbreak*. spectrum.ieee.org, 2015.

[5] CDC. *Zika virus and spread*. Center for Disease Control and Prevention, 2016.

[6] Nicolas Geard, Kathryn Glass, and James M McCaw. *Probabilistic graphic models applied to identification of diseases*. Epidemics Volume 13, December 2015, Pages 5664, 2015.

[7] David L Heymann and Guenael R Rodier. *Global Surveillance of Communicable Diseases*. 1998.

[8] Boston Children's Hospital. *Health Map Organisation*. http://www.healthmap.org/site/about, 2006.

[9] Teri Johnson. *Mathematical Modeling of Diseases: Susceptible-Infected-Recovered (SIR) Model*. University of Minnesota, Morris, 2009.

[10] IBM Africa Research Labs. *IBM applies data analytics, mobile technology and cloud computing to help fight the Ebola outbreak in West Africa*. 2014.

[11] Google Maps. *Google Timeline*. Google, 2014.

[12] AFMC Public Health Educators' Network. *Patterns of disease development in a population: the epidemic curve*. 2007.

[13] Renato Cesar Sato and Graziela Tiemy Kajita Sato. *Probabilistic graphic models applied to identification of diseases*. Einstein (So Paulo) vol.13 no.2, 2015.

An Accountable Access Control for E-health Clouds

Maode MA[1], **Qianqian ZHAO**[2], **Yuqing ZHANY**[2]
[1] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2] State Key Laboratory of Integrated Service Networks, Xidian University, China

Abstract - *Cloud Computing provides an efficient solution to share health information in the electronic health record (EHR) systems, which encourage the patients to upload their personal health information records (PHR) into the cloud servers. However, some security issues such as data confidentiality, identity privacy, scalability in key management, and user key accountability should be concerned in the application of the PHR. In this paper, we address the security issues in the deployment of the E-health clouds and propose a secure access control scheme to achieve the user key accountability. The performance and security analyses indicate that the proposed scheme is highly efficient and provably secure.*

Keywords: Personal health information record; Cloud computing; Attribute-based encryption; key accountability.

1 Introduction

The personal health record (PHR) system has emerged as a patient-centric platform for the health information exchange in recent years [1]. The PHR service allows the patients to create, manage, and control their personal health records in the external storage through the Internet, which has made the health information more convenient to be retrieval. Cloud computing has emerged as the most advanced computing paradigm in the information technology. It is a model for enabling ubiquitous, convenient, on-demand network access to configurable computing resources that can be rapidly provided and released with minimal management efforts or interaction from the service providers [2]. The PHR service based on cloud computing has drawn extensive attentions from both academia and industry in recent years. There are many security and privacy risks in the application of cloud computing based PHR service. The primary concern for users is the confidentiality of the shared data because the cloud servers are not entitled to access the shared data content for data confidentiality [3] and the assumption that the data owners and the servers to store their data are in the same trusted domain will no longer hold in the cloud computing based PHR systems. The access control caused by the encryption technology becomes a critical issue in the cloud computing based PHR systems. On one hand, the clients can store their data remotely and enjoy the on-demand cloud service without a burden of local hardware and software management, while it simply implies that they will lose the physical control of their local data [4]. Therefore, a trusted authorization center is required to implement the management. On the other hand, different users need to hold different access privileges with regard to the specific data. Furthermore, the data owners may not know who are going to access the shared data. Obviously, the traditional public key encryption mechanisms cannot satisfy the abovementioned requirements on the access control, which always need to perform much key management and key distribution work and require the data owners to stay online all the time. Moreover, the data owners may have to generate multiple copies of the encrypted data for the users with different keys. It may incur heavy storage overhead to make the data owners to pay more for the storage space [5]. It is clear that the design of secure and efficient access control scheme with appropriate encryption mechanisms has become an imminent demand.

The attribute-based encryption scheme (ABE) in [6] enabling one-to-many public-key encryption for the first time brings us a new approach for the fine-grained access control. The first key-policy ABE (KP-ABE) has been proposed in [7], by which the access policies are associated with users' private keys and the attributes are attached to the ciphertexts to describe the data. Another type of ABE is called Ciphertext-Policy ABE (CP-ABE) [8], by which the access policies are associated with the ciphertexts and the users are provided with the private keys corresponding to the attributes they own. Decryption is possible if and only if the attributes match the access policy. Hence, as long as the attribute authority issues the private keys according to that a user is entitled, the confidentiality of the data can be guaranteed. The attribute encryption mechanisms can greatly promote the development of secure and efficient access control schemes.

However, the escrow problem inherited from the identity based encryption (IBE) is another drawback of the attribute based encryption. It particularly implies that the existence of one single trusted authority (TA) is not feasible in the PHR systems. On one hand, it is not practical to delegate all the attribute management tasks including authenticating all users' attributes and generating the corresponding secret keys to the single trusted authority, which may create a bottleneck of traffic load. Different types of participants should have different authorization centers for authentication to make these centers to be unaware of the existence of each other. On the other hand, it may suffer from the key escrow problem severely because the single authority who assigns the private keys for the users can freely decrypt the ciphertexts. There is a necessity to design the multi-authority ABE to deal with these issues. The multi-authority attribute based encryption technique needs multiple trusted authorities in the system, which are jointly responsible for the authentication of the user attributes and issuing the private attribute keys.

Many multi-authority attribute based encryption schemes have been designed in the literature. An efficient scheme has been firstly presented in [9], which has only addressed the issue without a solid solution due to the existence of the central authority (CA). Some other multi-authority proposals have been put forward in [5], [9], [10] removing the central authority. The most revolutionary multi-authority CP-ABE scheme without a central authority has been constructed in [11-12], where the CA with different functions in the system is only responsible for the distribution of global secret key and global public key to each legal user in the system. However, the CA is not involved in any attribute management and the creation of secret keys that are associated with attributes. More importantly, it outsources major computation work of the attribute decryption by using a token-based decryption method. It has a strong allure for the users with limited computation power.

According to the characteristics of the ABE, the private key of one user is only related to the users' attributes without the involvement of his concrete identity. It will incur the issue of user key accountability, which means that one malicious user may share his private key illegally with other unauthorized users to gain some special benefits. Since the private key is only associated to the user's attributes, it is apparent that there could be many users owning the same attribute set. If a user's private key has been disclosed, it is difficult to affirm the user identity of this private key. A conclusion can be drawn that the user key accountability has the same security requirements with the traceability. The design of a secure access control scheme to achieve the user key accountability is an urgent challenge. Many solutions to handle the user key accountability have been proposed in [13-17]. The first two solutions for the single-authority have appeared in [13] and [14]. Both of them have not only achieved the user key accountability but also implemented partial hiding of the access policy. A tracing scheme has been proposed to achieve white-box traceability in [15] and black-box traceability in [16], respectively. They can support any monotone access policy and are efficient as one of the best existing non-traceable CP-ABE systems. But they are not applied to the multi-authority environment. A novel scheme to solve the user key accountability for the multi-authority has been first designed in [17], which can achieve the public tracking, by which any users and the attribute authority can trace the identity of a malicious user. But, the access policy has been implemented in the same way as the single-authority solutions without the expressiveness.

In order to efficiently solve the abovementioned security issues including data confidentiality, key management and distribution, the escrow problem and the user key accountability to achieve the fine-grained access control in the PHR system, we design a novel and efficient access control scheme based on the multi-authority ABE mechanism with the ability to handle the user key accountability, which is the traceable multi-authority access control scheme (MAAC). The contributions of this paper can be summarized as follows.

1) The proposed fine-grained access control scheme for the PHR system supported by the multi-authority ABE mechanism holds the distinctive feature that there are multiple attribute authorities assigning the corresponding attribute keys to the users.
2) The proposed scheme addresses the user key accountability effectively with the ability to achieve the public and black-box traceability.

The remaining of this paper is organized as follows. We first simply describe the system model a of the PHR systems with the preliminary in Section 2. In Section 3, we present the details of the proposed MAAC scheme for the PHR system. Furthermore, we demonstrate the security and performance analysis on the proposed scheme in Section 4. Finally, the conclusion is given in Section 5.

2 System Model and Preliminary

2.1 System Model

The PHR system under the study consists of five types of entities including the PHR service provider, the attribute authorities, the cloud server, the data owners, and the data consumers. The PHR service provider is a global trusted certificate authority. It sets up the system and accepts the registration of all the users and the attribute authorities (AA) in the system. The AA is a conventional attribute authority that is responsible for issuing the user's attributes according to their roles or identities in the domain. Each attribute is supervised by a single AA, while each AA will manage an arbitrary number of attributes. The cloud server stores the PHRs of all the participants and provides the data access for the users. The cloud server outsources the computation of the attribute decryption. The data owners define the access policy, by which the data will be encrypted before outsourcing. The way for the user himself to perform the policy of access control is extremely appealing. Only when the attribute set of the data consumers meets the access policy, they can access the shared data conveniently. The data consumers are the subjects who request the cloud server for the stored data. They will send the corresponding requirements to the cloud server when necessary.

It is assumed the cloud server is semi-trusted in the system, which is honest but curious. The server will honestly follow the operations in general, while it could also try to explore as much secret information as possible. In addition, an attacker can access the shared data regularly but may also share his private key illegally to other illegal clients for extra profits. The malicious user could easily generate a black-box pirate device by his reasonable private keys, which includes an attribute set, the hidden decryption keys and the decryption algorithm. Furthermore, it is assumed that the malicious user can collude with any other users to generate the pirate device, while the cloud server will perform his operation honestly.

2.2 Preliminary

Linear Secret Sharing Scheme

Since the linear secret-sharing schemes (LSSS) have been employed in the design of the proposed scheme, we review the definition first. A secret sharing scheme Π over a set of parties P is called linear (over $\mathbb{Z}_p$) if

1. The shares for each party form a vector over $\mathbb{Z}_p$.
2. There exists a matrix M called the shared generating matrix over Π. The matrix M has l rows and n columns. For all $i = 1, \cdots, l$, the i^{th} row of M is labeled by a party $\rho(i)$, where ρ is a function from $\{1, \cdots l\}$ to $\mathcal{P}$. When we consider one column vector $v = (s, r_2, r_3, \cdots \cdots, r_n)$, where $s \in \mathbb{Z}_p$ is the secret to be shared and $r_2, r_3, \cdots \cdots, r_n \in \mathbb{Z}_p$ are randomly chosen, then Mv is the row vector of l shares of the secret s according to Π. The shares of $(Mv)_i$ belongs to the party $\rho(i)$

It is shown that each linear secret sharing scheme also holds the linear reconstruction property, defined as follows. Suppose that Π is a LSSS for the access structure A. Let $S \in A$ be any authorized set, and let $I \subset \{1, \cdots, l\}$ be defined as $I = \{i: \rho(i) \in S\}$. Then, there exist constants $\{\omega_i \in \mathbb{Z}_p\}_{i \in I}$ such that $\sum_{i \in I} \omega_i \lambda_i = s$. If $\{\lambda_i\}$ are valid shares of any secrets according to Π. Furthermore, it is clear that these constants $\{\omega_i\}$ can be found in time polynomial in the size of the share generating matrix M. We note that for unauthorized sets, no such constants $\{\omega_i\}$ exist.

Classical Assumptions and Variants

The decisional bilinear Diffie-Hellman problem (DBDH[1]) in G :

Given (P, aP, bP, cP, U) for some $a, b, c \in Z_p^*$, and $U \in G$, where G is one cyclic group with the prime order p, output **Yes** if $U = abcP$ and **No** otherwise.

The modified decisional bilinear Diffie-Hellman problem (DBDH[1]-M) in G:

Given (P, aP, bP, Z) for some $a, b \in Z_p^*$, and $U \in G$, where all the parameters involved go the same by DBDH, output **Yes** if $U = ab^2P$ and **No** otherwise.

The decisional bilinear Diffie-Hellman problem (DBDH[2]):

Given (P, aP, bP, cP, Z) for some $a, b, c \in Z_p^*$ and $Z \in G_T$, where G and G_T are two cyclic groups with the prime order p, and $e: G \times G \to G_T$ is a bilinear map, P is the generator of G, and $g = e(P, P)$, output **Yes** if $Z = g^{abc}$ and **No** otherwise.

The mixed decisional Diffie-Hellman problem (MDDH):

Given (P, aP, a^2P, g^b, Z) for some $a, b \in Z_p^*$ and $Z \in G_T$, where all the parameters involved go the same with DBDH[2], output **Yes** if $Z = g^{ba^2}$ and **No** otherwise

3 The Proposed MAAC Scheme

In this section, we first provide an overview of the proposed MAAC scheme for the cloud computing based PHR service systems. Then, we describe the detailed design of the MAAC scheme.

3.1 Overview

The major design goal of the proposed MAAC scheme includes three major aspects including the data confidentiality, the access control, and the user key accountability. On the data confidentiality, it requires that the unauthorized users are incapable of getting the knowledge of the shared data. There are two types of requirements on the access control. The first one is that the unauthorized user cannot access the shared data absolutely. Another one is that the users themselves are the executors of the access control policy, which have a complete control on the access policy. On the key accountability, the system needs to ensure that if any user shares his attribute key illegally, the identity of this user would be disclosed.

By the proposed scheme, the CA first assigns a global user identity *uid* to each user and a global authority identity *aid* to each attribute authority in the system. Since the *uid* is globally unique, the secret keys issued by different AAs for the same *uid* can be tied together for the decryption purpose according to the unique *uid*. Furthermore, since each AA is associated with an *aid*, each attribute is distinguishable although some AAs may issue the same attribute. Thus, the collusion attacks can be resisted by using these special *aid* and *uid*. To achieve an efficient decryption, the token-based decryption outsourcing approach in [12] has been employed. In the attribute decryption phase, the user can submit his secret keys issued by the corresponding attribute authority and his global public key to the cloud server for re-encryption, and then the cloud server transmits the new ciphertexts to the user. Only when the user's secret keys satisfy the access policy in the ciphertexts, the cloud server can successfully generate the new ciphertexts. The user can then decrypt the new ciphertexts by using his global public key and global secret key.

3.2 Details

The proposed scheme has five phases including System Setup, Key Generation, PHR Encryption, PHR Access, and Traceability.

3.2.1 System Setup

Let S_A and S_U denote the set of attribute authorities and the set of users in the system, respectively. Given a security parameter $k \in Z$, a k-bit prime q is generated. Let G_1 be an

addition group and G_T be a multiplicative group with the same order q and define one bilinear map $e: G_1 \times G_1 \rightarrow G_T$. Let P be the generator of G_1. Let $H_1: \{0,1\}^* \rightarrow G_1$ and $H_2: G_T \rightarrow \mathcal{M}$ be two hash functions such that the security is in the random oracle, where $\mathcal{M}$ denotes all the choices of the plaintexts.

The CA first chooses two random numbers $a, z \in Z_q^*$ as the master key of the system, and then generates the system parameters: $Q = aP$; $g = e(P,P)$; $Z = g^z$. Meanwhile the CA establishes additional a pair of secret key and public key $(sk_{CA}; pk_{CA})$ to create the certificate for the registered users. Finally, the public parameter can be represented fully by $\{G_1, G_T, e, H_1, H_2, P, Q, g, Z\}$.

In the following, the CA begins to accept both user registration and AA registration. When a new user joins the system, the CA first authenticates the user to check whether the user has been registered before. If the user is legal in the system, the CA then assigns the user a global unique identity uid. Then, the CA chooses $x_{uid}, y_{uid} \in Z_q^*$. These two numbers are required to satisfy the equation $x_{uid} + a y_{uid} = z \pmod q$. It is clear that the number of such a combination of x_{uid}, y_{uid} is no greater than q according the theorem about the root of the congruence equation. The CA computes $GPK_{uid} = \pi_{uid} = y_{uid} P$ as the global public key of the user, and takes the corresponding $GSK_{uid} = x_{uid}$ as the global private key of the user. The CA also generates a certificate for this user by using its secret key sk_{CA} denoted as $Sig_{sk_{CA}}\left(uid, y_{uid}, \frac{1}{x_{uid}} P\right)$. The CA sends the global public key GPK_{uid}, the global secret key GSK_{uid}, the user's certificate $Sig_{sk_{CA}}\left(uid, y_{uid}, \frac{1}{x_{uid}} P\right)$ to the user together with uid. In the end, the CA adds the new $(uid, x_{uid} P, y_{uid} P = \pi_{uid})$ into the traceability list *T*. Finally, the CA makes a public traceability list as following: $\{uid_1, x_{uid_1} P, y_{uid_1} P\}$; $\{uid_2, x_{uid_2} P, y_{uid_2} P\}; \cdots \{uid_N, x_{uid_N} P, y_{uid_N} P\}$, where *N* is the number of all users and it is dynamically changing over time.

Each AA should also register itself to the CA in the system initialization. If the AA is a legal authority, the CA first assigns a global authority identifier *aid* to the AA. Then, the CA sends its public key pk_{CA} together with the system parameter. It is only the attribute authority, who can receive the public key of CA pk_{CA} rather than any other users.

The authority setup is implemented independently by each AA_k ($k \in S_A$). Let S_{A_k} denote the set of the attributes managed by this authority AA_k and $n_k = |S_{A_k}|$ denote the number of the attributes controlled by this authority. It chooses two random numbers $a_k, b_k \in Z_q^*$ as its secret key denoted as $SK_k = (a_k, b_k)$. For each attribute $x_k \in S_{A_k}$, the AA_k chooses a random number $A_k^{x_k} \in Z_q$, and computes $T_k^{x_k} = A_k^{x_k} P$, which can be regarded as the public key of this attribute generated by this attribute authority. Then, the AA_k generates its public key $PK_k = \left\{e(P,P)^{a_k}, \frac{1}{b_k} P, T_k^1, T_k^2, \cdots, T_k^{n_k}\right\}$ through its secret key $SK_k = (a_k, b_k)$, respectively. Finally, each owner can construct the integrate public keys as $PK = (g, Q, Z, \{PK_k\}_{k \in S_A})$.

3.2.2 Attribute Key Generation

For each user U_j ($j \in S_U$), each AA_k ($k \in S_A$) first checks whether the user is a legal user by verifying the certificate of this user. If the user is a legal user, which means that the AA_k can successfully decrypt the $Sig_{sk_{CA}}\left(uid, y_{uid}, \frac{1}{x_{uid}} P\right)$ by applying the pk_{CA}, the AA_k assigns the set of attributes $S_{j,k}$ to the user U_j according to his role in the domain of the attribute authority. For the concrete attribute set $S_{j,k}$, the AA_k generates the user's secret key $SK_{j,k}$ as

$$SK_{j,k} = \left(K_{j,k} = \left[\frac{a_k}{x_j} P + y_j Q\right], L_{j,k} = \left[\frac{b_k}{x_j} P\right], \forall x_k \in S_{j,k}, K_{j,x_k} = \left[\frac{b_k}{x_j} P + y_j b_k H_1(x_k) + y_j b_k T_k^{x_k}\right]\right)$$

3.2.3 PHR Encryption

First of all, the shared data $\mathcal{M}$ will be encrypted by the symmetric encryption with the private key $\mathbb{k}$, which is the core of the proposed access control solution. The public key PK, the private key $\mathbb{k}$ and one access structure (M, ρ) are taken as the inputs, where M is a $l \times n$ matrix and ρ is an injective function which associates each row of M to one positive attribute. It marks the involved set of the attribute authority as I_A over all the selected attributes according to the access structure. Then, the data owner chooses a random encryption exponent $s \in Z_q$ and creates a random column vector $v' = (s, y_2, \cdots, y_n) \in Z_q^n$, where $y_2, \cdots, y_n$ are used to share the encryption exponent s. It computes $\lambda_i = M_i v'$ for each $i \in (1, l)$, where M_i is the vector corresponding to the i-th row of M. Then, it randomly chooses $r_1, r_2, \cdots r_l \in Z_q$, and computes the ciphertext as follows:

$$CT = \left((M, \rho), c_1 = sP, c_2 = s^2 Q, c = \left(\mathbb{k} \oplus H_2\left(Z^{s^2}\right)\right), \left(\prod_{k \in I_A} e(P,P)^{a_k}\right)^s, \forall i = 1 \text{ to } l: C_i = \left[\lambda_i Q + (-r_i) T^{\rho(i)} + (-r_i) H_1(\rho(i))\right]; D_{1,i} = \frac{r_i}{b_k} P; D_{2,i} = -\frac{r_i}{b_k} P, \rho(i) \in S_{A_k}\right)$$

3.2.4 PHR Access

The PHR Access phase consists of two steps including cloud server re-encryption and user data decryption. First, a user has to get the shared data from the cloud server. In the process, the attribute decryption is outsourced to the cloud server. Finally, the user can gain the real plaintext through simple operations by using its global public key and global secret key.

Cloud server re-encryption: The user $U_j\ (j \in S_U)$ first sends its secret key $\{SK_{j,k}\}_{k\in S_A}$ to the server for requesting the corresponding ciphertexts. Only when the attributes the user U_j possesses satisfy the access structure in the ciphertext CT, the cloud server can dispatch the corresponding ciphertexts to the user. The cloud server will compute a decryption token for the ciphertext before re-encryption. Let I be $\{I_{A_k}\}_{k\in I_A}$, where $\{I_{A_k}\}_{k\in I_A} \subset (1,\cdots,l)$ is defined as $I_{A_k} = \{i: \rho(i) \in S_{A_k}\}$. Let $N_A = |I_A|$ be the number of AAs involved in the access in the ciphertext. If the attributes the user U_j possesses satisfy the access structure in the ciphertext, the cloud server can choose a set of constant $\{\omega_i \in Z_p\}_{i\in I'}$, and reconstruct the encryption exponent as $s = \sum_{i\in I'} \lambda_i\omega_i$, where $\{\lambda_i\}$ are valid shares of the secret s in the encryption process. First, the decryption token (TK) can be acquired as follows:

$$TK = \prod_{k\in I_A} \frac{e(c_1, K_{j,k})}{\prod_{i\in I_{A_k}} \left(e\left(C_i, GPK_{U_j}\right)\cdot e\left(D_{1,i}, K_{j,\rho(i)}\right)\cdot e\left(D_{2,i}, L_{j,k}\right)\right)^{w_i N_A}}$$

$$= \frac{e(P,P)^{asN_A y_j}\cdot \prod_{k\in I_A} e(P,P)^{s\frac{a_k}{x_j}}}{e(P,P)^{aN_A y_j \sum_{i\in I'} w_i\lambda_i}} = \prod_{k\in I_A} e(P,P)^{\frac{a_k s}{x_j}}$$

Re-encryption : After the cloud server successfully generates the decryption token TK, it will carry on the re-encryption procedure. The cloud server randomly chooses $l \in Z_q$, and takes it, the public parameter of the system and TK as the inputs. The new ciphertexts is generated as follows:

$$D = \left(c_1, c_2, c, d_1 = lP, d_2 = l^2 Q, d = TK \oplus H_2\left(Z^{l^2}\right)\right)$$

User Data Decryption : Upon receiving the new ciphertexts, the user first renews the decryption token applying its global public key GPK_{uid} and global secret key GSK_{uid}.

$$TK = d \oplus H_2\left(e\left(x_j d_1, d_1\right) e\left(\pi_j, d_2\right)\right)$$

$$k \oplus H_2\left(Z^{s^2}\right) = \frac{c}{TK^{x_j}}$$

$$k = \frac{c}{TK^{x_j}} \oplus H_2\left(e\left(x_j c_1, c_1\right) e\left(\pi_j, c_2\right)\right)$$

3.2.5 Traceability

Some clients find a pirate device composed by the private key of some user U^* in an illegal trading platform. There is only one public commitment accompanying with this pirate device. The pirate device promises that it can decrypt any ciphertexts stemming from some access policies which the attribute set S_D can meet with. In order to disclose the specific identity of the pirate device, the user first randomly chooses $s, s' \subset Z_q$ and c', where c' is selected randomly from the space of the symmetric encryption key. Then, the user sets

$$g^{uid_1} = e\left(x_{uid_1}P, s^2P\right) e\left(Q, s'^2\left(y_{uid_1}P\right)\right); \cdots; g^{uid_t} =$$

$$e\left(x_{uid_t}P, s^2P\right) e\left(Q, s'^2\left(y_{uid_t}P\right)\right); \cdots g^{uid_N} =$$

$$e\left(x_{uid_N}P, s^2P\right) e\left(Q, s'^2\left(y_{uid_N}P\right)\right)$$

applying the random number $s, s' \in Z_q$ and the traceability list

$\{uid_1, x_{uid_1}P, y_{uid_1}P\}$;

$\{uid_2, x_{uid_2}P, y_{uid_2}P\}; \cdots \{uid_N, x_{uid_N}P, y_{uid_N}P\}$. Then the user chooses one appropriate matrix $\mathcal{M}'$ as the access policy which depends on the attribute set S_D promised by the private device. According to the system parameters and the access policy, the user will generate a ciphertex, which is *CT'=*

$$\left\{(M', \rho), sP, s'^2 Q, c' \cdot \left(\prod_{k\in I'_A} e(P,P)^{a_k}\right)^s, \left(C_i, D_{1,i}, D_{2,i}\right)_{i=1}^{l'}\right\}$$

in the PHR Encryption phase. Like the data access process, the user outsources the ciphertexts into the cloud server for re-encryption. Then, the pirate device must request the cloud server for the new ciphertexts in order to obtain its desired shared data. The cloud sever can only perform the token generation when the attributes the pirate device possesses satisfy the access structure in the ciphertext CT'. After the cloud server has successfully obtained the decryption token TK', it will randomly chose $l \in Z_q$ and transfer the new ciphertexts,

$$\left(c_1 - sP, s'^2 Q, c - c' \cdot \left(\prod_{k\in I_A} e(P,P)^{a_k}\right)^s, lP, d_2 - l^2 Q, d = TK' \oplus H_2\left(Z^{l^2}\right)\right)$$

to the pirate device. If the final feedback of this pirate device is $d \oplus H_2(g^{uid_t})$, it is easy to conclude that the user with the global unique identity uid_t is the creator of this pirate device.

4 Security and Performance Analysis

4.1 Security Analysis

The security analysis is performed in terms of the ability of resisting collusion attacks, semantic in data confidentiality, and the traceability with the characteristics of black-box and public. They are demonstrated by the following analysis.

4.1.1 The proposal is secure against the collusion attacks.

By the proposed scheme, each user is assigned with a global unique identity uid, and all the attribute private keys generated for the same user by each different attribute authority are related to the unique *uid* of this user after the attribute authority has successfully verified the certificate of this user. Moreover, on the basis of the unique aid of each attribute authority, all the attributes are distinguishable although some attribute authorities may issue the same attribute, which can availably prevent the user from privately replacing some components of an attribute assigned by the one attribute authority with those components from other secret keys assigned by another attribute authority. The security functionality against the collusion attacks is easy to

be implemented based on the abovementioned technological measures.

4.1.2 The proposal can achieve the semantic security in data confidentiality

Let's assume that the scheme cannot achieve the semantic security against malicious adversaries. Then, there is an adversary $\mathcal{A}$ that, given the system parameter and the public key created by the attribute authority, can break the scheme with the advantage ε. If so, we can construct an algorithm $\mathcal{B}$ that breaks the MDDH problem. The algorithm $\mathcal{B}$ is given a random MDDH instance $(P, A = sP, B = s^2P, C = g^z, U)$, where U is a random element in G_T. Then $\mathcal{B}$ chooses randomly a and sets the system parameter $(g, Q = aP, Z = C)$. Meanwhile, $\mathcal{B}$ also generates the attribute key $\{PK_k\}_{k\in S_A}$ instead of the attribute authority AA_k. Then, it sends the system parameter and $\{PK_k\}_{k\in S_A}$ to $\mathcal{A}$. $\mathcal{A}$ randomly chooses two messages m_0, m_1 on which it excepts to be challenged. $\mathcal{B}$ picks a random $b \in (0,1)$ and the related parameters as the process of data encryption of the proposed scheme, and sends the corresponding ciphertext

$$CT' = \left\{(M', \rho), sP, s'^2Q, m_b \cdot \left(\prod_{k\in I'_A} e(P,P)^{a_k}\right)^s, (C_i, D_{1,i}, D_{2,i})_{i=1}^{l'}\right\}$$

to the cloud server. Then the cloud server generates the new ciphertexts

$$\left(c_1 = sP, c_2 = s'^2Q, c = m_b \cdot \left(\prod_{k\in I_A} e(P,P)^{a_k}\right)^s, lP, d_2 = l^2Q, d = TK \oplus H_2(Z^{l^2})\right)$$

through the interaction with $\mathcal{A}$, where the successful generation of the new ciphertexts fully relies on the strong assumption in regard to $\mathcal{A}$, and sends the new ciphertexts as the challenge to $\mathcal{A}$. Then $\mathcal{A}$ outputs a guess $b' \in (0,1)$. At this point, $\mathcal{B}$ returns **1** if $b' = b$ and **0** otherwise. It is clear that if $U = Z^{s^2}$, the ciphertexts CT' is the encryption of m_b. Otherwise, since H_2 is randomly selected from a universal hash function family, CT' is the ciphertext of a random message, hence $b' = b$ with the probability 50%. In turn, the adversary $\mathcal{B}$ has an advantage of $\varepsilon/2$ in solving the MDDH problem.

4.1.3 The proposal can achieve the traceability with the advantage of public and black-box

First of all, we prove that given a black-box access to the pirate device constructed by one of all the users, one can always decide which one of them has created it. One randomly selects $s', s \in Z_q$ and sets $u_i = x_{uid_i}s^2 + ay_{uid_i}s'^2$ applying the random number $s, s', \in Z_q$ and the list $\{uid_1, x_{uid_1}, y_{uid_1}P\}$; $\{uid_2, x_{uid_2}, y_{uid_2}P\}; \cdots \{uid_N, x_{uid_N}, y_{uid_N}P\}$. Then one submits one randomized invalid ciphertexts

$$CT' = \left\{(M', \rho), sP, s'^2Q, c' \cdot \left(\prod_{k\in I'_A} e(P,P)^{a_k}\right)^s, (C_i, D_{1,i}, D_{2,i})_{i=1}^{l'}\right\}$$

to the cloud server. The cloud server generates new ciphertexts by the interaction with the pirate device. As a result, the new ciphertexts

$$\left(c_1 = sP, c_2 = s'^2Q, c = c' \cdot \left(\prod_{k\in I_A} e(P,P)^{a_k}\right)^s, lP, d_2 = l^2Q, d = TK \oplus H_2(Z^{l^2})\right)$$

can be obtained by the pirate device. Finally, if the output is $c' \oplus H_2(g^{u_t})$, then one claims that uid_t is the generator of the pirate device.It can be concluded that the proposed scheme is black-box traceability against malicious attacks. In the same way, in order to execute the black-box traceability procedure, all the global public keys and global secret keys (x_{uid_j}, y_{uid_j}) will be called. So, the proposed scheme is public traceability

4.2 Performance Analysis

The scheme in [14] has its contributions to solve the security issues to achieve the user key accountability by the multi-authority attribute based encryption mechanism with some overheads. The best way to demonstrate the outstanding advantages of the proposed MAAC scheme is to carry out the efficiency comparison between the proposed solution and the work in [14] in terms of the corresponding computation cost and communication overhead with three metrics, which are storage overhead, communication traffic demand and computation cost. In order to simplify the comparison, we list all the notations used in the comparison are described as follows.

$$\begin{cases} N_A: \text{the number of the attribute authorities in the system} \\ S_U: \text{the number of the users in the system} \\ S_{n,k}: \textit{the total number of attributes managed by one } AA_k \\ S_{n,k,uid}: \text{the number of attributes assigned to the user with } uid \text{ from one } AA_k \end{cases}$$

The storage overhead is an important metric that determines whether the clients with limited resources will choose the corresponding application. By analysis, it is found the main difference between the two schemes on the storage overhead coming from the following aspects. One is the number of the public keys that each attribute authority needs to reserve. The other is the number of attribute private keys that the user obtains from the attribute authority. The comparison results based on the above analysis are shown in Table I.

Entity	Proposal in [14]	MAAC
AA_k	$3S_{n,k}$+3	$S_{n,k}$+2
Client	$3\sum_{k=1}^{N_A} S_{n,k,uid}$+1	$\sum_{k=1}^{N_A} S_{n,k,uid}$+2

The communication cost incurred by the MAAC scheme is rather lower than that of the solution in [14]. It is easily to find that the communication cost of our solution is linear to the number of AA_k. However, the communication cost of the scheme in [14] is much higher challenging for those users who access the shared data by mobile devices with the limited

communication ability. The comparison of communication cost can be described on account of the initialization and traceability operations between the two solutions for the systems, which is shown in Table II.

By the MAAC proposal, the cloud server generates the new ciphertexts on the basis of successfully accomplishing the

Operation	Proposal in [14]	MAAC
System initialization	$O(S_A{}^2)$	$O(S_A)$
Traceability	$O(S_U)$	$O(S_U)$

attribute decryption process, which has almost completely offloaded the computation burden from the clients with limited computing resource to the cloud server. But by the scheme in [14], the attribute decryption for the shared data access is performed by the user devices. It brings huge computational burden to the mobile devices, which have the limited computing ability. On one hand, due to the interference of the pseudorandom function, the process of generating private keys draws into a great deal of power calculations. On the other hand, the substance of the solution in [14] is to achieve the user key accountability by appending the identity test in the decryption algorithm. The repeated decryption attempts before a successful decryption bring additional power and bilinear mapping operations. Hence, it is no necessary to quantify the gap between these two schemes as the difference is so large. In a word, our proposed MAAC scheme has incomparable superiority in computation cost than the work in [14].

5 Conclusions

In this paper, we have proposed a secure and efficient access control scheme, MAAC, for the cloud computing based PHR service. Our proposed MAAC scheme has applied a new multi-authority CP-ABE scheme, by which the attribute decryption can be outsourced to the cloud servers. In addition, the proposed MAAC scheme can achieve the user key accountability in an effective way. Moreover, the proposed MAAC scheme is efficient with much less computation costs and communication costs.

6 References

[1] M. Li, S. Yu, Y. Zheng, K. Ren, and W. Lou, "Scalable and Secure Sharing of Personal Health Records in Cloud Computing Using Attribute-based Encryption," *IEEE Transactions on Parallel and Distributed Systems*, Vol. 24, No. 1, pp. 131-143, 2013,.

[2] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," US Institute of Science and Technology, Report at http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf, 2011.

[3] J. Anderson, "Computer Security Technology Planning Study," Air Force Electronic Systems Division, Report ESD-TR-73-51, http://seclab.cs.ucdavis.edu/projects/history/, 1972.

[4] S. Yu, C. Wang, K. Ren, and W. Lou, "Achieving Secure, Scalable, and Fine-Grained Data Access Control in Cloud Computing," *Proceedings of IEEE INFOCOM 2010*, pp. 534-542, 2010

[5] K. Yang, Z. Liu, Z. Cao, X. Jia, D. S. Wong and K. Ren, "TAAC: Temporal Attribute-based Access Control for Multi-Authority Cloud Storage Systems," *IACR Cryptology* ePrint Archive, pp. 651-651, 2012.

[6] A. Sahai and B. Waters, "Fuzzy Identity-Based Encryption," *Proceedings of EUROCRYPT*, Vol. 3494 of LNCS, pp. 457-473, 2005.

[7] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-Based Encryption for Fine-Grained Access Control of Encrypted Data," *Proceedings of ACM Conference on Computer and Communications Security 2006*, pp. 89-98, 2006.

[8] J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-Policy Attribute-based Encryption," *Proceedings of IEEE Symposium on Security and Privacy 2007,* pp. 321-334, 2007.

[9] M. Chase, "Multi-authority Attribute Based Encryption," *Proceedings of TCC2007*, Vol. 4392 of LNCS, pp. 515–534, 2007.

[10] V. Bozovic, D. Socek, R. Steinwandt and V. I. Villanyi, "Multi-authority Attribute Based Encryption with Honest-but-curious Central Authority," *Cryptology* ePrint Archive, Report 2009: 083, http://eprint.iacr.org/

[11] H. Lin, Z. Cao, X. Liang, and J. Shao, "Secure Threshold Multi-Authority Attribute Based Encryption without a Central Authority," *Proceedings of INDOCRYPT2008*, Vol. 5365 of LNCS, pp. 426–436, 2008.

[12] K. Yang, X. Jia, K. Ren, and B. Zhang, "DAC-MACS: Effective Data Access Control for Multi-authority Cloud Storage Systems," *Proceedings of IEEE INFOCOM2013*, pp. 2895–2903, 2013.

[13] J. Li, K. Ren, B. Zhu, Z. Wan, "Privacy-aware Attribute-based Encryption with User Accountability," *IACR Cryptology*, ePrint Archive 2009.

[14] J. Li, K. Ren, K. Kim, "A2BE: Accountable Attribute-Based Encryption for Abuse Free Access Control," *IACR Cryptology,* ePrint Archive 2009.

[15] Z. Liu, Z. Cao, and D. S. Wong, "White-box Traceable Ciphertext-policy Attribute-based Encryption Supporting Any Monotone Access Structures," *IEEE Transactions on Information Forensics and Security*, Vol. 8, No. 1, pp.76–88, 2013

[16] Z. Liu, Z. Cao, and D. S. Wong, "Black Box Traceable CP-ABE: How to Catch People Leaking Their Keys by Selling Decryption Devices on Ebay," *Proceedings of ACM Conference on Computer and Communications Security 2013*, pp. 475-486, 2013.

[17] J. Li, Q. Huang, X. Chen, S. S. M. Chow, D. S. Wong, and D. Xie, "Multi-authority Ciphertext-policy Attribute-based Encryption with Accountability," *Proceedings of ACM ASIACCS 2011*, pp. 386–390, 2011.

ISBN: 1-60132-437-5, CSREA Press ©

SESSION

HEALTHCARE AND PUBLIC HEALTH RELATED SYSTEMS

Chair(s)

TBA

ISBN: 1-60132-437-5, CSREA Press ©

ISBN: 1-60132-437-5, CSREA Press ©

e-Health Diaries for People at End-of-Life: "A crutch to lean on"

C. Wilson[1], P. Ormandy[1], C. Vasilica[1] and S. Ali[2]
[1]School of Nursing, Midwifery, Social Work and Social Sciences, University of Salford, Salford, Greater Manchester, United Kingdom
[2]Health and Social Care, University of Salford, Salford, Greater Manchester, United Kingdom

Abstract – *The objective of this article is to explore the use of e-Health diaries in palliative care. 68 patients from three hospices in the UK were involved in the study. A sample of 14 patients was selected for diary analysis and focus groups. The qualitative data was examined using thematic analysis and findings exposed the different ways patients used their diaries, who they wrote for and what they revealed about their personalized care needs. e-Health diaries are invaluable at providing the patient with a voice and creating rich information for healthcare professionals.*

Keywords: e-Health diary, blogging, end-of-life care, personalized care, qualitative data.

1 Introduction

e-Health is an emerging term relating to the use of the Internet and other technologies to enhance health care services and the information delivered to stakeholders involved in healthcare practice.[1] Consequently, e-Health's primary outcome is to improve patient well-being through communication and information technology.[2] An e-Health diary allows the patient to write a web or app based diary that is shared with their care team. The use of e-Health diaries is rare, although, they have been successful at improving the self-efficacy of patients with cardiovascular disease[3], Crohn's disease[4] and diabetes.[5]

Technologies, such as e-Health, have emerged in end-of-life care to enhance the communication between healthcare professional's (HCPs), patients and families, to expose and foster adherence to patients' care wishes.[6] These technologies include informative websites,[7] videos and telemedicine. However, very few interventions are patient focused,[8] with the exception of patient held medical records.[9] Most are aimed at providing information for caregivers and HCPs.[8,10] No known e-Health interventions involve patients nearing the end-of-life writing an e-Health diary, however, there are accounts of patients writing blogs about their medical conditions.

Blogging is a proficient social media feature enabling users to express themselves through online entries in a chronological order.[11-12] An important characteristic of blogging is the stimulation of communication,[13] allowing users to reflect on personal experiences, share and seek opinions, and release emotional tension. Users' blog for a number of reasons; three prominent types of blogs include, individual created entries, a mashup of information curated from other sources and knowledge entries.[14] Individual blogs known as personal journals, diary entries and online diaries embody the blogger's own experiences and views.[15]

In the health context, primarily blogs are used by patients to self-manage a health condition or achieve a specific health goal.[16] Patients also blog to communicate other issues rather than just health.[17] Evidence suggests that the therapeutic outcomes of blogging include being able to express emotion, decreased feeling of loneliness, emotions management and finding satisfactory information.[18-19] Furthermore, by sharing personal stories and reflecting on the process/services, users are able to support peers in decision-making practice[20], which in turn empowers people to actively be involved in their healthcare.[21]

Issues and concerns patients have regarding posting blogs range from a fear of being judged for their own opinions and behavior, maintaining their privacy, blogs being edited, receiving negative opinions and a lack of interest from HCPs.[22] However, blogs written by patients are often useful for physicians to gain a deeper understanding of their patients.[23] Additionally, clinicians can join their patient bloggers online to signpost users to information as a coping mechanism.[17]

e-Health diaries differ to blogs as they are not publically displayed on a website, instead they offer patients a platform to write a diary that can be shared with whomever the patient chooses. This can help patients overcome fears of being judged and loss of privacy.[22] Moreover, e-Health diaries ensure that HCPs take an interest in what the patient is writing by acting as a written communication between patient and HCP. Unlike most other e-Health interventions in end-of-life care, the diaries provide the patient with a means to express their wishes, fears, feelings and thoughts. This approach is innovative and as such, the qualitative research is exploratory in nature.

1.1 Research questions

This paper explores the use of an e-Health service, which provides patients with a diary writing function. The use of the diary by patients nearing the end-of-life will be analyzed by referring to the following research questions:

1) In what ways are end-of-life patients using e-Health dairies?
2) Who are the e-Health dairies being written for?
3) What do the diaries tell us about patients nearing the end-of-life?

2 Methodology

2.1 VitruCare™

The e-Health service that the patients were asked to trial is called VitruCare™. VitruCare™ is web and app based system that offers different patients different services, from action planning and goal setting for patients with long term conditions to well-being services for patients nearing the end-of-life. The diary forms one microapp available to patients in end-of-life care. It has a 3000 character allowance and can be written as often as the patient chooses. The diary is shared with the patients' chosen care team. The care team can include anyone the patient wishes from nurses, doctors, physiotherapists to family members and next door neighbors.

The patients were asked to trial VitruCare™ for three months, during which time they had access to a variety of microapps such as 'Introducing Me' and 'How I Feel Today'. The most popular microapp was 'How I Feel Today', which included the diary element of the service. The present paper explores the use of this online diary by patients and HCPs.

2.2 Sample

Sixty-eight patients were recruited from three different Sue Ryder hospices that offer palliative care in the United Kingdom (UK). These patients had been diagnosed as nearing the end-of-life according to the Gold Standard Framework, however, for the purpose of the trial they were perceived to have longer than three months to live by the clinician. A sample of 14 patients (7 male; 7 female) was selected to represent the use of the diary. This purposive sample was selected to represent a variety of ages, medical conditions, gender and type of diarist. The average age of the participants was 57 years with complex health conditions including neurological, cancer, and respiratory disease.

A group of 45 HCPs were involved with the 68 patients using VitruCare™, working for either the referral units or the hospices. Again a purposive sample of 11 HCPs were selected to take part in focus groups to gather qualitative data exploring their experiences of monitoring patients through online patient diary information entered on VitruCare™. The sample was selected to provide a range of HCPs from all three hospice sites, in a variety of roles such as day therapy nurses, clinical nurse specialists and consultants. They were aged between 31 and 62, with an average age of 49 and all were female as there were no male HCPs involved in the care of the patients.

2.3 Data collection

Following ethical approval from the UK National Research Ethics Committee and the Hospice research governance committee, qualitative data was collected from the sample of 14 patients through two different methods. Firstly, all patients consented to the data they entered into VitruCare™ being used for research purposes. As such, each participant's diary entries, for the three months they took part in the trial, were collated and anonymized through the Microsoft CRM system. There were 899 separate entries, which average at 64 entries per participant. Secondly, four focus groups each with two or three participants (well enough to attend) were held with 10 patients on the trial. There was a focus group at each of the three hospice sites, and two focus groups at one of the hospices where there were a larger number of users. All participants provided consent to be part of a focus group and for the discussion to be recorded. Patients were asked about their use of VitruCare™; how often they used it, what they used it for and which function they used most. They were also asked about the impact, if any, that the online diary had on their lives.

In addition, rich qualitative data was collected from the sample of 11 HCPs through focus groups. Three focus groups, one at each of the three hospice sites, were held with three to four HCP participants. All of the HCPs involved consented to attending the focus group and the discussion being digitally recorded. Similarly, the HCPs were asked about their use of VitruCare™; how often they monitored patient entries, what information they reviewed or was of interest and how they used this information. They were also asked about what impact the online diaries had on patient care and their role as an HCP.

2.4 Data analysis

All focus groups were transcribed and patient diary entries collated into date/time ordered qualitative data narratives. The data were analyzed by three experienced qualitative researchers using thematic analysis. Initially the researchers independently familiarized themselves with the data and developed preliminary coded themes alongside detailed notes of thematic meaning. These preliminary codes and meanings were then discussed, agreed and developed into a comprehensive coding guide, ensuring consensus meaning and understanding across the three researchers. The data were then coded using Nvivo 10 for Windows and second coded for validity and reliability by another researcher. From the coding process, overarching themes were developed that depict the data collected and answer the research questions. Under the research questions, three core themes: *Type of diarist*, *Who do the patient write for?* and *About the patient* will be presented and discussed in the results section.

3 Results and Discussion

3.1 Type of diarist

It became evident from the patient diary data that there were two main types of diarists. Patients appeared to use the diary for different reasons. The first of which was to log symptoms, daily activities and health indicators as a record for the patient care team or self-management of health. The second included using the diary therapeutically and as a narrative to expel thoughts, feelings and emotions. These diarists wrote their entries privately, excluding family and friends, but were still

aware that their HCPs could read it. A number of patients also began using their diary as a simple log but over time increased the detail in their entries to include emotions, feelings and intricate stories. These types of diarists tended to warm towards using the diary in their end-of-life care.

Out of the sample of diary entries, nine patients utilized their diaries to log moments in their health and in their lives, similar to people who develop blogs to self-manage their health[16] and previous e-Health diaries used to tackle diabetes[5] and cardiovascular disease.[3] These moments included daily activity, general mood, symptoms, appointments, hospice treatment, hospital treatment, medication and any self-managed treatments and demonstrate the complexity of patients' treatments and interactions with their care team. Brenda wrote daily in her diary:

Brenda: "Ached a bit today after physio, especially neck and diaphragm. Blood sugar down to 2.2 today with what felt like a dumping syndrome episode. Social worker came today and she will be ringing the hospice about the drop in, and what care I may need. Weight 39.8 Ventilator 2 x 20 mins."

Patients would write about what had happened in the immediate past; for instance, that day or that week and very rarely would they recollect events from the distance past or write about what may happen in the future. There was only one patient (Philip) who thought about the future and even then it was merely fleeting comments and not in too much detail:

Philip: "I've got to get info on making a Will I've put it off too long."

Philip: "I wonder what's gonna be next."

The entries were usually regular and often daily, to the extent that if a patient missed an entry, they felt like they should justify why, demonstrating that the diary acted more as a regular log of moments related to the patients' health. Jane's entry indicated a sense of obligation to update her diary regularly:

Jane: "Had a very bad day, (14/07/2015), that's why I had NOT filled in my diary, thought I had till I checked it now."

In accordance to tele-medically augmented palliative care,[24] additional reasons that patients struggled to complete their diaries included illness, death, tiredness, going on holiday, being admitted to hospital and having technical difficulties. Illness, death and tiredness were difficult to overcome as barriers to usage, however, other obstacles could be reduced. For instance, as suggested by Nemecek et al.,[24] there was readily available technical support for patients and technical problems were continuously being improved by VitruCare™ to prevent issues stopping patients using the service. Moreover, five of the 14 patients in the sample had admissions to hospital, which prevented them from using VitruCare™ due to poor connectivity issues. As digital health evolves and use of social media increases[25-26] certain web or app based systems will become paramount to patients' care plans. It is therefore important for the chosen e-Health service to be included in a patient's medical records and for hospitals and hospices to provide the necessary resources and appropriate internet access.

The second type of diarist is the patient who uses VitruCare™ as a reflective diary to record their ongoing thoughts and feelings. This emotion management has been previously adopted by people with cancer[18] and young adults with depression who blog about their experiences.[19] Five out of the sample of 14 diarists, were patients who utilized the diaries in this way. For these patients, the diary became less of a log of symptoms and treatments and more of a release of emotions about what was happening to them. Stephen in a patient focus group spoke about how he used the diary as an emotional outlet:

Stephen: "I'm using it as a sort of crutch to lean on…I kind of treat it as somewhere I can write down some of this stuff at the same time, just reading it, where I'm at and so on because I can't discuss it with my family at the moment."

These patients, through the recollection of stories[21] from the immediate past and the distant past develop a narrative that expresses: their fear of their illness and dying, frustrations at changes in lifestyle, feelings of guilt, sadness of being alone and happiness when interacting with others. Charlotte, who struggled with chemotherapy, felt too young to be dying and often expressed fears of what was going to happen:

Charlotte: "Feel sad that mobility is so impaired and frightened at the thought of going into hospital in the poor condition that I am currently in"

These emotive narratives tell us less about the daily symptoms and record of treatment but more importantly provide an invaluable insight of the emotions and mental well-being of people approaching the end-of-life. Chung and Kim[18] discovered that cancer patients and companions write blogs for four main reasons; prevention and care, problem-solving, emotion management and information sharing. Innovators and designers of e-Health for end-of-life care should consider these patient requirements when developing their software, especially as VitruCare™ has been used for similar reasons; to share their information with HCPs, friends and family and to manage emotions by writing emotive narratives. The system also provides a secure messaging facility for patients to ask about care and solve any health related problems.

3.2 Who do the patients write for?

Diaries are usually entirely private, written by the individual to reflect on moments in their life. They are therefore written for the person doing the writing however, a diary in an e-Health context is different. These diaries are completed by individuals as patients and as part of their healthcare.[4] The patients are aware that their care team can see the information that is being written and so their diary is not just written for themselves but

also for others. The following section explores who the patients are writing for and the themes connected with these interactions by exploring the focus group data.

All diarists can choose who they share their diary with by selecting people to be part of their extended care team. The patients who log symptoms, treatments, mood and activity tend to write their diaries for a mixed audience of HCPs, family and friends. These diarists are similar to bloggers and they use their diary to inform their care team of what is going on, mostly so that they do not have to constantly repeat information.[21] In a focus group, James spoke about the reasons he started using VitruCare™:

James: "I started using it because I was thinking of doing a blog just to get thoughts down somewhere, maybe share them and so what I was finding was that I was telling people things and some would say, how are you, and I couldn't remember who I had told what, who I needed to update. So it would have been a way to get the basic facts down somewhere."

For these diarists, the main difference between writing a blog and using VitruCare™, is that a variety of HCPs are looking at their information and would be able to see if any emerging problems and be able to act on these issues. This level of interaction that the online system provides, reassures patients:

Brenda: "I feel more secure in my health because of it. That you've got someone there, day by day rather than having to wait between appointments, which can be a bit scary"

Similar to previous studies on palliative care[27] and hospice volunteering,[28] HCPs describe the care they are providing their patients as giving time, not saving time. Therefore, even though learning the intricacies of a new IT system can be challenging and time consuming, nurses and clinicians were aware of how the information that the patient is providing can enhance the face to face interaction:

Clinical Nurse Specialist: "It's preparation isn't it, it's that enhanced communication because we may not have seen them for two weeks previously…you know if they've had an uneventful two weeks or whether they've had a rough weekend, so at least you're well prepared."

HCPs did, however, raise concerns about whether the service was being used correctly by patients. They did not want patients to enter emergency information in case the system was not regularly and immediately checked by the care team. These concerns were around whether the interactive nature of the diary function would prevent the patient from contacting the correct emergency channels, whether the patient would know it was an emergency and how the HCP would feel if they had missed something. Patients, did seem to be aware that VitruCare™ was not for emergencies:

Brenda: "I see the diaries as your day to day, this is how I am, there's nothing urgent, no real problems but this is a log."

Also from the diary entries it was clear that patients were only reporting emergency episodes after they had happened and as a log for their extended care team. In these recollections, it was evident that patients were still following the correct emergency procedures despite the communicative nature of VitruCare™:

Paul: "Woke from afternoon sleep, unable to move off the bed luckily phone to hand, phoned Judith at work, unable to communicate properly. Judith came straight home after calling ambulance, had managed to crawl into hallway where I was laid on floor when Judith & paramedics found me."

The patients who use their diaries to therapeutically express thoughts, feelings and memories, write for different reasons to patients who use their diary as a log. Although they are aware that HCPs can see what they write, they are writing the diary to help themselves come to terms with their situations, similar to people writing blogs about health conditions.[18-19,21] These diarists do not want to share their diary with friends or family:

Stephen: "I wouldn't want my family reading what I was putting down"

The e-Health diary, unlike blogs, provides a platform for patients who wish to express concerns and emotions about their illnesses privately, away from family and friends. These diarists do write, however, knowing that HCPs can read the information and often if a patient is low in mood or is experiencing depressive symptoms, they express this by reaching out to their HCP:

Doris: "feeling shitty, most of my life has being so hard, now I'm having a hard time dying, no I don't want to die like this, don't want to die at all yet, but in this pain is evil, but I will not kill myself I promised my kids last year I would not."

Knowing this information, which is not always verbalized by patients, HCPs can help make the right decision concerning a patient's care. Previously blogs by terminally ill patients have provided valuable information for HCPs, especially the nursing community, but often from past reflection on the written material.[13] These, e-Health diaries place HCPs in a unique position where they can see a live snapshot into a patient's life and can personalize healthcare accordingly. From having access to the diaries the hospice care team have been able to diagnose patients with depression and anxiety and effectively treat mental health alongside physical health.

3.3 About the patient

Throughout the diaries, whether used as logs or self-reflective narratives, it became apparent how important social interactions were for the patients. The patients who had difficult family circumstances or struggled to accept help from others, often expressed how alone they felt, finding symptoms and lifestyle management more difficult. Stephen had been diagnosed with bowel cancer and because of his fear of

hospitals, he postponed his treatment and surgery until it was too late. He was consequently given a terminal diagnosis. He writes in his diary about the pressures that this decision has put his family relationships, especially his wife. When he is alone, his mood and symptoms are worse and he finds it challenging to care for himself:

Stephen: "A little bit of reality hit home today... after an argument with my wife I spent the night in my own home only to discover that preparing a hot meal was stretching my abilities… I got tired just standing in the kitchen and out of breath walking to and fro to sit in the living room about 10 steps away to sit down. In the end I swallowed what little pride I had left and asked her to pick me up."

Doris, suffering from Congestive Obstructive Pulmonary Disease (COPD) and anxiety found it difficult to accept help from her relatives. Her daughter would regularly visit but Doris was aware of how she was taking her frustrations, of not being able to do things for herself, out on her daughter. She therefore felt a tremendous guilt, but was reliant on help. When her daughter was not with her and other family members had not contacted her, she felt alone and scared:

Doris: "…feel so alone, no one to tell about how I feel, scared, don't know why."

Patients, even if their symptoms and pain were bad, had improved mood and feelings of happiness if they had contact with another person. For Doris, her reassurance came from a positive experience of HCPs at the hospice and how *'lovely'* there were to her:

Doris: "I am in the hospice at mo, everyone is so very nice, there all so lovely and caring…just had a foot massage from carol it was very nice and relaxing she is so good with it, she is such a nice person anyway, she's smashing."

Mood and emotions were also dramatically improved for all patients who had had contact with family members, whether this had been a phone call or a face-to-face visit. Jane mentioned how her symptoms and energy were still low but her mood was lifted by the visit from her daughter:

Jane: "My daughter, Naomi, staying with me for a few days & we are REALLY close & miss each other. So even though my symptoms and energy are RED, I am happier cos she's here and she looks after me so much, my pain levels are lower, (can't do anything), pampered by my loving daughter."

HCPs, family and friends of patients with a terminal diagnosis cannot underestimate the importance of social interaction. A positive social interaction with minimal effort can hugely improve a patient's quality of life.[29] Kubler-Ross[30] convey that the communication of people with a serious illness depends on which stage of the grief cycle the patient is at (denial, anger, bargaining, depression and acceptance). Patients at different stages often need more or less social interaction from loved ones either through technology or face to face.[31] The e-Health diaries have revealed that mood, depression, symptoms and treatments can prevent patients from socializing but when interactions with HCPs or family and friends do occur, the patient's mood is uplifted.

The diary data, particularly the patient logs, indicated the level of individual patient activity. It was evident, as time progressed, how a patients' ability to be active and social reduced as their illness and bodily function[32] deteriorated:

Charlotte: "Frank decided to take me to the local town for different scenery...disaster as I was unable to walk any significant distance…felt very insecure in an unfamiliar environment."

Even though social interactions and activities were important for patients, many were frustrated at not being able to do what they used to. The activities that patients could no longer do included baking, cooking, working, running errands, going out for lunch and coffee. Both James and Cheryl expressed their frustration and disappointment at not having the same level of freedom as they used to:

James: "Very moody and snappy on Sunday…probably a mixture of worry about work, finances and lack of independence."

Cheryl: "My friends are coming for lunch so that will be nice...we are not going out for lunch because I have backache. I feel bad about letting them down because they were looking forward to it."

The frustration that patients' felt towards losing physical and cognitive function coupled with how influential social interaction was on patient mood, highlighted the importance of providing patients with the means to be dignified and independent for as long as possible.[33] Treatments for independence could be as simple as efficiently providing the right painkillers and assistive technologies. James was frustrated at not being able to drive, not being independent, having to wait for his power chair. Once he was given his power chair, he found a new feeling of freedom:

James: "Good day. First trip out in my power chair with Lorraine. Went well if not a little cold...it felt great to be able to get up to the shops without any trouble."

There were, however, still problems as James found that not many areas had wheelchair access including the entrance to his own home and his doctor's surgery. When treating patients nearing the end-of-life, these diaries reveal how essential it is to provide personalized care tailored to the individual. There needs to be an awareness of the daily problems that people face so that independence[33] and social interactions[31] can be maintained for as long as possible. Web and app based digital services, can help HCPs understand the individual needs of the patient by providing a window into their lives.[13] This does

not hinder the face to face interaction but enhances the quality of care provided.

4 Conclusions

VitruCare™ developed an e-Health service for people nearing the end-of-life. It was used by 68 terminally ill patients from three different UK based hospices. The most popular function of the e-Health service was the 3000 character diary, which could be shared with HCPs, family and friends. This phenomenon, alongside the literature on blogging, indicates the desire of patients who have been diagnosed as end-of-life to write about their situation.

Two different type of diarists emerged from the qualitative analysis. The first writes for information sharing and self-management of health[21] whilst the other writes for emotion management by creating a rich narrative.[18] Innovators of e-Health should consider these two different styles when developing online diaries or blogs for patients in palliative care. Designs would need to incorporate the needs of both by providing enough character space for storytellers but also providing tools for patients to log symptoms, activity, patient reported outcome measures and health indicators.[4] Additionally, the level of privacy of the diary would need to be controlled by the patient as information sharers would want more people to read their entries whilst therapeutic diarists would want to keep their entries private.

Both types of diarist wrote their information for HCPs to read, providing a feeling of reassurance that HCPs would be able to see any changes in the mental and physical health of the patients and plan care accordingly. This information can also prove invaluable to HCPs, as an insight into the lives of people nearing the end-of-life, similar to information from patient blogs.[13] The difference being that the 'real-time' entries allow HCPs to action according to patients' needs. e-Health could therefore be paramount in understanding the patient and developing personalized healthcare[34] for people in palliative care.

Insights provided by the analysis in this paper reveal that patients, however low they are feeling, place significant value in social interaction and relationships. Often low mood is connected with being alone and good mood is related with positive interactions with HCPs, family or friends. This finding has powerful implications for innovators of connective technologies[31] such as video calling, social media and web-based communities. It also supports the work of befriending schemes, volunteer[28] and therapeutic services in hospices.

There is limited research on the use of e-Health diaries and blogs by patients nearing end-of-life. These mediums offer therapeutic services for patients whilst delivering valuable information for HCPs. Consequently, future research should encourage the use of e-Health in palliative care to give the patient a voice and improve personalized patient-led care.

5 Acknowledgements

The authors wish to thank Innovate UK for funding the project, Dynamic Health Systems for developing VitruCare™ and Sue Ryder for granting access to the three UK hospices and providing an invaluable project team.

6 References

[1] Pagliari C, Sloan D, Gregor P, Sullivan F, Detmer D, Kahan JP, Oortwijn W, MacGillivray S. "What is eHealth (4): a scoping exercise to map the field"; J Med Internet Res, 7(1), e9, 2005.

[2] van Gemert-Pijnen J, Nijland N, van LM, Ossebaard HC, Kelders SM, Eysenbach G, et al. "A holistic framework to improve the uptake and impact of eHealth technologies"; J Med Internet Res, 13(4), e111, 2011.

[3] Wolf A, Fors A, Ulin K, Thorn J, Swedberg K, Ekman I. "An eHealth Diary and Symptom-Tracking Tool Combined With Person-Centered Care for Improving Self-Efficacy After a Diagnosis of Acute Coronary Syndrome: A Substudy of a Randomized Controlled Trial"; J Med Internet Res, 18(2), e40, 2016.

[4] Kim ES, Park KS, Cho KB, Kim KO, Jang BI, Kim EY, Jung JT, Jeon SW, Jung MK, Lee HS, Yang CH. "Development of a Web-based, self-reporting symptom diary for Crohn's Disease, and its correlation with the Crohn's Disease Activity Index"; Journal of Crohn's & Colitis, Oct 2015.

[5] Rossi MC, Nicolucci A, Pellegrini F, Bruttomesso D, Bartolo PD, Marelli G, Dal Pos M, Galetta M, Horwitz D, Vespasiani G. "Interactive diary for diabetes: a useful and easy-to-use new telemedicine system to support the decision-making process in type 1 diabetes"; Diabetes technology & therapeutics, 11(1), 19-24, 2009.

[6] Ostherr K, Killoran P, Shegog R, Bruera E. "Death in the digital age: A systematic review of information and communication technologies in end-of-life care"; Journal of palliative medicine, Dec 2015.

[7] Gustafson DH, DuBenske LL, Namkoong K, Hawkins R, Chih MY, Atwood AK, Johnson R, Bhattacharya A, Carmack CL, Traynor AM, Campbell TC. "An eHealth system supporting palliative care for patients with non–small cell lung cancer"; Cancer, 119(9), 1744-51, 2013.

[8] Walczak A, Butow PN, Bu S, Clayton JM. "A systematic review of evidence for end-of-life communication interventions: Who do they target, how are they structured and do they work?"; Patient Education and Counseling, 99(1), 3-16, 2016.

[9] Cornbleet MA, Campbell P, Murray S, Stevenson M, Bond S. "Patient-held records in cancer and palliative care: a randomized, prospective trial"; Palliative Medicine, 16(3), 205-12, 2002.

[10] Germain A, Nolan K, Doyle R, Mason S, Gambles M, Chen H, Smeding R, Ellershaw J. "The use of reflective diaries in end-of-life training programmes: a study exploring the impact of self-reflection on the participants in a volunteer training programme"; BMC Palliative Care, 15(1), 1, 2016.

[11] O'Reilly T, Battelle J. "Web squared: Web 2.0 five years on"; O'Reilly Media, Inc, 2009.
[12] Bacigalupe G. "Is there a role for social technologies in collaborative healthcare?"; Families, Systems, & Health, 29(1), 1, 2011.
[13] Watson J. "The rise of blogs in nursing practice"; Clin J Oncol Nurs, 16(2), 215–217, 2012.
[14] Herring SC, Scheidt LA, Bonus S, Wright E. "Bridging the gap: A genre analysis of weblogs"; InSystem Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on 2004 Jan 5. IEEE.
[15] Bolander B. "Disagreements and agreements in personal/diary blogs: A closer look at responsiveness"; Journal of Pragmatics, 44(12), 1607-1622, 2012.
[16] Adams SA. "Blog-based applications and health information: Two case studies that illustrate important questions for Consumer Health Informatics (CHI) research"; International journal of medical informatics, 79(6), 2481-2488, 2010.
[17] Ngwenya NB, Mills S. "The use of weblogs within palliative care: a systematic literature review"; Health Informatics J, 20(1), 13-21, 2014.
[18] Chung DS, Kim S. "Blogging activity among cancer patients and their companions: Uses, gratifications, and predictors of outcomes"; Journal of the American Society for Information Science and Technology, 59, 297-306, 2008.
[19] Marcus MA, Westra HA, Eastwood JD, Barnes KL, Mobilizing Minds Research Group. "What are young adults saying about mental health? An analysis of Internet blogs"; Journal of medical Internet research, 14(1), e17, 2012.
[20] de Boer M, Slatman J. "Blogging and breast cancer: Narrating one's life, body and self on the Internet"; Women's Studies International Forum, 44, 17-25, 2014.
[21] Kim S, Chung DS. "Characteristics of cancer blog users"; J Med Libr Assoc, 95(4), 445–450, 2007.
[22] Ressler PK, Bradshaw YS, Gualtieri L, Chui KK. "Communicating the experience of chronic pain and illness through blogging"; Journal of medical Internet research, 14(5), e143, 2012.
[23] Wiesenthal A, Ross J, Cai K, Sanchez-Reilly S, Lin L. "When Cancer Blogging Helps with Healing More than One (S774)"; Journal of Pain and Symptom Management, 47, 511-512, 2014.
[24] Nemecek R, Huber P, Schur S, Masel E, Porkert S, Hofer B, Watzke H, Zielinski C, Binder M. "Telemedically Augmented Palliative Care: Empowerment for Patients with Advanced"; E-Health and Telemedicine: Concepts, Methodologies, Tools, and Applications, 183, Sep 2015.
[25] Hopewell-Kelly N, Baillie J, Sivell S, Harrop E, Bowyer A, Taylor S, Thomas K, Newman A, Prout H, Byrne A, Taubert M. "Palliative care research centre's move into social media: constructing a framework for ethical research, a consensus paper"; BMJ supportive & palliative care, Jan 2016.
[26] Nwosu AC, Debattista M, Rooney C, Mason S. "Social media and palliative medicine: a retrospective 2-year analysis of global Twitter data to evaluate the use of technology to communicate about issues at the end-of-life"; BMJ supportive & palliative care, Sep 2014.
[27] Giuliani L, Piredda M, Ghilardi G, Marinis MG. "Patients' Perception of Time in Palliative Care: A Metasynthesis of Qualitative Studies"; Journal of Hospice & Palliative Nursing, 17(5), 413-26, 2015.
[28] Claxton-Oldfield S. "Hospice palliative care volunteers: The benefits for patients, family caregivers, and the volunteers"; Palliative and Supportive Care, 13(3), 809-13, 2015.
[29] Hansen DM, Higgins PA, Warner CB, Mayo MM. "Exploring family relationships through associations of comfort, relatedness states, and life closure in hospice patients: a pilot study"; Palliative and Supportive Care, 13(2), 305-11, 2015.
[30] Kübler-Ross E. "Death"; Simon and Schuster, Jun 1997.
[31] Wallbaum T, Timmermann J, Heuten W, Boll S. "Forget Me Not: Connecting Palliative Patients and Their Loved Ones"; InProceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems 2015 Apr 18 (pp. 1403-1408). ACM.
[32] Håkanson C, Öhlén J. "Meanings and experiential outcomes of bodily care in a specialist palliative context"; Palliative and Supportive Care, 13, 625-33, 2015.
[33] Meier EA, Gallegos JV, Thomas LP, Depp CA, Irwin SA, Jeste DV. "Defining a Good Death (Successful Dying): Literature Review and a Call for Research and Public Dialogue"; The American Journal of Geriatric Psychiatry, Jan 2016.
[34] Reeve J, Cooper L. "Rethinking how we understand individual healthcare needs for people living with long-term conditions: a qualitative study"; Health & social care in the community, 21(1), 27-38, 2016.

A Context Driven Human Activity Recognition Framework

(Regular Research Paper)

Shatakshi Chakraborty, Chia Y. Han, Xuefu Zhou, and William G. Wee
Department of EECS, University of Cincinnati, Cincinnati, OH, USA
han@ucmail.uc.edu

Abstract - *A four-layer hierarchical framework for representing human body movements and characteristic behaviors, augmented with probability theory and statistical learning methods to discover and process complex activity signatures in a domain is presented in this paper. A typical waiting lounge scene in a health-care facility is used to illustrate the implementation and the power of this framework. This work shows how specific human behaviors can be collected, processed, represented, and interpreted directly in real world context, thus affording health care providers to understand patient's experience, both qualitatively and quantitatively, and allowing them to make recommendations for making work flow in clinics or hospitals more efficient, and ultimately improving the quality of care in their facilities.*

Keywords: human behavior recognition, gesture-based activity representation

1 Introduction

There is a growing demand of activity recognition in different areas of everyday living. A particular domain of interest is in the health-care sector, where patients under care can be monitored and assisted even when care provider is not readily around. In this paper, we will provide a framework that uses activity recognition theory to understand how human behaviors can be monitored, and consequently to identify the needs of a patient at a particular context. To illustrate the ability of the framework, we will use a simplified scenario, patients in lobby during the waiting time for doctor's consultation. Waiting rooms in clinic or hospital constitute small cosmos of human behaviors, many of which are well defined by a sequence of activities in health care workflow.

In today's health-care system, patients wait about 22 minutes on average in doctor's offices, and more than four hours in some emergency departments [1]. As wait time increases, patient satisfaction drops. With a growing consumer-mindedness of instant gratification or satisfaction, health care providers and institutions or hospitals are looking ways to improve productivity, like shortening each patient's path through the health care system, perhaps, adopting measures such as clinics using kiosks, and not reception desks, for speedier check-in for returning patients, and taking measures to funnel visitors to the appropriate part of the clinic or hospital when appointments have been arranged earlier, while providing more attentive face-to-face care to those who are first timers to the system and in need [16]. The purpose of this study is to investigate a computer-based means to obtain useful data on typical human behaviors during visits to clinics.

We will use two prior works to implement our system. A framework, proposed by Ma [2], that defines a four-layer hierarchical framework, where computer vision is used to study and understand human behavior through body movements. A second framework, developed by Saguna [3-6], uses probability theory and statistical learning methods to discover complex activity signatures, which can be performed sequentially, concurrently and interleaving with time. Additional modalities of information, such as speech, face expression, time-based contextual information can also be incorporated to interpret various human behaviors and elicit the cognitive processes used in analyzing the workflow of normal activities or social engagements.

The first sections of the paper explores the background and theory of the two frameworks: First is the 4-layer hierarchical framework, namely, the four layers are: 1) Feature extraction, 2) Simple behavior, 3) Individual behavior sequence, and 4) Pair-wise social interaction. The second is the Situation and Context-Driven Activity Theory. In the following sections, we created a typical waiting lounge scene, which depicts some very simple and common activities that a patient may be involved. As such, the result of this work would give health care providers and facility managers an understanding of their patients' experience and use this knowledge to improve the quality of the services they provide.

2 Review of frameworks

In this section we will introduce the key elements for representing and processing sensory data into human behaviors and activities given a context.

2.1 Ma's social interaction framework

The framework we are using in this context is an extension of the framework that has been proposed by Ma [2]. It is a four-layer hierarchical framework which mathematically describes pair-wise and individual social interactions by deriving body motion features from sensor data. Figure 1 gives

a pictorial representation of the social interaction framework. A bottom up view of the framework is explained as follows. The first layer extracts body motion features from the skeleton data, and objects and features from RGB data as received from the *Kinect* sensor. The localized room furniture features such as tables, doors, chairs, cabinets can be included to constrain the environmental setting. The second layer recognized 9 basic gestures for upper limbs and lower limbs. We classify the lower limb as Type-1 and upper limb as Type-2 behaviors. It includes 3 simple Type-1 gestures - standing, walking and sitting and 6 simple Type-2 gestures: waving, talking over cell phone, reading book/ magazine, sleeping while seated, seated relaxed, and making hand gestures while talking. Along with the behaviors, the user height is also considered in this layer. Currently, we are considering very specific and simple social roles for each person.

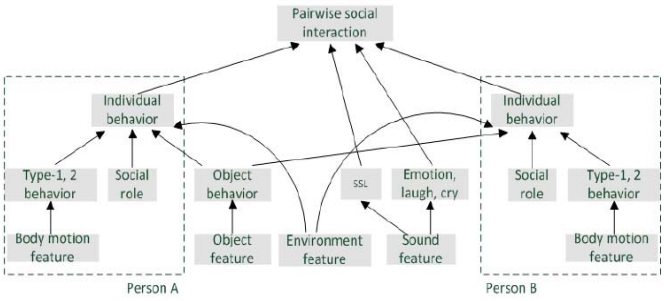


Figure 1. 4-layer hierarchical framework of social interaction

The third layer generates meaningful sequence of individual behaviors. The factors considered in this layer are derived from the Type-1 and Type-2 behaviors we defined in the previous layer alongside the surrounding environmental features, social roles and object behaviors. The fourth and the last layer generates pair-wise social interactions. It uses the same features as the third layer, but considers two persons in the frame at the same time. It also refers to the same Type-1 and Type-2 behaviors defined in the second layer. The framework is well defined to handle both individual and pairwise social interactions.

2.2 Saguna's situation and context-driven activity theory

As described by Winograd [7], the word "Context" has been derived from: *"con"* which means "with" and *"text"*. The use of *Context* is to infer *Atomic* activities and *Complex* activities based on situations. It has been observed that situations can be used to trigger actions by a person. *Complex Activities* occur when multiple atomic activities occur sequentially or interleaved in time, whereas, *Atomic Activities* are the simplified unit level activities that cannot be decomposed further given the application semantics. It can also be described as a leaf in the tree structure of the activity hierarchy. Figure 2 shows the low-level atomic activities forming into high-level complex activities. Another important concept is *Context attribute*. A context attribute C_i^t can be defined as any type of information *i* or data at time *t* that can define an activity or a situation(s).

The potential of activity theory lies in the attention that it gives to multiple dimensions of analyzing human engagement with the world. Saguna *et al.* proposed a model for recognizing multiple concurrent and interleaved complex activities using a Context Driven Activity Theory (CDAT) [3-6], which uses probabilistic data analysis for recognizing sequential, concurrent, and interleaved activities. Human activity recognition involves collection of training data, and applying activity recognition models based on different machine learning techniques to the training data sets to test the models.

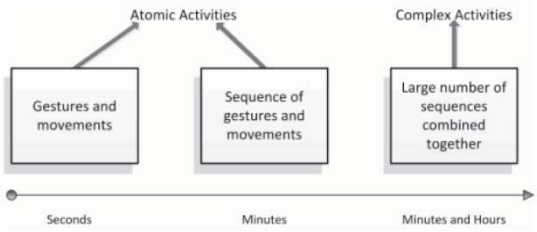


Figure 2. Atomic activities forming complex activities

Context/situation awareness is one of the major factors that describes human activity theory. Context can be defined as any information that that can be used to characterize the situation of entities [8]. Context can be distinguished in to two categories: *Physical context* and *Cognitive context*. *Physical* context can be defined as the environmental information or the sensor data, like, location, time, temperature, and more. *Cognitive* context includes mental states, preferences, tasks, and social affinities of the users [9]. Apart from contexts, another important factor to be considered for activity recognition is Situation. Every human activity is situation driven. According to Saguna *et al.*, "Situations are set of circumstances in which a human or an object may find itself".

2.3 Activity recognition approaches

The goal of activity recognition is to recognize mundane human activities in real life settings. Major sources of input data can come from a variety of sensors or cameras. Activities can be interpreted as a sequence of individual motions or gestures. The sequence can be linear or interleaved and even made up by concurrent subsequences. The start and the end of each subsequence are sometimes difficult to ascertain in complex activities. A common approach to activity recognition is the Hidden Markov Model (HMM), which is a generative probabilistic model, in which the sequence of observable states in time *t*, (y_1, y_2, ..., y_t) are generated by sequence of internal (unobserved) hidden states, (x_1, x_2, ..., x_t) [8]. As HMM makes modeling joint distribution only dependent on its immediate previous state and only on the current hidden state, there are still issues when dealing with complex concurrent and interleaved activities.

In reality, most of the activities are of non-deterministic nature, i.e., the atomic activities of a complex activity can be performed in any order and not necessarily in the same order every time. Another approach, the conditional random field (CRF) is more efficient in addressing such practical scenarios of non-deterministic activity recognition. CRF is a class of statistical modeling method for pattern recognition and machine learning problems [18]. In [18], a CRF is defined as a graph $G = (V, E)$ on observations X and random variables Y as follows: $Y = (Y_v)_{v \in V}$, so that Y is indexed by the vertices of G. Then (X, Y) is a conditional random field when the random variables Y_v, conditioned on X, obey the Markov property with respect to the graph: $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, where $w \sim v$ means that w and v are neighbors in G.

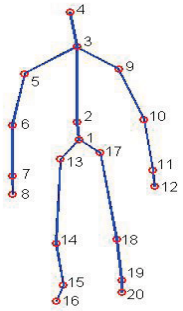
Though linear-chain CRFs are more flexible than HMM, the problem still exists that both CRF and HMM can only recognize sequential activities. In order to model more complex, interleaved and concurrent activities, more sophisticated models has to be used. Skip Chain CRF (SCCRF) is a linear chain that can be used to address the interleaving property of multiple goals. It uses multiple linear chains to capture activity variables with a larger distance between them or *long distance dependencies between the goals*. However, SCCRF is computationally expensive due to the high number of Skip Chains between goals [19].

Gu *et al.* in his paper proposed the emerging patterns based approach for activity recognition that can model sequential, concurrent and interleaved activities (epSICAR) [10]. Unlike other learning based models that uses training datasets for differentiating complex activities. An emerging pattern describes significant changes of two classes of datasets through feature vector for each complex activity. *Support* and *GrowthRate* are calculated for each attribute A. *Support*(A) is the ratio of the number of instances containing A in dataset and the number of instances in dataset. For two different datasets $D1$ and $D2$, the growth rate of attribute A from $D1$ to $D2$ is given as the following. *GrowthRate*(A) = 0 if Support1(A) = 0 and Support2(A) = 0; *GrowthRate*(A) = ∞ if Support1(A) = 0 and Support2(A) > 0; and *GrowthRate*(A) = Support2(A)/Support1(A), otherwise. These emerging patterns are mined from the sensor data and are used to compute interleaved and concurrent activities.

Recent work shows that Interleaved Hidden Markov Model(IHMM) [20], a variant of HMM, a variants of HMM, can be used to model sequential, interleaved and concurrent complex activities. Factorial Conditional Random Field (FCRF) [21], a variants of CRF, can also be used to recognize multiple concurrent activities for a single user, but the model cannot handle interleaved activities as well as multiple users. Like other training-based models, they require large training datasets to build the models for concurrent and interleaved activities. The issue being, in real life same activities are performed differently every time and hence gathering such huge training dataset can be difficult.

2.4 Body motion features

Body motion features can be interpreted by measuring the joint distance, joint angle or joint rotation speed. Microsoft *Kinect* sensors can provide body skeletons and 20 joint points [11]. Figure 3 shows the coordinates and the numbering of the joints. Joint distance can be typically useful in determining common actions, such as talking over cell phone. The distance between the joint 4(head) and the wrist left (7) or wrist right (11) becomes relatively less.



num	Definition	num	Definition
1	Hip Center	11	Wrist Right
2	Spine	12	Hand Right
3	Shoulder Center	13	Hip Left
4	Head	14	Knew Left
5	Shoulder Left	15	Ankle Left
6	Elbow Left	16	Foot Left
7	Wrist Left	17	Hip Right
8	Hand Left	18	Knew Right
9	Shoulder Right	19	Ankle Right
10	Elbow Right	20	Foot Right

Figure 3. Body feature points set and reference systems

Or, while clapping, the distance between the two wrist joints can be seen changing periodically. The typical joint distance pairs that can be considered for motion feature extraction are: (7, 3), (7, 4), (7, 5), (7, 9), (7, 10), (7, 11), (7, 1), (7, 14), (7, 15), (7, 18), (7, 19), (11, 3), (11, 4), (11, 5), (11, 6), (11, 9), (11, 1), (11, 14), (11, 15), (11, 18), (11, 19), (14, 18), (15, 19), (13, 15), (17, 19). The absolute position is the average value of joint points 1, 3, 5 and 9 with respect to the *Kinect* coordinates. Only x and z axis are considered. The absolute distance are converted to relative distance by dividing them by distance of skeleton (1, 3). Body motion features can also be detected based on the relative angles formed by the four limbs with respect to the torso plane. The four limbs can be given by the following pair of joint points: Right upper limb: (5,6), (6,7); Left upper limb: (9,10), (10,11); Right lower limb: (13,14), (14,15); and Left lower limb: (17,18), (18,19). Angular information between few feature geometries can be computed and used in determining the motions that characterize certain activity. The hands and feet joints of skeleton are very unstable and hence not considered for body motion feature recognition. Lastly, the joint rotation speeds can be derived by subtracting the rotation angle between the adjacent frames.

3 Our approach

In our research, we are using Microsoft Kinect V1 Sensor [12-15] to determine the context information. For simplicity and other limitations, physical context information such as time, temperature, weather conditions etc. and cognitive contexts such as preferences, social affinities are beyond the scope of the current work.

3.1 The complex activity recognition algorithm

Saguna et al. [3-6] proposed an algorithm for complex activity recognition which combines together the atomic activities and context information in order to infer a successful complex activity. The Complex Activity Recognition Algorithm (CARALGO) is given as shown in Figure 4. In it, symbols A_i, C_i, S_i represent Atomic Activities, Context Attributes, and Situations, respectively. CA_k is the Complex Activity. A_S, A_E, C_S, C_E are the start and the end of Atomic Activity and Context Attributes, respectively. T_L stands for life span of the activity. Set of Activities and Context Attributes are represented as γA_i and ρC_i. Sum of weights of Atomic Activity, Context Attributes, and Complex Activities are shown as, $w_{CA_k}^{A_i}$, $w_{CA_k}^{C_i}$, and w_{CA_k}, respectively. $T_{L_{max}}^{CA_k}$ is the maximum time taken for a complex activity and $w_{CA_k}^{T}$ is the threshold of weights of complex activity.

Input : A_i, C_i, S_i
Output: CA_k

1 **Initialization:**
2 findStartAtomicActivity(A_i, C_i);
3 check for current situation S_i;
4 findComplexActivitiesList(CA_{list});
5 **foreach** CA_k **do**
6 **if** ($A_i == A_S$) **then**
7 $add(CA_{list} \leftarrow CA_i = (\gamma A, \rho C, A_S, A_E, C_S, C_E, T_L)$
8 **end**
9 **end**
10 **return** CA_{list} ;
11 findComplexActivity(A_i, C_i);
12 **foreach** ($CA_{list} \leftarrow CA_k$) **do**
13 **while** *timecounter* $< T_{L_{max}}^{CA_k}$ **do**
14 **if** $A_i ==$ *element in* γA_i **then**
15 **add** $A_i \rightarrow \gamma A_i$ **and recalculate** $w_{CA_k}^{A_i}$ using recomputed weights
16 **end**
17 **if** $C_i ==$ *element in* ρC_i **then**
18 **add** $C_i \rightarrow \rho C_i$ **and recalculate** $w_{CA_k}^{C_i}$ using recomputed weights
19 **end**
20 **end**
21 **if** ($A_E, C_E found for CA_i$) ***and*** ($\rho C_i and \gamma A$ *are complete and* $w_{CA_k} \geq w_{CA_k}^T$) ***and*** *complex activity signature matched* **then**
22 $found CA_k$
23 **end**
24 **return** CA_k;
25 **end**

Figure 4. The CARALGO algorithm for activity recognition

CARALGO takes the list of atomic activities, context information and situations as input to infer complex activities. The algorithm starts by finding the initial start atomic activity and context and the current situation. It assigns a time window corresponding to the lifespan *TL* for each start atomic activity, A_S and start context attribute, C_S, that belongs to a complex activity CA_k. The algorithm also looks for duplicate lists of atomic activities, αA, contexts, ρC, and end atomic activity, A_E and end context, C_E within complex activity. It calculates the total weight w_{CA_k} according to Equation 1. If the weight is above the threshold weight, i.e. it matches Equation 2, it can be inferred that the complex activity has successfully occurred. All the time windows run laterally and the A_i and C_i are added for each complex activity CA_k at runtime until a successful match has been discovered. The initial weights that were assigned from domain knowledge, are then recalculated and updated and probabilities are analyzed accordingly. This helps remove the error of initial domain knowledge-based weight assignments.

3.2 Computing complex activity weight threshold and handling false positives

Each complex activity has a set of core atomic activities, CoreγA, and core context attributes, CoreρC, which determines the threshold weight, $w_{CA_k}^T$. The total weight of the complex activity w_{CA_k} should be more than the threshold weight for the complex activity to occur successfully. If the total weight w_{CA_k} is less than threshold weight, $w_{CA_k}^T$ can imply either of the following two reasons: *a*) The activity was started but abandoned before completion or *b*) The core atomic activities and context attributes did not occur for the particular complex activity. The initial value of the threshold weight $w_{CA_k}^T$ is simply assigned as the sum of the weights of the core atomic activities and core context attributes.

Complex activities can be inferred from atomic activities and context information. As mentioned earlier, each complex activity, CA consists of a set of atomic activities γA, and a set of contexts, ρC. The order of occurrence of an atomic activity for a complex activity is not considered in this work. This leads to a very practical situation that the atomic activities can be performed in any order for the complex activity to occur. Weights are assigned to each atomic activity A_i and context information C_i according to its importance for the complex activity CA_i to occur. We will denote the weight of atomic activity as $w_{CA_k}^{A_i}$ and weight of context as $w_{CA_k}^{C_i}$. The weights are assigned according to the following set of rules: 1) The core set of atomic activities and context are assigned higher weights than those which are of less importance; 2) If all the atomic activities are equally likely to take place for the complex activity, equal weights are assigned; 3) The sum of all the weights w_{CA_k} for each complex activity CA_k = 1; and 4) If an atomic activity, Ai or context, Ci does not occur in a complex activity CA_k, then $w_{CA_k}^{A_i} = 0$ and $w_{CA_k}^{Ci} = 0$. The sum of all weights, w_{CA_k} , must be above a threshold value, $w_{CA_k}^T$ for the complex activity CA_k to occur successfully. Thus, for a complex activity to occur successfully, w_{CA_k} is mathematically formulated as,

$$w_{CA_k} = \frac{\sum_{i=1}^{N} w_{CA_k}^{A_i} + \sum_{i=1}^{N} w_{CA_k}^{C_i}}{2} \quad (1)$$

where, $0 \leq w_{CA_k} \leq 1$, and

$$w_{CA_k} \geq w_{CA_k}^T \quad (2)$$

According to Saguna et al. [3], a complex activities can be defined by finding the associations between each atomic activities within the same complex activity as well as two different complex activities occurring together. Associations between atomic activity and complex activity can be determined by calculating probabilities of start, end and other atomic activities of the complex activity. We determine the

individual probabilities for the start and end atomic activities $(\alpha A_S, \beta A_E) \in A$ for complex activity CA_k

Probability of A_S, Pr(A_i) $\forall A_S$ in $\alpha A_S = \frac{total\ occurrence\ of\ A_i\ as\ A_S}{n}$

Probability of A_E, Pr(A_i) $\forall A_E$ in $\beta A_E = \frac{total\ occurrence\ of\ A_i\ as\ A_E}{n}$

Similarly, we calculate the individual probabilities for the start and end context attributes $(\alpha C_S, \beta C_E) \in C$ for the complex activity CA_k

Probability of C_S, Pr(C_i) $\forall C_S$ in $\alpha C_S = \frac{total\ occurrence\ of\ C_i\ as\ C_S}{n}$

Probability of C_E, Pr(C_i) $\forall C_E$ in $\beta C_E = \frac{total\ occurrence\ of\ C_i\ as\ C_E}{n}$

We consider all the atomic activities that lie between A_S and A_E, and all context attributes between C_S and C_E. The probabilities for every atomic activity: Pr(A_i, t) and all context attributes: Pr(C_i, t) is calculated by:

Probability of atomic activity, $\Pr(A_i) = \frac{total\ occurrence\ of\ A_i}{n}$, n being the sum of occurrences of all atomic activities.

Probability of context attribute, $\Pr(C_i) = \frac{total\ occurrence\ of\ C_i}{n}$, n being the sum of occurrences of all context attributes.

Associations are determined between different atomic activities within a complex activity by calculating conditional probabilities and transition probabilities ($p_{i,j}$) for different pairs of atomic activities (A_i,A_{i+1}) within a complex activity, CA = Pr(A_i | A_{i+1}, t). Markov chains are used to determine the associations between the atomic activities and context attributes in the complex activity.

4 A Typical set of complex activities

As stated, we have limited ourselves to just considering the patient's behavior at the waiting lounge. We have identified few. Figure 5 and 6 show a typical setting of the waiting lounge in a clinic. Table 1 presents typical common complex activities during the wait time that we have identified is such setting.

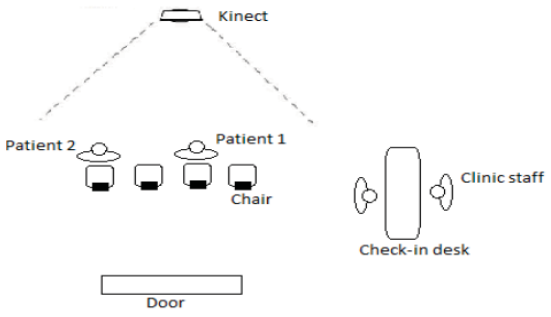


Figure 5. Typical setting of a waiting area in a clinic.

Figure 6. Captures of waiting area of a doctor's clinic

Useful context attributes will help constrain the understanding of the sequence of a given activity. In the instance of 'go get coffee' requires that person leave sitting area to go to drink area, and lights must be on, coffee machine must be on, mug and condiments available, etc. By observation, the activities of the patients depend on the environmental context as well as the amount of time spent waiting for his turn to go to exam room. A set of 20 context attributes are considered for the experiment. Sample attributes include: C_1-sitting area of waiting lounge, C_2-lights on, C_3-beverage center of waiting lounge, C_6-coffee condiments present, C_7-coffee mug available, C_9-coffee machine on, C_{14}-direction to sitting area.

	Complex Activitiy	Atomic Activities γA_l	Context Attributes ρC_l
CA_1	Fetching coffee	$A_{1,2,3,4,5,6,7,8,9,11}$	$C_{1,-1,2,3,9,13,7,6,14}$
CA_2	Talking/using a cell phone	$A_{11,1,2,10,12,13,14,15,16,17,32,33,45}$	$C_{1,2,-1,16,11,18}$
CA_3	Leaving for restroom	$A_{1,2,24}$	$C_{10,11,2,-1}$
CA_4	Walk to another person	$A_{1,2,31,17,24}$	$C_{1,2,19}$
CA_5	Fetching item from vending	$A_{1,2,27,28,24,29,30,31,20}$	$C_{1,4,-4,2,14,-1,15,5,8}$
CA_6	Talking to a neighbor	$A_{32,17,11,48}$	$C_{1,2,16,19}$
CA_7	Drinking or eating	$A_{11,23,47}$	$C_{1,2,16,14,-1,17}$
CA_8	Leaving for exam room	$A_{1,2,25,39}$	$C_{1,2,-1,12}$
CA_9	Filling out forms	$A_{11,2,35,36,37}$	$C_{1,2,16,-16}$
CA_{10}	Sleeping	$A_{11,21,41,42,43}$	$C_{1,2,16}$
CA_{11}	Reading	$A_{11,2,1,18,19,34,26,40,44}$	$C_{1,2,16,20}$
CA_{12}	Show discomfort/pain	$A_{11,46,22}$	$C_{1,2,16}$

Table 1. Common Complex Activities

5 Sample run and results

Each of the above activities belong to the Complex Activity class, consisting of one or more atomic activities and based on spatio-temporal information as context attribute. One of the major challenges for activity recognition is that the variations in the ways a user performs the same activity multiple times or different users perform one activity in various different ways. Users even tend to switch between different activities or perform them concurrently instead of an isolated manner. For our experiment, a dataset of 5170 frames from short video clips was collected in our lab. Figure 7 shows three typical behaviors in a waiting state - reading, falling asleep, and answering a call. In total, we conducted processing of the given 12 complex activities in order to validate the framework.

Figure 7: Sample video snapshots

5.1 A case study

We show an example case of "Patient performs the following complex activities concurrently and interleaved in time while waiting at the waiting lounge for checkup" and its related Atomic activity and Complex activity classes:
CA_1 = *"Fetching coffee"*
CA_2 = *"Uses cell phone"*
CA_3 = *"Talks to a friend who is accompanying"*
CA_4 = *"Walks over to reception for inquiry"*
The above complex activities (CA) can be represented as (CA_1|CA_2|CA_3|CA_4) as they can be performed concurrently or interleaving. For the lack of space, only the processing of the first activity is illustrated below.

$CA_k(w^T_{CA_k})$	$\gamma A(w^{A_i}_{CA_k})$	$\rho C(w^{C_i}_{CA_k})$	Core αA and ρC	A_S, C_S	A_E, C_E
Fetching coffee from coffee machine in waiting room (0.65)	A_1: walking (0.10), A_2: standing (0.10), A_3: coffee cup (0.15),A_4: coffee machine (0.15), A_5: dispense coffee (0.15), A_6: creamer (0.08), A_7: sugar (0.10), A_8: stirrer (0.04), A_9: lid (0.04), A_{10}: sitting in waiting lounge(0.08)	C_1: beverage area of waiting room (0.20), C_2: lights on (0.10), C_3: coffee machine on (0.25), C_4: cups available (0.20), $\neg C_1$: beverage area of waiting (0.20), C_5: coffee condiments present (0.05), C_6: sitting area of waiting room	A_1, A_2, A_3, A_4, A_5 and C_1, C_3, C_4	A_3, A_4 and C_1, C_3, C_4	A_{10} and C_6, $\neg C_1$

Table 2. Complex Activity example

In our example in Table 2, let us consider complex activity CA_1 = *"Fetching coffee from coffee machine in waiting room"*.
Start atomic activities $A_S = A_3, A_4$
Probability of A_S, $\Pr(A_3) = 0.40$
Probability of A_S, $\Pr(A_4) = 0.60$
End atomic activity $A_E = A_{10}$
Probability of A_E, $\Pr(A_{10}) = 1.0$
A_S = max $\Pr(A_i)$, where $A_i \in \gamma A_K$
A_E = max $\Pr(A_i)$, where $A_i \in \gamma A_K$
Hence, for our complex activity CA_1, $A_S = A_4$ and $A_E = A_{10}$
Start context attribute $C_S = C_1, C_3, C_4$
C_S = max $\Pr(Ci)$, where $C_i \in \rho C_K$
C_E = max $\Pr(Ci)$, where $C_i \in \rho C_K$
Similarly, for CA_1, $C_S = C_3$ and $C_E = \neg C_1$

The procedure for activity recognition involves the following major steps: identifying the CA_S, the A_S and C_S; carry out probabilistic analysis of CA_S, based on the probabilities, recalculating the weights of A_S and C_S; performing analysis to discover signatures of CA_k. Given that there may be several paths for the activity sequences, Markov chains are used to determine the activity signatures represented by the states of the paths. Various path probabilities can be analyzed to define the possible behaviors according to the known contexts. Table 3 shows the signature of this activity.

Complex Activity CA_k	Complex Activity Signature with Atomic Activities γA_i	Complex Activity Signature with Context Attributes ρC_i	Path Probability (γA_i) (ρC_i)
Fetching coffee (CA_1)	$A_{11} \rightarrow A_2 \rightarrow A_1 \rightarrow A_3 \rightarrow A_4 \rightarrow A_5 \rightarrow A_7 \rightarrow A_6 \rightarrow A_8 \rightarrow A_9 \rightarrow A_{23} \rightarrow A_1 \rightarrow A_{11}$	$C_1 \rightarrow C_2 \rightarrow C_{13} \rightarrow \neg C_1 \rightarrow C_3 \rightarrow C_7 \rightarrow C_9 \rightarrow C_6 \rightarrow C_{14} \rightarrow C_1$	(0.86) (0.91)

Table 3. Signature of the Complex Activity CA_1 example

Concurrent and interleaved activities can be further studied by using heat map (Figure 8) to see the most relevant occurrences in the path combinations. Clearly, we can see that the complex activity pairs (CA_2, CA_7), (CA_7, CA_{11}), (CA_6, CA_7) have the highest values for concurrency and interleaving. Other complex activities such as *"Talks to person sitting beside"* and "*Reading a book/ magazine"* are also often performed concurrently or interleaving in time with other complex activities.

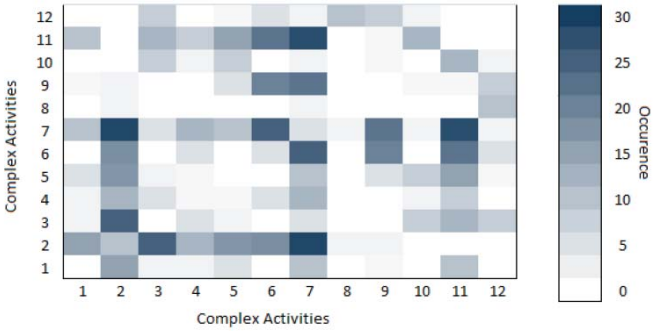


Figure 8 Heat map for viewing best path combinations

To validate the result visually, our system provides a simple user interface, showing the frame characterizing the state and overlaid with the input data, represented by the input skeleton joint points. Shown in Figure 9 is the state of a patient reading book/magazine. It is recognized as such. The outcome of the activity recognition is displayed on the right side of the viewing screen.

Figure 9. The system console showing the result of experiment.

6 Discussion and conclusions

We have developed a framework for human activity analysis and recognition. We presented a case study to show how a domain can be captured and represented. To our best knowledge, no other study has been conducted thus far to understand patient's satisfaction during a clinical visit based on real-time body movements and gestures of the patients in the waiting lounge. This common scenario can be extended to other settings involving subjects in any venue of our daily life. Many other human behavior study can thus be developed by using this framework.

7 References

[1] Entisar K. Aboukanda, Muhammad Latif, "The Effect of Patient Behavior on Wait Times in Emergency Departments", International Journal of Business and Commerce Vol. 3, No.6, pp. 18-31, Feb 2014

[2] Tao Ma, "A Framework for Modeling and Capturing Social Interactions", Ph.D. Dissertation, University of Cincinnati, December 2014.

[3] Saguna, Arkady Zaslavsky and Dipanjan Chakraborty, "Building Activity Definitions to Recognize Complex Activities Using an Online Activity Toolkit", IEEE 13th International Conference on Mobile Data Management, pp. 344-347, Aug. 2012

[4] Saguna, Arkady Zaslavsky and Dipanjan Chakraborty, "Recognizing Concurrent and Interleaved Activities in Social Interactions", IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, pp. 230-237, Apr. 2011

[5] Saguna, "Inferring Multiple Activities of Mobile Users with Activity Algebra", IEEE 12th International Conference on Mobile Data Management, pp. 23-23, Jun. 2011

[6] Saguna, Arkady Zaslavsky and Dipanjan Chakraborty, "Complex Activity Recognition Using Context-Driven Activity Theory and Activity Signatures", ACM Transactions on Computer-Human Interaction, Vol. 20, No.6, Article 32, Dec. 2013

[7] Terry Winograd, "Architectures for Context", Computer Science Department, Stanford University

[8] Hidden Markov Model, Wikipedia. Available: http://en.wikipedia.org /wiki/HiddenMarkov model

[9] Dey, A.K., Abowd, G.D. and Salber, D.2001. "A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware-Applications", Human-Computer Interaction, Volume 16, Number 2, pp. 160-175

[10] Tao Gu, Zhanqing Wu, Xianping Tao, Pung, H.K. "epSICAR: An Emerging Patterns based approach to sequential, interleaved and Concurrent Activity Recognition", Pervasive Computing and Communications, IEEE International Conference, March, 2009

[11] Microsoft, Coordinate Spaces, [online]. Available: https://msdn.microsoft.com /enus/library/hh973078.aspx

[12] Kinect, Wikipedia, http://en.wikipedia.org/wiki/Kinect

[13] Implementing Kinect Gestures, online. Available: http://pterneas.com/2014/01/27/implementing-kinect-gestures/

[14] Joint Orientation, [online]. Available: https://msdn.microsoft.com/en-us/library/hh973073.aspx

[15] Microsoft, Tracking Modes (Seated and Default), https://msdn.microsoft.com/enus/library/hh973077.aspx

[16] Clifford Bleustein, MD, MBA; David B. Rothschild, BS; Andrew Valen, MHA; Eduardas Valaitis, PhD; Laura Schweitzer, MS; and Raleigh Jones, MD, "Wait Times, Patient Satisfaction Scores, and the Perception of Care"

[17] J. C. B. Christopher, "A tutorial on support vector machines for pattern recognition", Data Mining and Knowledge Discovery, 1998.

[18] Lafferty, J., McCallum, A., Pereira, F. (2001). "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". Proc. 18th International Conf. on Machine Learning. Morgan Kaufmann. pp. 282-289.

[19] Derek Hao Hu and Qiang Yang. "CIGAR: Concurrent and Interleaving Goal and Activity Recognition", Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008)

[20] Joseph Modayil, Tongxin Bai, and Henry Kautz, "Improving the recognition of interleaved activities," Research note, in Proc. of the Tenth International Conference on Ubiquitous Computing (UBICOMP08), Seoul, South Korea, September 2008.

[21] Tsu-yu Wu, Chia-chun Lian, and Jane Yung-jen Hsu, "Joint recognition of multiple concurrent activities using factorial conditional random fields", In Proceedings of AAAI Workshop on Plan, Activity, and Intent Recognition, California, July 2007.

[22] Jason Weston, "Support Vector Machine (and Statistical Learning Theory) Tutorial", NEC Labs America, Princeton, USA

Measuring the Quality of Numeracy Skill Assessment in Health Domain

Mandana Omidbakhsh, Olga Ormandjieva
Computer Science and Software Engineering, Concordia University, Montreal, Canada

Abstract - *Healthcare represents a traditional personal service sector with established importance of patient numeracy profiles. This is of importance because patient numeracy skill is one of the key factors associated with poor understanding of treatment decisions and non-adherence to therapies. This paper aims to introduce a novel quality model for the evaluation of patient numeracy assessment methods. The new quality is used to compare numeracy assessment methods on a pilot study that helped us to establish the place of our previously published confidence-based numeracy assessment method C-PNA among the other numeracy assessment methods.*

Keywords: Accuracy, Healthcare, Quality Model, Numeracy Assessment, Subjective Characteristic, Objective Characteristic.

1 Introduction

Numeracy assessment in healthcare domain is noticeably an attractive topic which concerns the evaluation of the level of patients' numerical skill enabling them to understand and perceive the information related to their health. We proposed a confidence based adaptive testing model [1] that assesses the patients' numeracy skill by integrating the parameter of confidence in the adaptive assessment. In [2], we introduced our goal-driven modeling for Confidence-based Patient Numeracy Assessment named C-PNA.

Although a number of numeracy assessment methods have been used in the health domain, a key limitation of selecting the right method is that no quality model for evaluating numeracy assessment methods is available.

This paper describes the development of a novel model to measure different objective and subjective quality characteristics of numeracy assessment methods, inspired by the latest standard ISO/IEC 25022[3]. It provides a framework for the comparison of our method with any other existing numeracy assessment method.

The objective of this paper is two-fold: i) to present our new quality model, aligned with the ISO/IEC 25022 and adapted specifically to numeracy assessment in health domain, and ii) to use the new quality model to compare numeracy assessment methods on a pilot study. We conducted a pilot study, in which our Confidence Based adaptive Testing (CBT) method of C-PNA model was compared with existing Non-Confidence Based Testing (NCBT) methods for patient numeracy assessment. The results of our study demonstrate that our confidence-based adaptive testing method for the assessment of numeracy level of patients, C-PNA, has higher patient satisfaction, discretionary usage and trust than existing related work along with the same accuracy, but greater usage efficiency and remarkable effectiveness.

The organization of this paper is as follows. In Section 2, the research methodology is explained; we define our research problem, our objectives and the steps to achieve them. The quality model is introduced in Section 3. Our Website and support tool are presented in Section 4 and the pilot study is described in Section 5. Section 6 summarizes the literature on quality modeling for numeracy assessment methods and explains in what ways they are similar or different from our approach. Finally, in Section 7, we conclude the paper and outline the directions of our ongoing research.

2 Methodology

The first objective of this paper is to present our new quality model, aligned with the ISO/IEC 25022 and adapted specifically to our numeracy assessment method. The model is then used to conduct experiments aiming at evaluating the quality of the new CBT numeracy assessment method as compared to two classical NCBT methods: Lipkus [4] and NUMi [5].

To achieve the research objectives, we followed the steps as outlined below:

Step 1: Quality Model. In order to evaluate an assessment model, we had to determine the appropriate objective quality characteristics, which mostly influence on the numeracy assessment namely accuracy, effectiveness, productivity and usage efficiency. Furthermore, we identified the subjective characteristics related to the research problem such as satisfaction, discretionary usage and trust. The measurements designed to quantify these objective and subjective characteristics were also determined.

Step 2: Tool Support. Secondly, we designed and developed a web-based application for the quality evaluation of numeracy assessment to carry on the experiments [6]. After a series of experiments, the quality model was revised and then pilot tested using the Web-based application as described next.

Step 3: Pilot Study. The last step of our methodology is concerned with the empirical validation. We selected two classical numeracy assessment methods, Lipkus and NUMi, to enable a pairwise comparative measurement of the quality characteristics, and then we designed and conducted the experiments. Data were collected and validated during the execution of the experiments. These data were then analyzed and a comparison was performed with the results obtained using the alternative methodologies Lipkus and NUMi.

The pilot study provided the evidence about our theory and helped us establish the place of our confidence-based numeracy assessment method among the other numeracy assessment methods. Our novel quality model is introduced next.

3 Quality model

In developing numeracy assessment methods, not only high quality, reliable and efficient assessment is required, but also high personal satisfaction of the users should be taken into the consideration [3]. ISO/IEC 25022 standard provides a quality model definition, which could serve as a customer satisfaction model to ensure that all characteristics of quality are covered from the perspective of each stakeholder.

Here, we introduce our quality model, which is designed specifically for the purpose of numeracy assessment. The quality model is tailored in a way that facilitates the evaluation of such assessment systems in terms of accuracy, effectiveness productivity, usage efficiency, satisfaction, discretionary usage and trust.

For our study, we employed the quality characteristics both objective and subjective. The former is associated to sets of data, which depend only on the object that is measured, however, the latter not only depends on the object that is measured, but also on the viewpoint from which it is taken. The former includes accuracy, effectiveness, productivity, and usage efficiency that only depend on the object being measured. On the other hand, the latter includes comfort, pleasure, understandability, satisfaction, discretionary usage and trust that rely on the viewpoint from which they are taken as well.

In our hierarchical quality model, the quality characteristics are delineated through several layers. At the root of this structure, there is a division of characteristics into objective and subjective ones.

3.1 Objective characteristics

The quantification of the objective characteristics is based on numerical rules to ensure fairness of the assessment. In other words, it is assured that users produce same measurement results every time the measurement is undertaken on the same source and in the same context. This consistency of measurement is considered very important [7].

Each of the objective characteristics is defined as below.

3.1.1 Accuracy

Accuracy is the percentile of the numeracy assessment test results that are similar to the threshold (standard) test results. In other words, accuracy is indicated in terms of similarity of the results.

Generally, accuracy is described by answering to the question of: "What percent of our prediction were correct?" So, if we base the definition on the truthfulness of the reality and the prediction, accuracy is calculated as the ratio of prediction values that are the same as reality values over the total values true or false [8].

For our study, we took Lipkus as a standard for numeracy skill assessment (Reality) and then we compared the results of method CBT as a variation for numeracy skill assessment (Prediction) with the results of Lipkus. We calculated the percentile of users who fall in the same numeracy skill level in CBT as in Lipkus. For this purpose, we first obtained the scores of each user in the tests; we used box-plotting technique for categorizing their level of numeracy skill. There are three levels in this categorization: low, medium and high. We compared the results of each user in both tests and find the overall number of the similar results.

3.1.2 Effectiveness

Effectiveness is defined in terms of the coverage of categories of numeracy questions. Difficulty Level (DL) is a number assigned to each question in the question bank and it varies depending on the type of the question. It is calculated as the number of DLs covered without explicitly asking related questions to each DL. If all DLs are covered in the test, the test covers all types of questions, all categories of numeracy questions, and it means that the set of questions of the test is effective.

3.1.3 Productivity

Productivity is the number of questions asked in a specified test relative to the time taken to answer them by users. Generally, productivity is the output over input which here is the number of questions answered over time. We say users are more productive using the test if they answer more questions per unit of time.

3.1.4 Usage efficiency

The usage efficiency based on ISO/IEC DIS 25022 of the test is measured as an objective been achieved over a specific time. It is calculated as *the average time to cover one DL*. Our objective is to cover more DLs meaning obtaining more

coverage on different types of questions. Usage Efficiency is the time required to cover one DL, one category of numeracy question types.

Table 1 shows the definition of each of the objective characteristics discussed above along with their indicators. Table 2 introduces the base measures required for calculation of the objective characteristics with their measurement formulas and the measurement data interpretation as represented in Table 3.

Figure 1 depicts the objective characteristics of our hierarchal quality model, which is composed of four characteristics: accuracy, effectiveness, productivity and usage efficiency.

TABLE 1. OBJECTIVE CHARACTERISTICS

Objective Characteristic	Indicator	Definition
Accuracy	accuracy_ind	The percentile of our test results that is similar to the standard test results. The number shows the percentile of the users who fall in the same category in two different tests.
Effectiveness	effectiveness_ind	Number of DLs covered without explicitly asking related questions to each DL.
Productivity	productivity_ind	Number of questions answered in a specified test relative to the time taken by the user.
Usage Efficiency	usageEfficiency_ind	The usage efficiency of the test is measured as an objective been achieved over a specific time. Our objective is to cover more DLs meaning obtaining more coverage on different types of questions.

TABLE 2. DEFINITION OF BASE MEASURES

Base Measure	Definition
A	Answer to each question for each individual
DL	Number of DLs covered by each test for each individual
Q	Number of questions required to complete a test for each user
TNP	Total Number of Users
TNS	Number of Users in the same Category
T	Time required for the user to complete a test

TABLE 3. OBJECTIVE CHARACTERISTICS MEASUREMENT FORMULA

Indicator	Measurement Formula	Interpretation
accuracy_ind	= (TNP-TNS)/ TNP*100	Results close to 100% are ideal. Higher values indicate more accurate results.
effectiveness_ind	= DL	Results close to 100% are ideal. Higher numbers indicate higher effectiveness.
productivity_ind	= Q / T	Higher numbers show higher productivity.
usageEfficiency_ind	= DL / T	Higher numbers show higher efficiency in terms of usage.

We assume DL, T and TNP are always greater than zero.

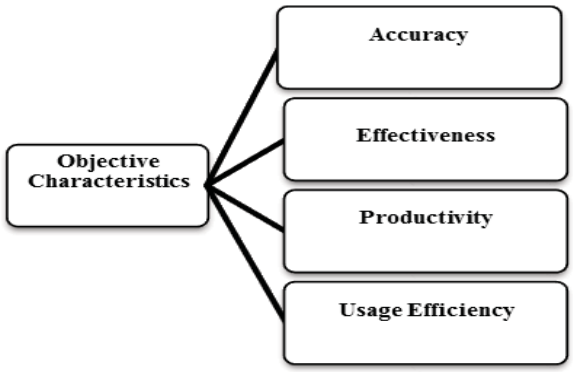


Fig.1.Objective Characteristics of Quality Model for Evaluation of Numeracy Assessment System

3.2 Subjective characteristics

Subjective characteristic measurements reflect the viewpoint of whom it is measured by. Basically, the viewpoints of users are obtained from the questions on the questionnaires presented to them after their experience using the system. To collect this qualitative data, users indicate the ratings on an ordinal scale. Consequently, this subjective characteristics quantification is engaged with human judgment [3].

Our subjective characteristics include: i) satisfaction characteristic which in turn concerns mainly on the comfort in answering the questions, the pleasure in writing the test, the understandability of the questions on the test, ii) the trust on the test results, and iii) the discretionary usage between two tests performed in one session. The subjective characteristics are measured on a Likert scale; the users are asked to rate their reaction to a statement along a scale for a type of survey question from a range of responses often from a positive rating to a negative rating with a neutral score in between. These subjective characteristics are listed as:

3.2.1 Satisfaction measures

Satisfaction measures based on ISO/IEC DIS 25022 assess the degree to which user needs are satisfied when a system is utilized in a specified context of use. The value of satisfaction can be an overall measure of satisfaction produced by combining measures of individual sub-characteristics, which could be in turn weighted according to the importance of them to the overall satisfaction. Users answer each question on the questionnaire by choosing one of the values on a scale ranging from strongly agree to strongly disagree. The sum of all sub-characteristics could be also transformed into a percentage.

Here, we in turn defined the users' level of satisfaction as a result of the pleasure in writing the test, the comfort in answering the questions, and the understandability of the test questions in each session. Table 4 shows the definition of satisfaction measures.

TABLE 4. SATISFACTION MEASURES

Measure	Description	Measurement function	Method
User Satisfaction	The overall satisfaction of the user	X = S(Xi) Xi sub-characteristics of satisfaction	Questionnaire

Table 5 summarizes our satisfaction measure for the purpose of our study.

TABLE 5. SATISFACTION INDICATOR

Measure	Description	Measurement function	Method
satisfaction_ind	The overall satisfaction of the user	X = Pleasure + Comfort + Understandability	Questionnaire

a.Comfort Measures:

Comfort measures based on ISO/IEC DIS 25022 assess the degree to which users' needs for physical comfort are satisfied. Physical comfort can be influenced by position or actions that the user has to make to use the computer system, and by the environment in which the system is used. It is shown as Table 6.

TABLE 6. COMFORT MEASURES

Measure	Description	Measurement function	Method
Physical Comfort	The extent to which the user is comfortable compared to the average for this type of system	X = A A = Psychometric scale value from a comfort questionnaire (See Table11)	Questionnaire

b. Pleasure Measures:

Pleasure measures based on ISO/IEC DIS 25022 assess the degree to which user needs for pleasure are satisfied. The needs of users encompass their desire to obtain new knowledge and skills, to communicate their personal identity, to provoke new pleasant memories and to be involved in the interaction. Table 7 shows the definition of pleasure measures.

TABLE 7. PLEASURE MEASURES

Measure	Description	Measurement function	Method
User Pleasure	The extent to which the user obtains pleasure compared to the average for this type of system	X = A A = Psychometric scale value from a pleasure questionnaire (See Table11)	Questionnaire

c. Understandability Measures:

Understandability measures assess the degree to which user understands the content of the questions on the test as defined in Table 8.

TABLE 8. UNDERSTANDABILITY MEASURES

Measure	Description	Measurement function	Method
Understandability	The satisfaction of the user with Understandability of system	X = A A= Response to a question related to understandability (See Table11)	Questionnaire

3.2.2 Trust measures

Trust measures based on ISO/IEC DIS 25022 assess the degree to which a user has confidence that a product or system will behave as intended. It is shown as Table 9.

TABLE 9. TRUST MEASURES

Measure	Description	Measurement function	Method
User Trust	The extent to which the user trusts the system	X = A A = Psychometric scale value from a trust questionnaire (See Table11)	Questionnaire

3.2.3 Discretionary usage

Discretionary Usage on the basis of ISO/IEC DIS 25022 is defined as the proportion of users who prefer one method to the other one as depicted in Table 10.

TABLE 10. DISCRETIONARY USAGE MEASURES

Measure	Description	Measurement function	Method

Discretionary Usage	The proportion of potential users choosing to use a system	X = A/B A= Number of users using a specific system B = Number of potential users who could have used the specific system (See Table11)	Measure user behaviour or automated data collection

The templates of Table 4 to Table 10 are inspired by ISO-IEC25022.

Table 11 shows the corresponding statements on the questionnaire for each of these subjective characteristics.

TABLE 11. SUBJECTIVE BASE MEASURE DEFINITIONS

Base Measure	Definition
Pleasure	The whole test was a pleasant experience to me.
Comfort	I felt comfortable going through the sequence of the questions in the test.
Understandability	It was easy to understand the questions.
Discretionary Usage	Personally, on the result of which method you prefer to have your numeracy skill assessed?
Trust	I trust the result of CBT.

Figure 2 demonstrates the subjective characteristics of our hierarchical quality model which is composed of satisfaction, discretionary usage and trust at one layer and satisfaction, itself, is included of comfort, pleasure and understandability at the next layer.

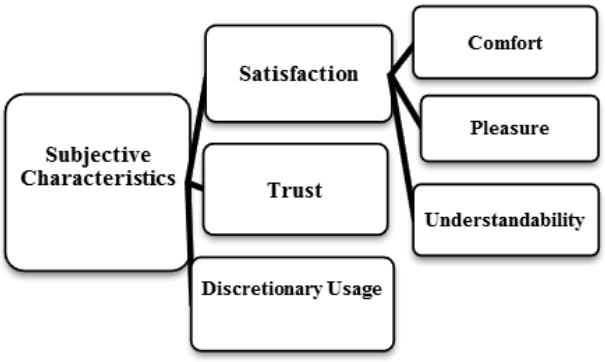


Fig.2. Subjective Characteristics of Quality Model for Evaluation of Numeracy Assessment System

4 Tool support

For the purpose of evaluating the quality of the new C-PNA method, we designed an online Web application [6]. The application enables us to create, run test sessions, and then save the results of the test sessions for further analysis. It facilitates the process of designing different test sessions with CBT and NCBT methods and adjusting the questionnaires based on the type of the tests and facilitates the comparison of results of different methods.

Our system has user and administrator levels. It has the following functionalities at administrator level:

i) Manage/list/add questions to question bank
ii)Add /rename question types to questions in question bank
iii) Manage/list/add new type of tests
iv) Manage/list/add test sessions
v)Manage/list/add survey questions to survey bank
vi) Present result information about users, sessions, and surveys
vii) Import/export results

At the user lever, it is possible for users to create an account to sign in and also to continue the test sessions if already started and signed in.

5 Pilot study

In order to evaluate our patient numeracy assessment method we performed an empirical investigation and conducted a controlled experiment. We adapted the six level process model described in [7] for this investigation. The objective of our controlled experiment was to determine, how differences in the numeracy skill assessment method could affect the result of the assessment (conception level).

The results of the formal empirical study demonstrated that our confidence-based numeracy assessment method excels the non-confidence assessment method in terms of objective and subjective characteristics.

6 Related work

There have been several methods for numeracy assessment in the literature. Some are considered as more established standard pioneers and some are developed in the more recent years.

In [10], Schwartz et al. assessed patients' numeracy with three questions and scored it as the total number of correct responses. In [11], Lipkus et al. evaluated a set of eleven questions that compose more questions that directly evalute the patients' ability of risk understanding.

Rapid Estimate of Adult Literacy in Medicine (REALM) [12] measures the individual's ability to read common

medical words and lay terms for parts of body and illness. WRAT-3 (Wide Range Achievement Test), [13] assesses basic skills in reading, arithmetic, and spelling. The test takes approximately 30 minutes to administer.

Test of Functional Health Literacy in Adults (TOFHLA) is designed in two parts: 17-item numeracy questions and 50-item reading comprehension questions with three passages. It uses actual health-related materials such as prescription bottle labels and appointment slips. S-TOFHLA, a shortened version of TOFHLA, consists of 4 numeracy questions and 36 reading comprehension with two passages. It needs half a time for administration compared with TOFHLA.

Medical Data Interpretation Test (MDIT) [14] assesses the individual's ability to interpret and understand medical statistics and understand concepts regarding risk. The test includes 18 questions based on the individual's daily encounter with health information. The Newest Vital Sign [15] is another functional test it consists only six questions based on the nutrition label states.

Numeracy Understanding in Medicine Instrument (NUMi) [5] is based on using item response theory scaling methods. The test has 20 items with an item bank calibrated with 1000 users.

The existing numeracy methods have some limitations [16]: first, none includes full set of skills and knowledge associated with numeracy. Second, potential confounders such as test anxiety, and distress are not taken into account and at last high-end means of communication and technology are not considered in the assessment.

To obtain reliable measurement, specific health education interventions should be individually tailored for patients [17] and the numeracy level of patients should be assessed.

7 Conclusions

We designed and conducted an empirical study that proves that our CBT method produces results as accurate and productive as standard numeracy assessment methods; however, our method is more effective and has higher Satisfaction, Trust and Discretionary Usage compared to NCBT methods. We developed a web-based/portal application to assess numeracy level, which includes a database system and withholds information about the patients and results of the surveys.

Our future work includes testing the method with different categories of patients within the hospital-ward domain.

References

[1] M. Omidbaksh, and T. Radhakrishnan,"An adaptive testing model for assessment of patient numeracy," in CBMS, NY, 2014.

[2] M. Omidbaksh, and O. Ormandjieva, "Goal-driven modeling for confidence-based patient numeracy assessment: C-PNA", EUSPN/ICTH 2015 pp.213-220, 2015.

[3] ISO/IEC DIS 25022. System and Software Engineering-Systems and Software Quality, 2015.

[4] I. M. Lipkus, G. Samsa and B. K. Rimer, "General performance on a numeracy scale among highly educated samples," Medical Decision Making, vol. 21, pp. 37, 2001

[5] M. M. Schapira, C. M.Walker, K. J. Capparet, P. S. Ganschow, K. E. Fletcher, E.L. McGinley, S. Del Pozo , C. Schaure . S. Tarima, and E. A. Jacobs, "The numeracy understanding in medicine instrument: a measure of health numeracy developed using item response theory", Med Decis Making. 2012 Nov-Dec;32(6):851-65. doi: 10.1177/0272989X12447239. Epub 2012 May 25

[6] http://www.assessnumeracy.com, 2015, December.

[7] N. Fenton, J. Bieman, Software metrics: A rigoorous and practical approach, third edition, 2014.

[8] N. Bettenburg, R. Premraj, T. Zimmermann, and S. Kim, "Extracting structural information from bug reports", In MSR '08: Proceedings of the 2008 international working conference on Mining software repositories, pp. 27–30, 2008.

[9] C. Wohlin, P. Runeson, M. HöstC. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, "Experimentation in Software Engineering: An Introduction", Norwell, MA, USA: Kluwer Academic Publishers, 2000.

[10] S. Woloshin, L. M. Schwartz, S. Byram, B. Fischhoff and H. G. Welch, "A new scale for assessing perceptions of chance: a validation study," Medical *Decision Making,* vol. 20, pp. 298, 2000

[11] I. M. Lipkus, G. Samsa and B. K. Rimer, "General performance on a numeracy scale among highly educated samples," *Medical Decision Making,* vol. 21, pp. 37, 2001.

[12] T.C. Davis, M.A. Crouch, S.W. Long, R.H. Jackson, P. Bates, R.B. George, and L.E. Bairnsfather, "Rapid assessment of *literacy* levels of adult primary care

patients", journal of Family Medicine, vol.23(6), pp.433-440,1991.

[13] S. Weintraub, "Neuropsychological assessment of mental state", Principles of behavioural and cognitive neurology, pp.121-173, 2000.

[14] S. Woloshin, "Can Patients Interpret Health Information? An Assessment of the Medical Data Interpretation Test", *Med Decis Making,* 25, pp. 290, 2005.

[15] B.D. Weiss, M.Z. Mays, W. Martz, K.M. Castro, D.A. DeWalt, M.P. Pignone, J. Mockbee, and F.A. Hale, "Quick assessment of literacy in primary care: the newest vital sign", Annals of Family Medicine, vol.3 (6), pp.514-518, 2005.

[16] R. L. Rothman, M. M.Huizigna, A. J. Carlsile, Kerri Cavaaugh , D., Davis, and R. Gregory, "Literacy, numeracy and portion-size estimation skills", AJPM, 2009.

[17] T. Davis, M. William, E. Marin, R. M. Parker, and J. Glass, "Health Literacy and cancer communication",CA,Vol.52,(3), pp.134-149, 2002.

Dependability Evaluation of WBAN-based Home Healthcare Systems using Stochastic Activity Networks

Manuel Ríos
Grupo SISTEMIC
Depto. de Ingeniería Electrónica y Telecomunicaciones
Facultad de Ingeniería
Universidad de Antioquia UdeA, Calle 70 No 52-21
Medellín, Colombia
manue.rios@udea.edu.co

Fredy Rivera
Grupo SISTEMIC
Depto. de Ingeniería de Sistemas
Facultad de Ingeniería
Universidad de Antioquia UdeA, Calle 70 No 52-21
Medellín, Colombia
falexander.rivera@udea.edu.co

***Abstract*—Wireless Body Area Networks (WBANs) have emerged as a great technological alternative for medical applications that require continuous monitoring of vital signs in patients with diseases of low progression and long duration. One of the most challenging problems in this kind of applications is related to preserve dependability attributes (reliability and availability), since a failure in the system might cause false alarms, alterations in medical diagnosis and, in a worst case, dead of patient. The absence of methodologies for evaluating and predicting correct system operation during design stage is common when developing this kind of devices. In this paper a Stochastic Activity Network (SAN) methodology is proposed, in order to assess dependability of a wireless equipment aimed to support vital signs monitoring for patients who meet the conditions for home healthcare. In SAN models the failures associated to hardware and wireless protocol communication were considered, for the purpose of estimating availability and reliability for typical operational scenarios. The results obtained allow system designers to identify dependability bottlenecks on which concentrate to reduce risks and threats to this fundamental system property.**

Keywords: *WBAN, Dependability, Reliability, Availability and Stochastic Activity Networks*

I. Introduction

The particular characteristics of Wireless Body Area Networks (WBANs) have consolidated this technology as a great alternative for improving life conditions of chronic disease patients, by providing support for continuous and remote monitoring of physiological variables without limiting their normal activities. WBANs are based on small, low power sensors located in or on the human body, that have the capability to collect, store, process, and transmit medical information that can be used to trigger first aid assistance, and to detect emergency situations [1]. Despite these advantages, there are still several system attributes that represent a challenge in WBANs, namely those related to dependability. Issues such as node failure, body shadowing, environmental interference, fault network propagation, quality of hardware and software design [2], need to be further developed in order to guarantee dependable nodes for medical applications. According to Avizienis et al. [3], reliability, safety, integrity, availability and maintainability are five attributes that gathered together describe dependability, defined as the ability to avoid more frequent and more severe system failures. All attributes associated with dependability are important in a general evaluation of any system. However, this research was focused on reliability and availability modeling and assessment, considering the requirements of medical devices used in home healthcare scenarios, where it is necessary to have trustworthy and reliable equipment, i.e. capable to finish the task entrusted, without experiencing a failure (reliability), and available to provide a correct service when it is required (availability). Despite the importance of evaluating this attributes during design stage, there are few papers that deal with the evaluation of dependability of WBAN related to sensor nodes hardware, communication protocol, and the effectiveness of use of the techniques proposed to overcome the threats to availability and reliability.

Designing dependable systems is not an easy task, partly due to the great difficulty to evaluate the effectiveness of different design options, taking into account the complexity and high cost of implementing prototypes to evaluate reliability and availability with high accuracy. Several methods like *fault rate models*, *reliability graphs*, *fault trees*, *Markov chains* and *Petri Nets* have been used to evaluate dependability of computing and communication systems [4]. However, they suffer from the fact that the system to be modeled must be described at the state level, and the number of states that must be considered can be huge [5]. Stochastic Activity Networks (SANs) are a modeling formalism widely used to assess performance, dependability and performability in complex computation systems, with an acceptable accuracy [6]. SANs are Petri nets extensions based on directed graphs, that provide the necessary tools to specify high-level atomic

models, reducing the complexity to specify thousand of states and their interaction, using graphics primitives. Taking advantage of the benefits of SANs, in this research they were used to assess the attributes of reliability and availability of a wireless sensor nodes aimed to support physiological signs monitoring in home healthcare scenarios.

The rest of the article is divided as follows: section II presents an overview of related work on identification of reliability and availability threats in WBANs, as well as the techniques commonly used to evaluate these dependability attributes. Section III includes a detailed description of the wireless sensor node architecture. Section IV contains the description of the SAN model of the sensor node. Section V shows the experimental results of the reliability and availability evaluation. Finally, section VI presents conclusions and future research.

II. Related Works

Threats to dependability are a major challenge in WBANs for medical scenarios, considering the characteristics of this kind of applications, that collect, store, and transmit patient's vital and extremely confidential information. In this way, the evaluation of reliability and availability are fundamental non-functional requirements to be considered during design stage in order to guarantee the effective insertion of WBAN in home healthcare monitoring scenarios. In this sense, the authors in [2] have identified the potential causes of failure on a WBAN. Issues like unreliable hardware, limited or power loss, environmental interference, and network failure are addressed. They propose several schemes aimed to increase the dependability (related to reliability, availability and security attributes), being the inclusion of redundant nodes the recommended solution. However, this option implies an increase in cost while reducing the ergonomics and mobility of the user. In the same work, reliability assessment in the design stage is not addressed by the authors. In [7], the authors propose a methodology for reliability assessment on a wireless sensor node, based on an automatic generation of fault trees, when permanent faults occur on network devices. This proposal supports any topology, different levels of redundancy, network reconfigurations, criticality of devices, and arbitrary failure conditions, allowing to optimize in design stage parameters that could affect reliability and availability requirements. It was tested in typical industrial scenarios, and the results obtained show that is possible to identify dependability bottlenecks. Another approach to evaluate reliability is reported in [8], where the authors besides identifying threats to reliability and availability of Wireless Sensor Networks (WSNs), also propose a flexible framework for dependability evaluation and analysis using Stochastic Activity Networks (SAN). The classification of potential hazards and risks is performed using a Failure Mode and Effect Analysis (FMEA) for a WBAN in order to obtain the failure model of a single sensor node. It is noteworthy the use of external libraries in the SAN model, that allows to modify the failure behaviour at execution time according to network dynamics. However, the results lack of schemes or means that increase the values of reliability and availability. SAN formalism has been used to evaluate reliability and availability in different areas. The authors in [9] evaluate the failure probability of systems implemented on SRAM-FPGA technology. The results obtained with the SAN model allow to predict the probability of failures according to the maximum number of faults that can be injected, and the input signal probability of data signals. SAN modeling was also used to estimate the reliability of a Controller Area Network (CAN)-based system [10], whose results allow to quantify the reliability achievable by highly-reliable CAN-based systems that rely on a replicated bus topology. The authors were able to compare the reliability attainable by a replicated CAN bus with the one what would be achieved by a simplex CAN bus, and a replicated CAN star. In [11] the authors assess the operational reliability of an aircraft before and during a fly mission, using a SAN constructed from a meta-model that describes, at a high level, the behaviour of the mission. The SAN model considers real operation conditions after component failures in order to support aircraft maintenance planning. The results provide the schedule of repairs taking into account some optimization criteria: cost, remaining useful life and operational risks, according to reliability values after several missions.

III. WBAN Node Description

The system called *HealTICa* is composed of a couple of wireless sensor nodes located on the patient's body, in order to acquire physiological variables of interest such as heart beat rate, oxygen saturation, body temperature, blood sugar levels, as well as the electrocardiographic (ECG) signal. These variables are wirelessly transmitted to a sink node, in this case a smartphone, that features a dongle that integrates wireless transceiver and serial communication circuits, also allowing to store and visualize medical variables, while providing the functionality of triggering alerts in emergency situations, originated by algorithms that analyse medical information, or requested by the user using the panic button provided in the mobile application. Figure 1 shows the system architecture, highlighting the *HealTICa node* (ECG-PPG sensor node), responsible of acquiring, processing and transmitting the referred variables, except the blood sugar level, which is collected by another wireless glucometer. The HealTICa node is formed by an analog front-end for ECG signal acquisition, an infrared (IR) sensor for measuring the patient's body temperature, and a pulse oximeter sensor that calculates heart beat rate and oxygen saturation (SpO_2) values from a photoplethysmogram (PPG) sensor. All sensors are orchestrated by a low energy consumption microcontroller, responsible of executing data acquisition and processing algorithms, as well as communicating to the transceiver, that is configured according to the standard for low-rate wireless Personal Area Networks (PAN), IEEE 802.15.4 with beacon enabled. The PAN is formed by a coordinator, in this case a smartphone (equipped with an IEEE

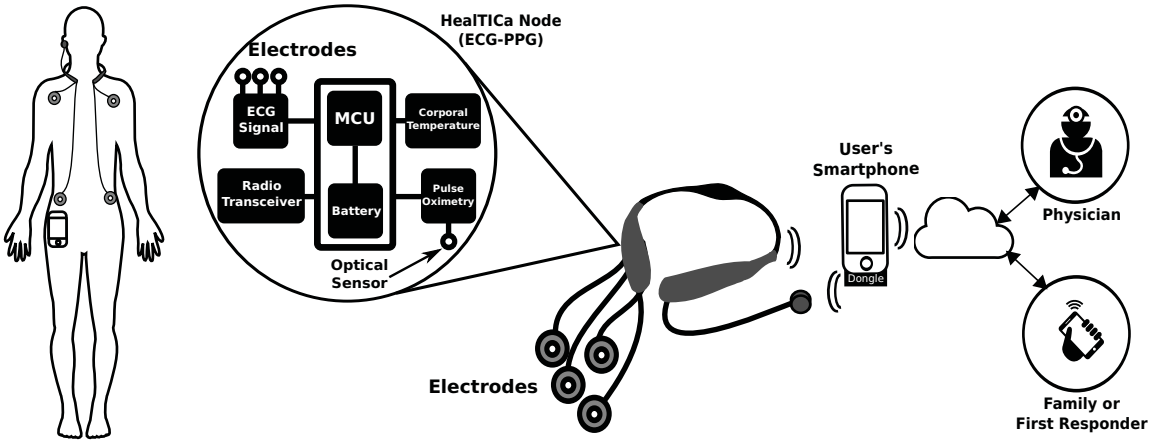


Fig. 1. Architecture and hardware block diagram of the HealTICa system

802.15.4 dongle), that is in charge of managing the whole network. The IEEE 802.15.4 MAC protocol is used in beacon enabled mode for saving energy [12]. The CSMA/CA (Carrier Sense Multiple Access with Collision Avoidance) algorithm is used in this protocol for channel access. Reliability of WBAN is affected by issues associated to this algorithm as reported in [13], [14] and [15]. A common solution to these issues consists in adjusting the standard parameters related to the maximum number of retransmissions (*macMaxFrameRetries*), and the number of backoff stages (*macMaxCSMABackoffs*) as it is intended in this work.

IV. Reliability and Availability Evaluation

In order to evaluate the reliability and availability of the HealTICa system, the Stochastic Activity Network modeling formalism is used. SAN is a probabilistic extension of Petri nets, and together with reduced base model construction techniques, has the potential to avoid the state space explosion for dependability evaluation of parallel and distributed complex systems [6]. The HealTICa system SAN model was created using the framework Möbius [16], which is an extensible dependability, security, and performance modeling environment for large-scale discrete-event systems. It provides multiple modeling formalisms and solution techniques, facilitating the representation of each part of a system, and providing different solution methods that allow the estimation of the system behaviour. Möbius offers the typical SAN graphical elements that allow to create high level models that are close to the actual stochastic behaviour of this kind of systems with a reduced number of states compared to those ones requiered for an analytic solution. SAN models include input gates, output gates, places, activities and arcs. Places, graphically represented by circles, can be seen as a state of the modeled system. Each place of a SAN contains a certain number of tokens, which represent the marking of that place.

Figure 2 shows the atomic SAN model of the HealTICa node hardware. In this model each place represents the status of hardware components. Places related to sensors are

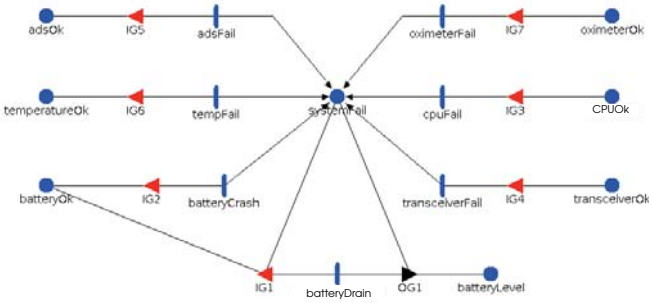


Fig. 2. Atomic SAN model for the HealTICa node hardware

adsOk (ECG signal sensor), *temperatureOk* (corporal temperature sensor), and *oximeterOk* (pulse oximeter sensor). There are also places related to the status of the microcontroller (*CPUOk*), communication transceiver (*transceiverOk*), and battery (*batteryOk*). If any of these places have only one token, this means that there have been no failures in those components. Evolution in time of the SAN model is governed by activities, input gates (IG) and output gates (OG). Activities (transitions in Petri net terminology) represent actions of the modeled system that could take some specified amount of time to complete. There are two types: timed and instantaneous activities. Timed activities have durations that affect the performance of the modeled system. Instantaneous activities represent actions that complete immediately when enabled in the system. In the HealTICa system SAN model presented in Figure 2, there are seven activities, graphically represented by thick vertical lines, associated with failures in each component of the node. Activation time follows an exponential distribution, which is associated with failure rates of hardware components reported by manufacturers. The *adsFail* activity represents the time between fails in the ECG sensor following an exponential distribution. This likewise occurs with the remaining *transceiverFail* (communication transceiver), *tempFail* (corporal temperature sensor),

TABLE I
ENABLING PREDICATES IN THE HEALTICA NODE HARDWARE SAN MODEL

Input gate	Enabling predicate	Input function
IG1	(systemFail->Mark()==0)&& (batteryOk->Mark()==1)	;
IG2	batteryOk->Mark()==1	batteryOk->Mark()=0;
IG3	(*CPUOk* ->Mark()==1) && (systemFail->Mark()==0)	*CPUOk* ->Mark()=0;
IG4	(transceiverOk->Mark()==1)&& (systemFail->Mark()==0)	transceiverOk->Mark()=0;
IG5	(adsOk->Mark()==1)&& (systemFail->Mark()==0)	adsOk->Mark()=0;
IG6	(temperatureOk->Mark()==1)&& (systemFail->Mark()==0)	temperatureOk->Mark()=0;
IG7	(oximeterOk->Mark()==1)&& (systemFail->Mark()==0)	oximeterOk->Mark()=0;

batteryCrash (Li-Ion battery), *oximeterFail* (pulse oximeter sensor), and *cpuFail* (microcontroller) activities. Activation of activities depends on boolean enabling predicates of input gates. An input gate is graphically represented in a SAN model by a triangle. If an input gate enabling predicate is true, its associated activity is activated, initiating its corresponding time distribution function. At the time an activity is completed, the input function of the corresponding input gate updates the marking of the network. In the HealTICa node SAN mode presented in Figure 2, when an activity is completed, one token is located in the *systemFail* place. If there is a token in this place, it represents a general fail of the system. The *systemFail* place is considered in the reward function for reliability and availability assessment. The marking of the network can be also updated by the output function of an output gate. Output gates, as input gates, are also graphically represented by triangles in a SAN model. When an activity terminates, the output function of the corresponding output gate updates the marking of the network. Table I presents the enabling predicates and input functions for each input gate of the HealTICa node SAN model illustrated in Figure 2. Meanwhile, Table II presents the output function for the output gate *OG1*. In both tables, the *Mark()* function returns the number of tokens of the corresponding place. The output function in *OG1* changes the marking of the *batteryLevel* place, reducing the number of tokens, according to the energy consumption per hour (*batteryDrain* activity) of the HealTICa node represented by the variable *TotalBatteryDrain*. When the level is less than 50 *mAh* (threshold from where there is a high probability of electronic components failure) a token is located in the *systemFail* place . The output function in Table II is compared with the marking in the *batteryLevel* place. If *batteryLevel* place is less than 50 *mAh*, the *systemFail* place receives a token indicating a failure in the system. This also occurs if a token is located in the *batteryCrash* place, that represents a critical battery failure.

The atomic model of the HealTICa system dedicated to

TABLE II
OUTPUT GATE IN THE HEALTICA NODE HARDWARE SAN MODEL

Gate	Output function

OG1	batteryLevel->Mark()=batteryLevel->Mark()-totalBatteryDrain;
	if (batteryLevel->Mark()<50){
	systemFail->Mark()=1;
	}

the wireless protocol communication is based on the standard IEEE 802.15.4 configured with the default parameters for the number of *backoff* stages (*MaxCSMABackoffs* = 4), and the maximum number of retransmissions when the ACK is not successfully received (*MaxFrameRetries* = 3), as it is modeled by the output function of the output gate *OG3*, and shown in Table IV. This model considers the CSMA/CA algorithm related to frame drops due to congestion. If the channel is busy, represented by the *chBusy* place in Figure 3, the Contention Window (*CW*) (number of *backoff* periods during which the channel must be sensed idle before accessing the channel) is re-initialized to $CW = 2$ i.e. the Clear Channel Assessment (CCA) is made twice, and the Number of Backoffs (*NB*) that represents the number of times the CSMA/CA algorithm was required to *backoff* while attempting to access the channel is incremented. If the maximum number of *backoffs* (*macMaxCSMABackoffs* = 4) is reached, the algorithm reports a failure, and a token is located in the *discardPacket* place, as it is shown in the *OG4* output function in Table IV. Otherwise, it goes back to the CCA. If the channel is sensed as idle, *CW* is decremented. The CCA is repeated if $CW \neq 0$. This ensures performing two CCA operations to prevent potential collisions of acknowledgement frames. If the channel is again sensed as idle, the node attempts to transmit until the number of *MaxFrameRetries* = 3 is reached, when the packet is dropped. The atomic model of the HealTICa system wireless communication protocol presented in Figure 3 considers a transmission rate represented by the *txPacket* activity. This activity fires if the *idleRadio* and *transceiverOk* places have at least one token. When a failure in the transceiver occurs, the token in the *transceiverOk* place is removed. The *waitACK* and *chBusy* places receive a token according to the probability of transition of the *txPacket*. Meanwhile, *txStatus* and *chStatus* activities correspond to a delay associated to the ACK before the process of retransmission, and the period of *backoff* in the CSMA/CA algorithm, respectively [17]. These cases in the activities represent the probability of failure in the Clear Channel Assessment (CCA), and the probability of successfull ACK reception. It is considered that a failure in the system occurs if the number of tokens in *discardPacket* place is greater than 40 data frames per hour. In this case, a token is located in the *systemFail* place. This value represents the maximum number of tolerable discarded packets in order to successfully reconstruct the ECG signal by the application installed in the smartphone. The output functions of output gates *OG3* and *OG4*, described in Table IV, check that the number of retransmisions *MaxFrameRetries* (number of tokens

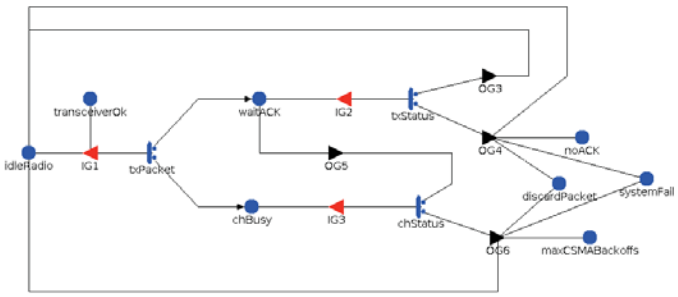


Fig. 3. Atomic SAN model for the HealTICa node wireless communication protocol

TABLE III
INPUT GATES DESCRIPTION FOR THE HEALTICA NODE WIRELESS COMMUNICATION SAN MODEL

Input gate	Enabling predicate	Input function
IG1	(transceiverOk->Mark()==1)&& (idleRadio->Mark()==1)&& (systemFail->Mark()==0)	idleRadio->Mark()=0;
IG2	(waitACK->Mark()==1)&& (systemFail->Mark()==0)	waitACK->Mark()=0;
IG3	(chBusy->Mark()==1)&& (systemFail->Mark()==0)	chBusy->Mark()=0;

in the *noACK* place) is less than 3, and the number of *macMaxCSMABackoffs*, that are represented by the number of tokens in the *macMaxCSMABackoffs* place, is less than 4. The *noACK* and *macMAxBackoff* places are incremented according to the evolution in time of the model. If the marking reaches the limit, the output gates *OG3* and *OG4* increment the number of tokens in the *discardPacket* place.

Atomic models can be replicated and joined together to form a complete, or composed, model, allowing to share global variables. The HealTICa system SAN composed model is formed by the atomic models presented in Figure 2 (sensor node hardware) and Figure 3 (wireless communication protocol), and it is not illustrated here for the sake of space.

TABLE IV
OUTPUT FUNCTIONS IN THE HEALTICA NODE WIRELESS COMMUNICATION SAN MODEL

Gate	Output function
OG1	waitACK->Mark()=1; maxCSMABackoffs->Mark()=0;
OG2	idleRadio->Mark()=1; noACK->Mark()=0;

OG3	noACK->Mark()++;
	waitACK->Mark()=1;
	if(noACK->Mark()>3){
	discardPacket->Mark()++;
	}

OG4	chBusy->Mark()=1;
	maxCSMABackoffs->Mark()++;
	if(maxCSMABackoffs->Mark()>4){
	discardPacket->Mark()++;
	}

V. EXPERIMENTAL SETUP AND RESULTS

Different scenarios were considered in order to test the HealTICa system SAN model, representing typical situations under which it can operate. These experiments aim to identify the impact of different factors on the HealTICa system reliability and availability attributes, according to each scenario. The HealTICa system SAN model was depicted and simulated to estimate its reliability and availability using the Möbius software tool [16], running on a Laptop with an Intel Core i5-3337U processor (1,80GHZ, 3MB cache), 4 GB RAM memory, and Ubuntu Linux 14.04 as operating system. Table V shows the constant global variables of the HealTICa system SAN composed model. ECG, oximeter and corporal temperature sensors failure rates are taken from datasheets published by manufacturers. Data trasmission rate (*txRate*) was established in 900 frames per hour, according to the number of data frames required to reconstruct and analyze the ECG signal.

TABLE V
CONSTANT GLOBAL VARIABLES IN THE HEALTICA SYSTEM SAN MODEL

Variable name	Value	Description
adsFailRate	$2{,}6\times10^{-10}$	ECG sensor failure rate
oximeterFailRate	$3{,}4\times10^{-6}$	Oximeter sensor failure rate
tempFailRate	$1{,}8\times10^{-6}$	Corporal temperature sensor failure rate
txRate	900	Data transmission rate (frames per hour)

The experimental setup and results for five different scenarios intended for HealTICa system SAN model reliability evaluation are presented as follows:

- *relCommTypical scenario*: This scenario is composed by typical values of congestion in the 2,4GHz band, in which IEEE 802.15.4 works. This band is shared with Wi-Fi, bluetooth, cordless phones and many other devices, making the sensor node susceptible to interferences generated by the mentioned technologies. This scenario recreates a current home environment where the HealTICa system can be deployed. The values of the variables associated with ACK reception (*receiveACKProb*) and channel congestion (*chBusyProb*) were consequently configured as shown in Table VI. Figure 4 shows an exponential decreasing of the reliability, associated to the loss of data frames. Although the impact is negligible for less than 100 hours, it is important to take into account these values in order to schedule preventive maintenance when the reliability values are close to 95%, that is after 900 hours of operation. For this scenario, the computation time for solving the HealTIC system SAN model was 1064 seconds, with a SAN state space of 7564 states.

- *relCommNoisy scenario*: The configuration of this experiment is similar to the previous one, adding

high interference, which is represented by a high probability of packet loss and low probability of channel access. Hence, the values of variables associated to congestion (*chBusyProb*) and packet loss probability (*receiveACKProb*) are shown in Table VI. This behaviour was experimentally verified when the nodes operate in environments with more than ten Wi-Fi hotspots. It is observed in Figure 4 that the drop on reliability associated to noisy environments is pronounced, reaching 90% of reliability after 1000 hours of system operation. These results suggest the implementation of techniques for reducing the packet loss such as changing the default parameters *MaxFrameRetries* and *macMaxCSMABackoffs*. For this scenario, computation time for solving the HealTICa system SAN model was 321 seconds, with a SAN state space of 182796 states.

- *relMCULow* and *RelMCUTypical scenarios*: Third and fourth scenarios contemplate different failure rates for commercial micro-controllers (*mcuFailRate*), as shown in Table VI. MCU is a key component of the sensor node, because its failure represents a severe system error. Manufacturers of MCUs offer a wide variety of devices according to the application requirements. Typical values of general purpose microcontrollers provided in manufacturers datasheets were used in these scenarios. From the experimental results is possible to identify MCU failure as the main issue that concerns reliability. Figure 4 shows a pronounced decrease in reliability after 500 hours of system operation when using MCU with short Mean Time Before Failure (MTBF) values (*relMCULow* scenario). This means six months of usage, with the system working eight hours a day. On the other hand, MCUs with large MTBF values (*relMCUTypical* scenario) yields to a very important reliability improvement, reaching 93% of reliability after 1000 hours of operation, as can be seen in Figure 4. In the *relMCULow* scenario, the computation time for solving the HealTICa system SAN model was 1055 seconds, while for *RelMCUTypical* scenario was 1108 seconds. For both scenarios the number of SAN states was 7564.

- *relBatteryTypical scenario*: This scenario consists of varying the value of reliability for commercial batteries, usually indicated in hours of operation. According to [18], batteries of Li-Ion have an average cycle of life of 40000 hours. The graph associated to this experiment in Figure 4 presents a smooth decline in reliability, reaching values greater than 92% after 1000 hours of system operation. In this sense it is important to consider battery preventive maintenance for life cycles near to 1000 hours. For purposes of evaluating a most severe scenario, a battery with a life cycle of 20000 hours were considered. The results are not shown in Figure 4 because its time response is very close to the scenario with a low reliable MCU (*relMCULow*), i.e. reliability values are lower than 90% after 850 hours of operation. Hence, reliable batteries are a major requirement to take into account in this kind of applications. The computation time of the HealTICa system SAN model for this scenario was 1059 seconds, with 7564 SAN states.

TABLE VI
VALUES OF EXPERIMENTAL SETUP FOR RELIABILITY EVALUATION IN DIFFERENT SCENARIOS

Experimental scenario	Variable	Value
relCommTypical	receiveACKProb	0,95
	chBusyProb	0,05
relCommNoisy	receiveACKProb	0,85
	chBusyProb	0,15
relMCUTypical	mcuFailRate	1×10^{-8}
relMCULow	mcuFailRate	1×10^{-6}
relBatteryTypical	batteryFailRate	1/40000

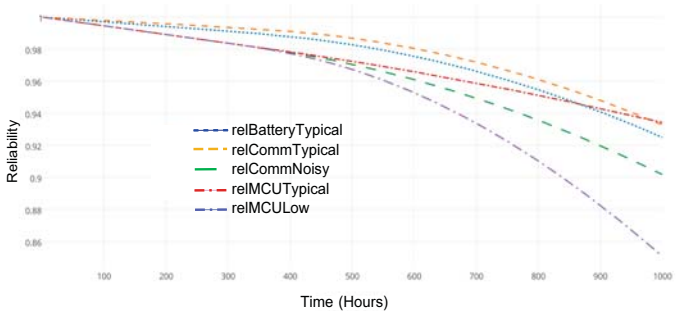


Fig. 4. HealTICa reliability node for different scenarios

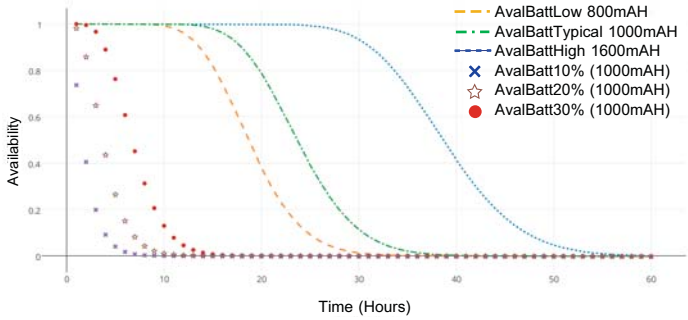


Fig. 5. HealTICa node availability for different battery capacities and percentages of charge

Finally, two sets of scenarios were created to evaluate the HealTICa system SAN model availability. By means of these experiments it is possible to estimate the system autonomy, and the minimum percentage of battery charge required to guarantee two hours of continuous operation, which is a medical

requirement for the system. In the first set, for exploring system autonomy, three different values of charge for commercial batteries were considered: 800 *mAh* (*availBattLow*), 1000 *mAh* (*availBattTypical*), and 1600 *mAh* (*availBattHigh*). According to Figure 5, although the battery with larger charge produces longer availability, it is necessary to consider its cost and size that could affect system viability and ergonomics. In the second set, for determining the minimum battery charge required to guarantee two hours of continuous operation, a 1000 *mAh* battery was considered. Three experiments were performed with charges of 10% (*avalBatt10%*), 20% (*avalBatt20%*), and 30% (*avalBatt30%*) of full capacity. Figure 5 shows that is necessary to have at least 30% of full capacity when using a 1000 mAh battery to guarantee two hours of continuous operation. In these scenarios the average computation time for solving the HealTICa system SAN model was 750 seconds, and the number of SAN states was 54208.

VI. Conclusions and Future Work

In this study, the availability and reliability of a wireless sensor node for medical applications called HealTICa was evaluated using the formalism named Stochastic Activity Networks. The results obtained from high level atomic models provide information about hardware quality, power supply and environment conditions necessary to ensure proper system operation, evaluated in terms of system dependability, specifically in its availability and reliability attributes. The direct relationship between the battery charge and the availability of the system was verified under different operational scenarios. During the design stage it is necessary to consider not only the battery capacity in *mAh* but also the battery reliability in life cycles to guarantee system availability. Low reliability microcontrollers turn themselves as a system reliability bottleneck. In terms of communication protocol, it was evidenced the interference issue in the IEEE 802.15.4 2,4 GHz band because it is shared with countless devices in typical operational scenarios. WBANs represent a serious alternative for developing home healthcare devices. This work presented an approach to evaluate the dependability (in term of its availability and reliability attributes) of a WBAN-based home healtcare system during the design stage, in order to early detect risks and threats in different operational scenarios.

For future work is being considered to design a reliable hardware and software architecture for home healthcare applications in order to have highly dependable systems, adding more attributes as maintainability, safety, confidentiality and integrity. This architecture will be modeled with the SAN formalism in order to avoid space state explosion, and to evaluate its effectiveness prior to system prototyping.

VII. Acknowledgment

This work has been supported by the project *Plataforma tecnológica para los servicios de teleasistencia, emergencias médicas, seguimiento y monitoreo permanente a los pacientes y apoyo a los programas de promoción y prevención* of the *Centro de Excelencia ARTICA (Alianza Regional en TICs Aplicadas)* funded by the *Sistema General de Regalías* (Colombia).

References

[1] J. Elias, A. Jarray, J. Salazar, A. Karmouch, and A. Mehaoua, "A reliable design of wireless body area networks," in *Global Communications Conference (GLOBECOM), 2013 IEEE*, 2013, pp. 2742–2748.

[2] Y. Hovakeemian, K. Naik, and A. Nayak, "A survey on dependability in body area networks," in *Medical Information & Communication Technology (ISMICT), 2011 5th International Symposium on*, 2011, pp. 10–14.

[3] A. Avižienis, J.-C. Laprie, B. Randell, and C. Landwehr, "Basic concepts and taxonomy of dependable and secure computing," *Dependable and Secure Computing, IEEE Transactions on*, vol. 1, no. 1, pp. 11–33, 2004.

[4] A. M. Johnson Jr and M. Malek, "Survey of software tools for evaluating reliability, availability, and serviceability," *ACM Computing Surveys (CSUR)*, vol. 20, no. 4, pp. 227–269, 1988.

[5] W. H. Sanders and L. M. Malhis, "Dependability evaluation using composed SAN-based reward models," *Journal of parallel and distributed computing*, vol. 15, no. 3, pp. 238–254, 1992.

[6] W. H. Sanders and J. F. Meyer, "Stochastic activity networks: Formal definitions and concepts," in *Lectures on Formal Methods and PerformanceAnalysis*, 2001, pp. 315–343.

[7] I. Silva, L. A. Guedes, P. Portugal, and F. Vasques, "Reliability and availability evaluation of wireless sensor networks for industrial applications," *Sensors*, vol. 12, no. 1, pp. 806–838, 2012.

[8] M. Cinque, D. Cotroneo, C. Di Martinio, and S. Russo, "Modeling and assessing the dependability of wireless sensor networks," in *Reliable Distributed Systems, 2007. SRDS 2007. 26th IEEE International Symposium on*, 2007, pp. 33–44.

[9] C. Bernardeschi, L. Cassano, and A. Domenici, "Failure probability of SRAM-FPGA systems with stochastic activity networks," in *Design and Diagnostics of Electronic Circuits & Systems (DDECS), 2011 IEEE 14th International Symposium on*, 2011, pp. 293–296.

[10] M. Barranco, F. Pozo, and J. Proenza, "A model for quantifying the reliability of highly-reliable distributed systems based on fieldbus replicated buses," in *Emerging Technology and Factory Automation (ETFA), 2014 IEEE*, 2014, pp. 1–8.

[11] K. Tiassou, K. Kanoun, M. Kaâniche, C. Seguin, and C. Papadopoulos, "Aircraft operational reliability—a model-based approach and a case study," *Reliability Engineering & System Safety*, vol. 120, pp. 163–176, 2013.

[12] R. Wavage and A. Kaushik, "Performance analysis of beacon enabled IEEE 802.15. 4 using GTS in Zigbee," *International Journal of Computer Science & Applications (TIJCSA)*, vol. 2, no. 12, 2014.

[13] G. Anastasi, M. Conti, and M. Di Francesco, "The MAC unreliability problem in IEEE 802.15.4 wireless sensor networks," in *Proceedings of the 12th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems*, 2009, pp. 196–203.

[14] S. Wijetunge, U. Gunawardana, and R. Liyanapathirana, "Data transmission reliability of IEEE 802.15. 4 based wireless sensor networks with synchronised periodic data," in *Computer & Information Science (ICCIS), 2012 International Conference on*, vol. 2, 2012, pp. 619–624.

[15] D. Gomes and J. A. Afonso, "Improving the communication reliability of body sensor networks based on the IEEE 802.15.4 protocol," *Telemedicine and e-Health*, vol. 20, no. 3, pp. 261–268, 2014.

[16] T. Courtney, S. Gaonkar, K. Keefe, E. W. Rozier, and W. H. Sanders, "Möbius 2.3: An extensible tool for dependability, security, and performance evaluation of large and complex system models," in *Dependable Systems & Networks, 2009. DSN'09. IEEE/IFIP International Conference on*, 2009, pp. 353–358.

[17] A. Koubaa, M. Alves, E. Tovar, and Y.-Q. Song, "On the performance limits of slotted CSMA/CA in IEEE 802.15.4 for broadcast transmissions in wireless sensor networks," 2006.

[18] S.-W. Eom, M.-K. Kim, I.-J. Kim, S.-I. Moon, Y.-K. Sun, and H.-S. Kim, "Life prediction and reliability assessment of lithium secondary batteries," *Journal of Power Sources*, vol. 174, no. 2, pp. 954–958, 2007.

Regulation-driven Verification of Vein-to-vein Blood Transfer Safety

[1]Noha Hazzazi, [1]Duminda Wijesekera, and , [2]Jasem Albasri
[1]Department of Computer Science, George Mason University, Fairfax, VA, 22030.
[2]Centeral Military Laboratory & Blood Bank, PSMMC, Riyadh, KSA.

Abstract— *Minimizing Transfusion risks is being researched by hematologist and blood bank professionals for many years. The Food and Drug Administration (FDA), American Association of Blood Banks (AABB) and other set safety standards and update them regularly to achieve the same objective in the presence of newer pathogens and advances in transfusion technology. We have formally modeled the vein-to-vein blood transfusion supply chain as an executable workflow, translated the FDA and AABB requirements as Temporal Logic formulas and verified them against our formal model. We show this in by explaining the main theorem and lemmas in our paper, proving that the vein-to-vein transfusion safety hold in modeled processes.*

Keywords: Blood safety, FDA, AABB

1. Introduction

Blood transfusion is a common procedure used all over the world. Due to safety related issues such as blood quality, contamination, aging, etc. United States and other countries have placed regulations and standards in place to ensure that the blood delivered for transfusion is safe for donors and recipients. Between donation and transfusion, blood products go through many steps in a supply chain. Decomposing donated blood into commonly transfused components, testing them against disease agents and storing them to satisfy safety of the recipient are some standard steps of this supply chain. Thus, the ultimate safety of the transfused blood depends on being safe at every step of the supply chain, most of which are governed by regulation. Traditionally the safety at each stage is validated by performing regulated checks and by having an inspection process to ensure regulatory compliance of these processes themselves. This paper describes a verification method for safety in the blood storage and transfusion processes that contribute to the safety of the vein-to-vein blood supply chain. Consequently, the paper adds semi-automated verification of the processes used in individual steps and their choreography to existing validation processes. Our previous paper [1] showed a method for verifying the safety of the methodology for the donor registration process that would contribute to the safety of the donor and the donated blood bags. This process starts with registering the donor into the system, check if the donor is suitable and collecting units of blood for a specific purpose(i.e. Whole Blood vs. Apheresis). In this paper, we extend our previous verification method to cover the safety of the rest of the blood supply chain. Specifically, we extend our verification process against standards and regulated requirements in the laboratory (unit and sample processing), storage, post donation, and transfusion. Consequently, taken together, our previous publication [1] and this work presents a verification method that covers regulation based verification of the vein-to-vein blood supply chain safety. In order to do so, we model the blood supply chain as a workflow, where each step is modeled as a process carried out by an individual or a machine and their choreography is modeled as workflow constructs. Then we consider safety regulations specified primarily by the FDA [2] and the AABB [3] and associate relevant regulations that should be satisfied by the appropriate components of the modeled workflow. We then manually translate these regulations to statements in Temporal Logic [4], thereby creating a workflow model where appropriate components satisfying a temporal logic formula (that formally states the safety condition) should ensure that the blood supply chain complies with the mandated safety regulations. Lastly, we extend an automated translator that translate the workflows and the temporal logic formulas attached to appropriate fragments of the workflow and use the theorem prover to validate that all mandated safety conditions are satisfied by the blood processing supply chain.

The rest of this paper is written as follows. Section 2, covers the regulations we considered. Section 3 provides our verification using LTL. Section 4 describes related work. Section 5 has our concluding comments.

2. The Blood Supply Chain and Relevant Safety Regulations

This section describes the vein-to-vein blood supply chain and mandatory regulations that apply to the safety of blood recipients and donors. We model the former as a workflow as shown in Figure 1. The boxes in Figure 1 represent tasks or process in the workflow and arrows represent process transitions. Double-lined boxes in Figure 1 represent compound processes (i.e. those that have sub-processes) in the modeling system. We used YAWL an executable workflow modeling language to model our workflow [5].

Figure 1 shows the main processes of the supply chain consisting of registration, suitability, collection, post donation, Blood Transfusion Services (BTS) Laboratory, and transfusion. As shown in Figure 2, complex processes (pro-

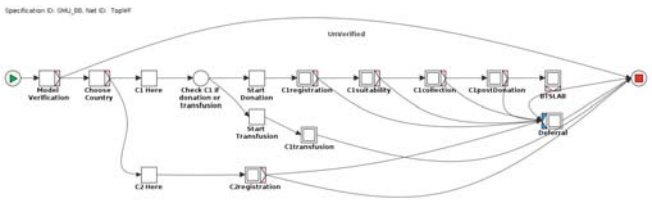


Fig. 1: Main Processes in the Main Net

cess with sub-processes) are shown in a big box including small boxes (the sub processes). The fragment of this large model analyzed in this paper covers the numbered boxes of the diagram. Each box should complete all (sub) processes within prior to commencing to execute the next process. In Figure 2 bold red arrows show the flow of work between the main processes, navy blue arrows to show transfusion flow and purple arrows to show the flow of deferrals. The flow of work is as follows: Starts with (1) Registration then suitability and collection, (2) post-donation (3) BTS Lab, or starts with (9) Transfusion request (10) Transfusion Process (11) Post-transfusion. This section will focus on (a) Updates in the Registration processes, (b) Post Donation, (c) Blood Transfusion Laboratory Processing, (d) Storage, (e) Transfusion.

The post donation process is shown in box 2 of Figure 2. This processes models tests done on the donor after blood collection and report any adverse events [3], [2]. This section also covers the relevant regulations and standards from the FDA and AABB that specify the safety requirements for the laboratory, storage and transfusion processes. We describe the regulation rules against processes in Table 1.

Table 1: Regulations and Standards

Process	FDA regulation	AABB Standard	AABB Technical Manual
Unit Processing	CFR640	5.7.4.7	Chapter 6
	Whole blood has special requirements when processing each component such as RBC, PLT, FFP, Cryo		
Sample Processing	Subpart E, CFR660, Subpart C	5.8.1, 5.8.2, 5.8.3, 5.9.3, 5.9.4, 5.9.5, 5.14	Chapter 11, table 11-1
Storage	CFR610.53	-	Chapter 9, table 9-1, Chapter 13

2.1 Blood Transfusion Services Laboratory

The first step after the donation is to take the blood components to the processing laboratory (BTS Lab) and process the bag as shown in the purple box 3 of Figure 2. This process consists of many interconnected sub-processes such as the *Unit Processing* and *Sample Processing* and *Discarding*. Each of these sub-processes has their own sub-processes as shown inside the purple box in teal, orange, light green and red colors in Figure 2. *Unit Processing* and *Sample Processing* are executed simultaneously and explained in the subsequent section.

2.1.1 Unit Processing:

Donated whole blood units are separated and processed into different blood product depending on the blood bank setup and the current demand for blood products within the BTS lab, as shown in teal color in box 4 of Figure 2. The BTS lab receives the blood units then are visually inspected and other parameters such as time, temperature, color and blood volume are checked to ensure that they are within acceptable safety ranges. Units with a lower than standard blood volume are automatically sent to generate packed red blood cells, and discard other components. Description of the volume requirements are in the AABB Technical Manual Chapter 6 and AABB standard 5.7.4.7 [3], [6]. Concurrently the FDA has also defined blood processing requirements for each of the blood products such as Whole blood, Red blood cells, Plasma, Platelets and Cryo described in CFR 640 [2]. The requirements are to ensure that the received whole blood is in a proper container with the right temperature and set of bags connected to the unit depending on the set of products to be generated.

2.1.2 Sample Processing:

Sample processing checks for the existence of infectious agents within the blood sample and is shown in orange color in Figure 2. Sample processing consists of multiple sub-processes, dedicating one sub-process for each test. Examples of these processes include ABO Rh and Antibody screening, and serologic and molecular testing for detection of infectious disease agents such as Human T-lymphotropic virus (HTLV), human immunodeficiency virus(HIV), Hepatitis (HBV, HCV).

The serology testing requirements check the donated unit against screening tests for infectious diseases that are followed based on the FDA CFR sub-part E and AABB standard and technical manual [3], [6], [2]. Box 6 in Figure 2 shows the serology testing process consisting of screening tests and confirmatory tests. Also, Figure 2 shows the deferral period as one of the requirements assigned to a donor if the serology testing results in a positive outcome. Identifying the donor's blood group is another sub-process shown in light green in Figure 2. There are several blood group systems, where the most common one is ABO, which we used. In ABO testing, the donor's blood can be interpreted as either blood type A, B, AB, or O. This is done by detecting the presence of ABO antigens on the Red cells and the absence of their corresponding antibodies in the plasma (using so-called forward and reverse blood grouping). However, in some situations, a person can develop an unexpected antibody against antigens in the other blood group systems which we modeled at a high-level in our system to capture regulatory requirement CFR 640, 660 [6], [2]. Our body immune system generates antibodies as a result of blood transfusion, pregnancy, or bone marrow/stem

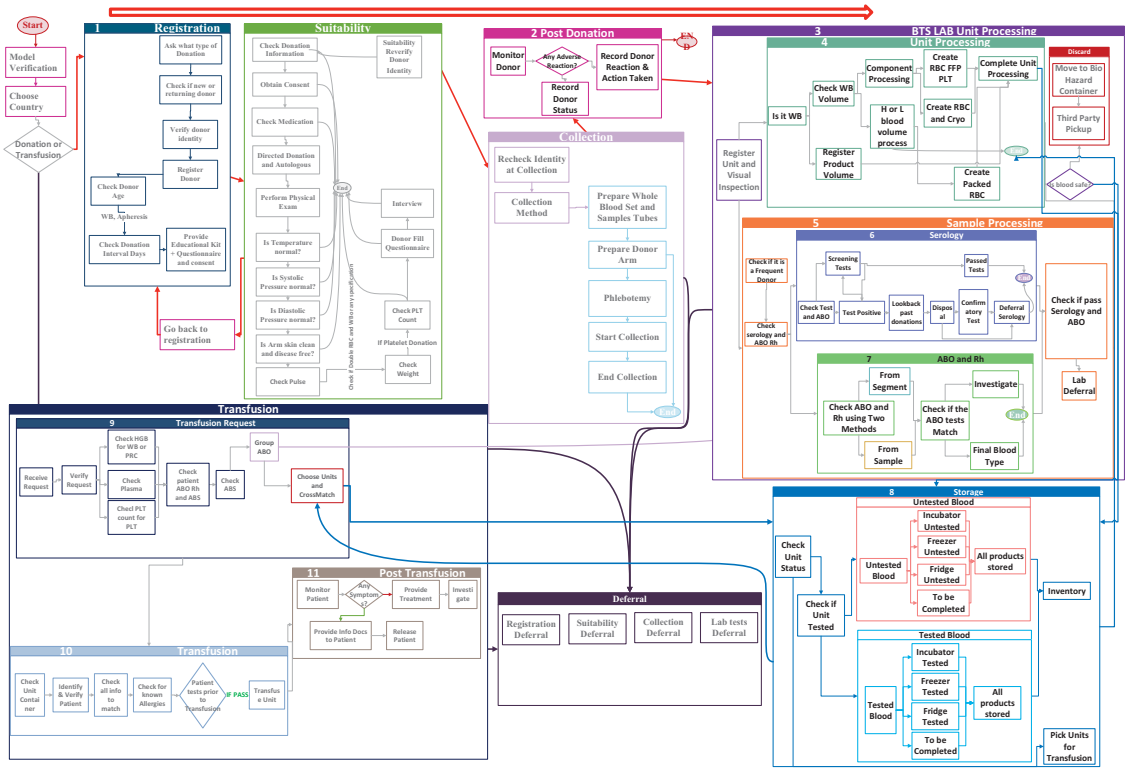


Fig. 2: High Level View of the FDA Regulations for the Blood Supply Chain.

cell transplantation. Antibodies are generated by the human immune system to fight any foreign antigen.

We modeled the blood grouping using two complex processes shown in light green box 7 of Figure 2. The first box shows a test done using the donated samples and the second test uses a unit segment. These test complying with FDA Sub-part C and CFR640 [2] where the FDA require the tests from the donor samples. However, as an extra precaution, the ABO/Rh blood group is modeled to check the blood group in two methods shown in Figure 2, box 7 in light blue and golden colors. Testing for the ABO/Rh from the segment and samples ensure accurate determination of donor's blood group and minimize the possibility of a mix-up between the unit and samples during the labeling of the samples and units prior to collecting the blood from the donor.

After completing the ABO/Rh processes from the segment and samples then the results are compared. Only if the ABO/Rh are the same results, then the final ABO is stored as is required and stated by the AABB standard 5.12 and 5.13 [3]. Otherwise, an investigation process is held to find any errors.

2.1.3 Storage

This process stores blood temporarily and then decomposed into different blood components such as Red Blood Cells (RBC), Platelets (PLT), Fresh Frozen Plasma (FFP) and Cryoprecipitated AHF (Cryo). Each of the processed blood components has specific storage requirements such as time, expiry date, temperature, usage and status as explained in FDA CFR 610.53, CFR 640, AABB technical manual Chapter 9 and AABB stand. 5.7.4.9 [3], [6], [2]. As whole blood is directly stored when it has been processed into components (Completing the Unit Processing Box 4) it is stored in the untested storage first. After the Sample processing shown in Box 5 of Figure 2 is completed and passed, all passing units are moved from the so-called *untested storage* to *tested storage*. These storage locations are kept in separate refrigerators. The modeled blood bank currently holds two of the following refrigerators, incubators and freezers to ensure tested and untested units are kept separated. Thirdly, any of the box 4 or 5 fails the unit is automatically discarded.

3. Verifying the Workflow for Safety Requirements

We translate FDA and AABB safety requirements to Linear Temporal Logic (LTL) as properties that can verify against our workflow. For this, we model the workflows as state transitions between so-called *states* of the blood supply chain. This view of the state space has many states. Figure 3 shows the complexity in our states. This diagram shows our updated main states in black, newly added states in pink and gray one were described in our previous work, of which

we omit details. Our new additions have expanded verifiable properties to 10, of which we discuss only the new ones.

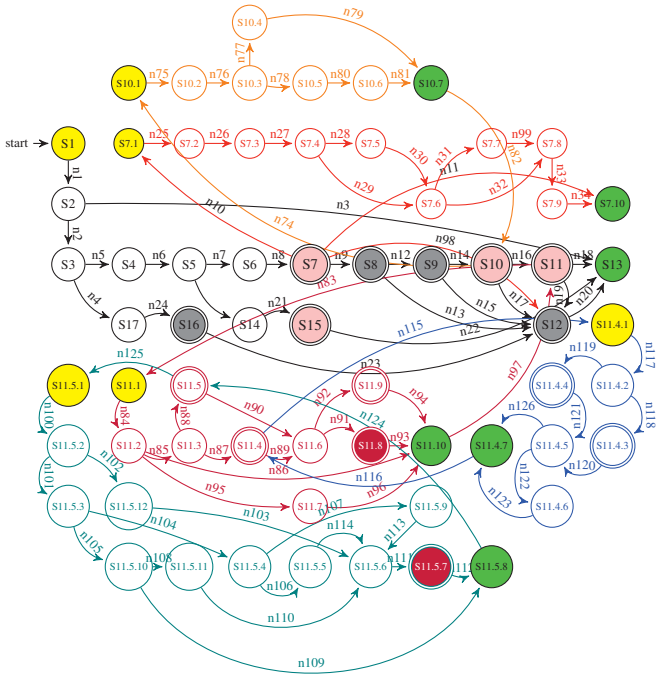


Fig. 3: Full State Diagram

3.1 Syntax

We now specify LTL syntax. Let $VAR = \{\vec{x_i}; i \geq 0\}$ be a set of variables, $CONST = \{\vec{c_i}; i \geq 0\}$ be a set of constants and $\Phi = \{p_i : i \geq 1\}$ be a set of atomic predicate symbols.

We say that $p_i(\vec{x_{i_j}})$, $p_i(\vec{x_{i_j}}) \wedge p_k(\vec{x_{k_j}})$, $p_i(\vec{x_{i_j}}) \vee p_k(\vec{x_{k_j}})$, $\neg p_i(\vec{x_{i_j}})$, $p_i(\vec{x_{i_j}})$, $\exists \vec{x_{i_j}} p_i(\vec{x_{i_j}})$, $\forall \vec{x_{i_j}} p_i(\vec{x_{i_j}})$, $p_i(\vec{x_{i_j}}) \rightarrow p_k(\vec{x_{k_j}})$ and $\Diamond p_i(\vec{x_{i_j}})$, $\Box p_i(\vec{x_{i_j}})$, $\mathcal{X} p_i(\vec{x_{i_j}})$ (sometimes this next-time operator $\mathcal{X}$ is written as $\bigcirc p_i(\vec{x_{i_j}})$) are predicates. Following standard convention, a fully instantiated predicate is one in which all variables are replaced by constants where we write $p_i(\vec{c_{i_k}})\vec{x_{i_j}}$ to indicate that the variables $\vec{x_{i_j}}$ in $p_i(\vec{x_{i_j}})$ have been replaced with constants $\vec{c_{i_k}}$.

3.2 Semantics

We now summarize the commonly used semantics of temporal logic. Let $S = \{s_i : i \geq 0\}$ be a collection of states (sometimes referred to as worlds) and an accessibility relation among states as $R \subseteq S \times S$. We assume that there is a mapping (referred to as an assignment of the fully instantiated instances of the predicate symbols), say $Inst = \{inst_k k \geq 0\}$ with the mapping $AtMap = M : Inst \mapsto \wp(S)$. Then we define the satisfaction relations for the predicates in the states as follows:

- $s_i \vDash inst_k$ if $s_i \in AtMap(in_k)$.
- $s_i \vDash inst_k \wedge inst_j$ if $s_i \vDash inst_k$ and $s_i \vDash inst_j$.
- $s_i \vDash \neg inst_k$ if $s_i \nvDash inst_k$.
- $s_i \vDash inst_k \vee inst_j$ if $s_i \vDash inst_k$ or $s_i \vDash inst_j$.
- $s_i \vDash \forall x\, p_k$ if $s_i \vDash inst_k$ for every instance $inst_k$ of p_k and the only free variable of p_k is x.
- $s_i \vDash \exists x\, p_k$ if $s_i \vDash inst_k$ for some instance $inst_k$ of p_k.and the only free variable of p_k is x.
- $s_i \vDash \Diamond inst_k$ if $s'_i \vDash inst_k$ for some $s'_i \in R^*(s_i)$, where R^* is the reflexive transitive closure of R.
- $s_i \vDash \Box inst_k$ if $s'_i \vDash inst_k$ for every $s'_i \in R^*(s_i)$, where R^* is as stated above.
- $s_i \vDash \mathcal{X} inst_k$ if $s'_i \vDash inst_k$ for some $s'_i \in R^*(s_i)$,

Through the updates made to the translator and model, the sample version expanded to 299 states $S = \{s_1, s_2 \cdots s_{16}\}$ in main states with layers of sub-states, 237 predicates labeled $P = \{t_1, t_2 \cdots t_{237}\}$ and fifty six constants. Our model checker uses X for $\bigcirc$. Sample safety requirements that are shown do not use the connective $\Box$. We show verification related to the post donation, and BTS Lab (Unit Processing, Sample Processing, and Storage).

3.3 Verification

This section describes our state space and show that state transitions satisfy associated LTL properties that model safety regulations. Each state is described below with their transitions depending on the requirements specified in our safety properties. As there are 299 states, this section will only discuss states, transitions and associated LTL properties shown in Figure 3, focusing on new states filled in pink color. They are states (S_7, S_{10}, S_{11}, S_{15}). States filled in gray (S_8, S_9, S_{12}) were covered in detail in our previous paper [1] and consequently will not be described here. In this diagram, States such as S_7, S_{10}, S_{11}, S_{15}, S_{16}, $S_{11.4}$, $S_{11.5}$, $S_{11.8}$, $S_{11.9}$, $S_{11.5.7}$ are compound states. All states ending with (1 or .1) or colored in yellow in Figure 3 are a starting state, while the colored green are ending states.

- S_1, S_2, S_4, S_6 and S_7: The user starts the workflow in states S_1 and in S_2 checks the model through automating the verification. Passing the verification the system transition the user to state S_3 the user enters the country and the system will route the user country in states S_4 or S_{17} as shown in 3. By clicking on the donation process, state S_6 is started followed by an automatic transition to S_7 to register the donor.
- S_4: This is an empty state that when entered automatically transition to Country1 workflow at S_5.
- S_5: This is a decision state to allow the user to specify if it is a Donation or Transfusion by clicking on donation or transfusion on the system. Choosing a donation process will route the user to state S_6 directly followed by state S_7. Choosing transfusion will route the user to state S_{14} directly followed by state S_{15}.
- S_8: In this composite state, the system checks if the donor is suitable for the donation by checking the donor vitals. Details are covered in [1].

- S_9: In this composite state, the system checks if the collection process's safety controls have been applied such as re-verifying the donor [1].
- S_{10}: This is a composite state shown in Figure 3. The blood bank staff enters if a donor has any adverse reactions t_{220} if not then a transition n_{16} to S_{11}. Otherwise, transition n_{17} goes to state S_{12} for deferral. This state output [donorReactionN]
- S_{11}: This is a composite state consisting of sub-states $S_{11.1}$ to $S_{11.10}$ checks the blood unit for infectious diseases and donor blood grouping. The composite state then outputs the following to verify the donated unit is safe (t_{221} , t_{222} , t_{223} , t_{224} , t_{225} , t_{226} , t_{227} , t_{228} , t_{229} , (t_{230} , (t_{231} , t_{232} , (t_{233}) or (t_{236} , t_{237})))) on trigger transitions n_{18} to S_{13}. This state output [HTLV, AntiHBc, HBsAG, AntiHCV, HIVNAT, HCVNAT, HBVNAT, Syphillis, AntiHIV, recordedVolume, UnitBagVolumeCap].
- S_{12}: In this composite state the system routes the deferrals based on the type of deferral as specified by the safety controls and standards. This paper covers only registration, collection, post donation and BTS lab deferrals which are shown in Figure 3 as transitions n_{13}, n_{15}, n_{17}, n_{19}, and n_{23}. The composite states consist of $S_{12.1}$ and $S_{12.2}$
- S_{13}: This is an empty state where all transitions end.
- S_{14}: This is an empty state that transition to state S_{15} for transfusion.
- S_{15}: This composite state checks the transfusion request, apply the transfusion and ensure the patient has no adverse reactions. Once the transfusion is complete, a transition n_{22} to state S_{13} to complete and end of transitions.
- S_{16}: This is a composite state Country 2 registration.
- S_{17}: Empty state to transition to Country 2 workflow.

State S_7 decomposes into 10 sub-states labeled $S_{7.1}$ through $S_{7.10}$. This complex state has been updated from our previous work as it routes the donor correctly through checking any donations history. Also, ensuring all donors (new or returning) are safe to donate through adding an extra checkpoint to check for donation intervals for specific donors.

State S_{10} decomposes sub states $S_{10.1}$ through $S_{10.10}$ where donors are monitored to ensure that no adverse reactions occur and if they do all symptoms should be recorded.

State S_{11} decomposes into sub-states $S_{11.1}$ through $S_{11.10}$. The composite state S_{11} checks the donated unit and process the unit into different products. Due to the page limitation, we will focus on two sub-states $S_{11.4}$ and $S_{11.5}$. These states show the transitions needed for testing blood for infectious diseases/blood grouping and processing the unit to different blood products.

$S_{11.4}$: is a composite state consists of states $S_{11.4.1}$ through $S_{11.4.10}$ used to model sample Processing.

- $S_{11.4.1}$: This state starts the sub-states. When $S_{11.4}$ is entered this state it imports [donationID] to be utilized in the sub-states. Then it transitions to $S_{11.4.2}$.
- $S_{11.4.2}$: This is an empty state that transitions into two composite states $S_{11.4.3}$ and $S_{11.4.4}$ simultaneously.
- $S_{11.4.3}$: This composite state checks for infectious disease serology from the donors sample. This state inputs [donationID] and output [RIBA, HBVNAT, antiHIV, HCVNAT, HIVNAT, HBsAG, antiHc, HTLV, Syphillis, Treponomal, antiHCV]. On trigger (t_{218} - t_{227}) transition n_{120} to state $S_{11.4.5}$
- $S_{11.4.4}$: This composite state checks the donor's blood group the presence of ABO antigens on the red cell and the absence of their complimentary antibodies in plasma cells. This state outputs [finalABOType].
- $S_{11.4.5}$: This state checks the outputs from states $S_{11.4.3}$ and $S_{11.4.4}$ by checking infectious diseases in all of the serology tests are negative which then transitions to state $S_{11.4.7}$. A positive serology will transition to state $S_{11.4.6}$.
- $S_{11.4.6}$: This is an empty state that on start input n_{123} transitions to $S_{11.4.7}$.
- $S_{11.4.7}$: This state end a composite state and output [finalABOType, RIBA, HBVNAT, antiHIV, HCVNAT, HIVNAT, HBsAG, antiHc, HTLV, Syphillis, Treponomal, antiHCV] to state $S_{11.4}$.

$S_{11.5}$: This composite state consist of component states $S_{11.5.1}$ through $S_{11.5.10}$ to model unit processing.

- $S_{11.5.1}$: This state start gets the inputs [donatioType, donationTypeSpecf, donationID, UnitBagVolumeCap, ForProcessingTo] from state $S_{11.5}$
- $S_{11.5.2}$: This state input [donationType] to check if the donation is for Whole blood or Apheresis. Whole blood donations are transitioned to state $S_{11.5.3}$ while Apheresis are transitioned to state $S_{11.5.4}$.
- $S_{11.5.3}$: This state checks if the donated unit of blood is within the regulated FDA safety standard for the donated unit volume, depending on the unit bag capacity used which is transitioned to state $S_{11.5.4}$. Donated units with higher or lower than standard volume are transitioned to state $S_{11.5.10}$.
- $S_{11.5.4}$: This state checks the blood components the user specified for the whole blood processing. The user is required to input [ForProcessingTo] the system checks this variable to identify if blood to be processed as *RBC FFP PLT* transitions to state $S_{11.5.5}$ or *RBC and Cryo* that transitions to state $S_{11.5.9}$. On trigger (t_{191}) transition n_{106} to state $S_{11.5.5}$.
- $S_{11.5.5}$: This state creates new products for RBC, FFP, PLT and checks the volume of each bag and then outputs [productID01-03, processedTo01-03, productRecordedVolume01-03].
- $S_{11.5.6}$: This end of the sub-process outputs [fi-

nalABO, donationID, unitTested, processedTo01-03, productID01-03, anticoagulant, productVolume01-03] to state $S_{11.5}$

- $S_{11.5.7}$: This composite state stores the blood products. It input [finalABO, donationID, unitTested, processedTo01-03, productID01-03, anticoagulant, productVolume01-03]. This state recognizes that the unit has not been tested and so [unitTested] is false.
- $S_{11.5.8}$: This state completes all unit processing. As the ABO is not tested here, the donor blood group [finalABO] is generated as *none*.
- $S_{11.5.9}$: This state creates new products for RBC and Cryo and checks the volume of each bag then it output the following [productID01-02, processedTo01-02, productRecordedVolume01-02].
- $S_{11.5.10}$: This state checks the donated unit volume. On input [unitBagVolumeCap, recordedVolume], it compares if the unit capacity versus the recorded volume if it is lower it transition to state $S_{11.5.11}$ to created packed RBC. Higher volume transition to state $S_{11.5.6}$, an empty state that completes the unit processing and transitions to state $S_{11.5.7}$. On trigger (t_{186}, t_{187}, t_{189}, t_{190}, t_{193}, t_{194}, t_{196}, t_{197}) transition n_{108} to state $S_{11.5.11}$.
- $S_{11.5.11}$: This state creates packed RBC and records the volume and productID and outputs [productRecordedVolume01, productID01]
- $S_{11.5.12}$: This state registers apheresis products. Currently, an empty state left for future expansion.

3.4 Lemma Verification

In this section, we explain how each lemma verifies a set of safety requirements described in each state as specified in the previous section. The collection of these lemmas show that the property holds in each state combines as a proof of correctness of the main theorem that end-to-end safety of the modeled workflow holds for FDA and AABB regulations quoted in the previous sections. The structure of our Lemmas are shown in Figure 4. The lemmas are numbered using the same counter as the one used in the LTL properties in the model checker. In this paper, we will focus Lemma 7 and Lemma 8, due to space limitations.

Table 2: Explanation of the Lemmas

Theorem/Lemmas in English	Lemma
Lemma7: This lemma holds in state $S_{11.5.1}$ where it checks if the whole blood units are processed correctly based on the FDA and AABB standards and regulations in all of its states from $S_{11.5.1}$ to $S_{11.5.12}$.	#property XX(t184 && X(t185 && t186 && t187 ‖ t188 && t189 && t190) && XXt191) ‖ XX(t184 && X!(t185 && t186 && t187 ‖ t188 && t189 && t190) && XX(t192 && t193 && t194 ‖ (t195 && t196 && t197)))
Lemma8: This lemma holds in state $S_{11.4.1}$ to check donor samples for infectious diseases and blood grouping.	#property XXXX(t198 ‖ t199 ‖ t200 ‖ t201 ‖ t202 ‖ t203 ‖ t204 ‖ t205 ‖ t206 ‖ t207 ‖ t208)

Main Theorem: This theorem verifies that all states are safe as sepcifed in the regulations and standards

- **Lemma 2** ··· **5:** This Lemma hold in states $S_{7.1}$, $S_{8.1}$, $S_{9.1}$, $S_{12.1}$ and verifies the donor registration processes
- **Lemma 7:** This lemma holds in state $S_{11.5.1}$ that checks if the whole blood units are processed correctly based on the FDA and AABB standards and regulations in all of its states in between $S_{11.5.1}$ and $S_{11.5.12}$
- **Lemma 8:** This lemma holds in state $S_{11.4.1}$ that verifies donor sample check for infectious diseases and blood grouping
- **Lemma 9:** This lemma holds in $S_{11.4.3.1}$ that verifies donor blood sample check for infectious diseases based on the FDA and AABB regulation and standard
- **Lemma 10:** This lemma holds in state $S_{11.4.4.1}$ where donor blood group is checked using two methods (1) ABO test from unit segment (2) ABO test from donor sample

Fig. 4: Lemmas Structure

*Theorem 1 (*Main Workflow Verification*):* #property t1&&X(t59&&Xt57&&XX((t11 && t16)‖ (t12 && t17) ‖ (t13 && t18) ‖ (t14 && t19) ‖ (t15 && t147)) && (t8 && t9 && t10) && (t2 && t54) &&XXX(t41 && ((t26 && t19) ‖ (t27 && t23) ‖ (t28 && t23) ‖ (t29 && t25) ‖ (t30 && t20) ‖ (t31 && t20) ‖ (t32 && t21) ‖ (t33 && t21) ‖ (t34 && t22) ‖ (t35 && t22) ‖ (t36 && t21) ‖ (t37 && t21) ‖ (t38 && t21) ‖ (t39 && t21) ‖ (t40 && t24)) && t42 && (t148 && t46) ‖ (t2 && t45)) && t47 && t48 && (t9 && t10 && t8) && ((t3 && t52 && t49) ‖ (t3 && t53 && t50) ‖ (t4 && t52 && t51) ‖ (t5 && t52 && t51) ‖ (t6 && t52 && t51) ‖ (t7 && t52 && t51) ‖ (t4 && t53 && t51) ‖ (t5 && t53 && t51) ‖ (t6 && t53 && t51) ‖ (t7 && t53 && t51)) && XXXX(t8 && t9 && t10 && t55 && t56 && Xt220 && XX(t221 && t222 && t223 && t224 && t225 && t226 && t227 && t228 && t229 && (t230 && (t231 && t232 && (t233) ‖ (t236 && t237))))))

Proof: ■

```
[noha@localhost tools]$ ./divine verify HIMS16.prop1.dve -p assert
 input: HIMS16.prop1.dve
 property assert: assertion safety
 ---------------- Reachability ----------------
  searching...  37 states, 96 edges
 ========================================
              The property HOLDS
 ========================================
```

*Lemma 7 (*Unit Processing Workflow Verification*):* #property XX(t184 && X(t185 && t186 && t187 ‖ t188 && t189 && t190) && XXt191) ‖ XX(t184 && X!(t185 && t186 && t187 ‖ t188 && t189 && t190) && XX(t192 && t193 && t194 ‖ (t195 && t196 && t197)))

Proof: ■

*Lemma 8 (*Sample Processing Workflow Verification*):* #property XXXX(t198 ‖ t199 ‖ t200 ‖ t201 ‖ t202 ‖ t203 ‖ t204 ‖ t205 ‖ t206 ‖ t207 ‖ t208)

```
[noha@localhost tools]$ ./divine verify HIMS16.prop7.dve -p assert
 input: HIMS16.prop7.dve
 property assert: assertion safety
 ---------------- Reachability ----------------
  searching...  37 states, 89 edges
 ============================================
```

Proof: ■

```
[noha@localhost tools]$ ./divine verify HIMS16.prop8.dve -p assert
 input: HIMS16.prop8.dve
 property assert: assertion safety
 ---------------- Reachability ----------------
  searching...  37 states, 91 edges
 ============================================
              The property HOLDS
 ============================================
```

4. Related Work

Henneman et al. [7] state the importance of formal modeling in transfusion therapy to increase patient safety. The authors claimed that Little-JIL Process Language they invented is a strong rigorous, precise usable for transfusion therapy verification. However, the modeled in the paper shows high level detail in transfusion that is missing several safety requirements.

Avrunin et al. [8] utilized the modeling tool (Little-JIL). As described by the authors. The tool provides superior functionality in precision and evaluation. The paper described transfusion processes example, where they were able to capture a deadlock in the process that aided a nurse when teaching the material. This was done through using Little-JIL for modeling and PROPEL for describing the properties.

Lu et al. [9] used many risk analysis techniques such as fault tree analysis, reporting analysis, etc··· to identify weaknesses in processes. This paper has shown a level of detail and critical points in blood transfusion in a flow chart that was not captured in many publications. Although the presented flow chart is not a formal model, it provides a high level view of actual transfusion processes and identifying risk priority number in blood transfusion.

Damas et al. [10] emphasize the importance of formally modeling in medical process and models a simple real cancer therapies model which is then analyzed using g-HMSC modeling tool [10]. The g-HMSC modeling tool uses four different operators: Union, Restriction, Focus and Merge. These operators are used to composition (for sub-processes), decomposition (for going back to higher level processes). On the other hand, YAWL is a user friendly and expressive modeling tool that is executable.

Bertolini et al. [11] describe the importance of health information systems which has lead them to create a collaborative health-care workflows (CHWF). The workflow modeling was created using CSP for verification. Also, they created UML models to aid visualizing the system [12].

5. Conclusions

This paper shows that many important steps such as the BTS Lab, Storage, and Transfusion can be verified against the FDA and AABB regulations that govern them. To ensure blood/transfusion safety, it is important to ensure vein-to-vein processes comply with the safety regulations and standards to provide safe blood. This paper completes the blood bank supply chain from vein-to-vein and verifies the workflow against the regulatory FDA and AABB standard in safety requirements. This formal model is an executable model.

6. Acknowledgment

This work is sponsored in part by the Saudi Arabian Cultural Mission (SACM) and King Abdulaziz University (KAU). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, by the Government of the Kingdom of Saudi Arabia.

References

[1] N. Hazzazi, B. Yu, D. Wijesekera, and P. Costa, "Using temporal logic to verify blood supply chain safety."

[2] *CFR Mini-Handbook*, 2015th ed. AABB, 2015.

[3] J. Levitt, *Standards for blood banks and transfusion services*, 29th ed. Bethesda, MD: AABB, 2014.

[4] V. Goranko and A. Galton, "Temporal Logic," in *The Stanford Encyclopedia of Philosophy*, winter 2015 ed., E. N. Zalta, Ed., 2015.

[5] Y. Foundation. (2012, 3) Yawl - user manual.

[6] Mark K. Fung, Brenda J. Grossman, Christopher Hillyer, and Connie M. Westhoff, Eds., *Technical Manual*, 18th ed. AABB, 2014.

[7] E. A. Henneman, G. S. Avrunin, L. A. Clarke, L. J. Osterweil, C. Andrzejewski, K. Merrigan, R. Cobleigh, K. Frederick, E. Katz-Bassett, and P. L. Henneman, "Increasing patient safety and efficiency in transfusion therapy using formal process definitions," *Transfusion Medicine Reviews*, vol. 21, no. 1, pp. 49–57, 2007.

[8] G. S. Avrunin, L. A. Clarke, L. J. Osterweil, S. C. Christov, B. Chen, E. A. Henneman, P. L. Henneman, L. Cassells, and W. Mertens, "Experience modeling and analyzing medical processes: Umass/baystate medical safety project overview," in *Proceedings of the 1st ACM International Health Informatics Symposium*, ser. IHI '10. New York, NY, USA: ACM, 2010, pp. 316–325.

[9] Y. Lu, F. Teng, J. Zhou, A. Wen, and Y. Bi, "Failure mode and effect analysis in blood transfusion: a proactive tool to reduce risks," *Transfusion*, vol. 53, no. 12, pp. 3080–3087, 2013.

[10] C. Damas, B. Lambeau, and A. van Lamsweerde, "Transformation operators for easier engineering of medical process models," in *Software Engineering in Health Care (SEHC), 2013 5th International Workshop on*, May 2013, pp. 39–45.

[11] A. Seyfang, K. Kaiser, and S. Miksch, "Modelling clinical guidelines and protocols for the prevention of risks against patient safety." *Studies In Health Technology And Informatics*, vol. 150, pp. 633 – 637, 2009.

[12] C. Bertolini, M. Schäf, and V. Stolz, *Foundations of Health Informatics Engineering and Systems: First International Symposium, FHIES 2011, Johannesburg, South Africa, August 29-30, 2011. Revised Selected Papers*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, ch. Towards a Formal Integrated Model of Collaborative Healthcare Workflows, pp. 57–74.

Design of Health Care System for Disease Detection and Prediction on Hadoop Using DM Techniques

Dingkun Li, Hyun Woo Park, Erdenebileg Batbaatar, Yongjun Piao, Keun Ho Ryu
Database/Bioinformatics Lab, School of Electrical & Computer Engineering,
Chungbuk National University, Cheongju, South Korea

Abstract - *Apache Hadoop MapReduce is a well-known software framework for developing applications that process vast amounts of data. Combined with traditional Data Mining (DM) techniques, it provides a more powerful way to handle data with high speed, safety and accuracy. In our work, we took advantages of both Hadoop and DM techniques to design a comprehensive, real-time and intelligent mobile healthcare system for disease detection and prediction. It provides an assistant system for user self-healthcare as well as a complementary system for doctors' diagnosis on their daily work. Due to the time limit, the whole system has only been partially implemented, but the whole design work has been finished, the 4-node Hadoop experiment environment has been setup in the lab to do some experiments for further analysis and the experiment result is promising.*

Keywords: Hadoop, Data Mining, Healthcare System, Risk Factor, Disease Detection, Disease Prediction

1 Introduction

Hadoop is one of the most important and popular techniques during last few years with the emergence of the cloud computing concept. It has a great power to handle a huge amount of data of any kind. Data Mining (DM) is one of the most popular and promising techniques of discovering the meaningful information from varies massive data. The most exciting part is taking advantages of using both Hadoop and DM techniques to provide a greater powerful way to handle data with high speed, safety and accuracy.

Recent years, DM techniques have been widely used in healthcare field due to its efficient analytical methodology for detecting unknown and valuable information in health data as well as detection of the fraud in health insurance, availability of medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods. It also helps the healthcare researchers for making efficient healthcare policies, constructing drug recommendation systems, developing health profiles of individuals [1].

A lot of research works have been done for healthcare by using DM techniques. In [3, 4, 5] the authors use classification, regression techniques to predict Cardiovascular Disease, Heart Disease etc. In [6, 7], it provides integrated DM techniques to detect chronic and physical diseases. Some other research works [8, 9] developed new methodology and framework for healthcare purpose but all these researches took the advantages of the DM techniques.

Meanwhile, In last decade, cloud computing services developed very quickly and provided a new way to establish new health care system in a short time with low cost. The "pay for use" pricing model, on-demand computing and ubiquitous network access allows cloud services to be accessible to anyone, anytime, anywhere [2].

Hadoop framework on cloud computing [10] has been developed for delivering healthcare as a service. A wide variety of organizations and researchers have used Hadoop for healthcare services and clinical research projects [11]. Taylor, R.C. gave a detailed introduction to how Hadoop is used in bioinformatics [12] and Schatz M.C. developed an OSS package named CloudBurst that provides a model for parallelizing algorithms using Hadoop MapReduce [13]. Indeed there are many important works made great contributions to healthcare field by using Hadoop framework.

The purpose of our work is to takes advantages of both Hadoop and DM techniques to design a comprehensive, real-time and intelligent mobile healthcare system for disease detection and prediction. It is designed to provide an assistant system for user self-healthcare as well as a complementary system for doctors' diagnosis on their daily work.

The contributions of our system are: (1) We designed a comprehensive healthcare system which covers main aspects of the healthcare like disease detection and prediction. (2) We explored the possibility of using Hadoop and DM techniques on healthcare big data. (3) The system provides flexible communication between system and users. (4) The system guarantees real-time data transaction in very low cost.

The rest of the paper is organized as follows: We describe the related work in section 2. An overview of our system will be introduced in section 3 and its implementation detail will be described in section 4 followed by experiment result in section 5. Section 6 concludes our work and depicts future work.

2 Related Work

This section briefly describes the DM, Cloud platform services, Hadoop and other related services.

2.1 Data Mining (DM)

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the web, other information repositories, or data that are streamed into the system dynamically.

Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. In general, such tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks include association, clustering, summarization etc. characterize properties of the data in a target data set. Predictive mining tasks include classification, regression etc. perform induction on the current data in order to make predictions [6].

2.2 GCM and GCSql

Google Cloud Messaging (GCM) for Android is a service that allows you to send data from your server to your users' Android-powered device, and also to receive messages from devices on the same connection. The GCM service handles all aspects of queueing of messages and delivery to the target Android application running on the target device, and it is completely free.

Google Cloud SQL (GCSql) uses MySQL deployed in the Cloud and therefore the user gets all the benefits of using Explore Analytics with MySQL. Explore Analytics provides direct connectivity to Google Cloud SQL for live reporting, allowing you to deliver superb data analysis, visualization, and reporting. The data resides in Google Cloud SQL instance and there's no need to transfer the data to Explore Analytics.

Both Cloud services are available on Google Cloud Platform [15].

2.3 Hadoop

Diagram in Figure 1. shows the architecture of Hadoop2.

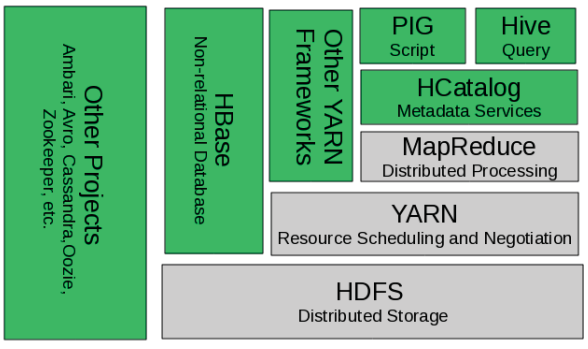


Figure 1. Hadoop 2 Architecture

Hadoop consists of HDFS (Hadoop Distributed File System), HBase, and Hadoop MapReduce which can analyze big data [16]. It is an open source framework that writes and implements an application program for processing big data.

HDFS is made up of a Master Node and several Slave node. The Master Node consists of a Name Node that controls an access to a client file and a Job Tracker which accomplishes the scheduling about the given jobs. The Master Node also manages the name space of HDFS [17].

The MapReduce [18] is a Distributed and Parallel processing model of data based on a Key/Value pair. It provides a scalability responding to data growth caused by distributed and parallel processing and minimizes network traffic caused by data movement among nodes. The MapReduce in Figure 1 generates an intermediate result with the key/value by accomplishing MapReduce based on input data. The intermediate result grouped by key value is transferred to a Reduce Task. The Reduce Task integrates all the intermediate keys and transfers the final result to the HBase.

2.4 Healthcare Management

A fully integrated and comprehensive healthcare management that includes the integrated interconnection and interaction of the patient, health care provider, utilization reviewer and employer so as to include within a single system each of the essential participants to provide patients with complete and comprehensive pre-treatment, treatment and post-treatment health care and predetermined financial support therefor [14]. The system developed should have the ability to connect all participants together for the purpose of providing high quality healthcare services.

3 Design of Main Framework

We give an overview of the whole system design in this section, and then we will describe the design detail in section 4. The system architecture is depicted in Figure. 2.

1) Data Collection Module

The system is designed to use three ways to collect data.

A key aspect of health application is the acquisition of people's health data through the use of the internet combined with all kinds of the mobile devices such as phone, watch, ring, etc. This system is called mobile health sensor network (MHSN) which collects sensor data like heart rate, walking speed etc. as well as sport information like bicycling, walking steps etc.

Meanwhile, the system collects data from the user input by using mobile devices. The system app concludes sport, nutrition intake input activities to obtain related information.

What's more, the system provides interface for data import. The data such as patients' basic information,

disease history, examine result, clinical data, etc. obtained from the hospital and Korean national health care center is imported by using web service. Also downloaded public data like twitter, facebook data with disease information can also be imported to the system.

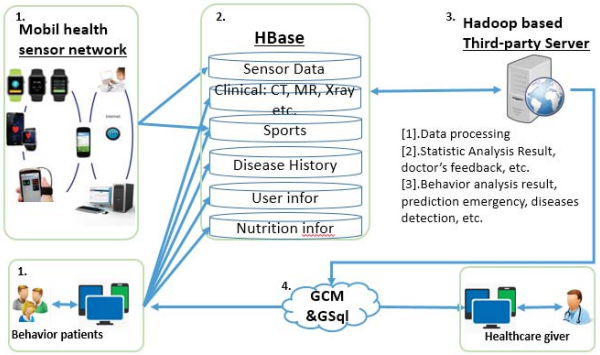


Figure.2 System Architecture

2) Data Storage Module

The data collected by the system is of three types: structured the semi-structured, and the unstructured data. Firstly, all these three kinds of data will be stored in HBase as it is quiet suitable for mass data preprocessing and storage. Then this data should be converted into structured data for further processing.

3) Third-party Server Module

Hadoop based Third-party Server (TPS) Module is the key module of the whole system, all the data processing and analysis work will be done by this module. It responses for data statistical analysis, patient emergency detection, disease prediction and detection.

Hadoop MapReduce is essentially a scheduling framework that processes data that can be sliced into different splits of disease detection, prediction etc. tasks. Each Map task works on its own input data split without having to interact with other Map tasks at all. While each Reduce task collects all the analysis result, sorts and stores it in the HBase.

TPS also responses for message like data analysis results generation. These result will be sent to Cloud Service Module.

4) Cloud Service Module

After receiving result and processed data from the TPS, the Cloud Service Module will be used to store data and transfer data. This model is implemented by using GCSql and GCM Cloud services. When receiving the requests from the TPS, Cloud model responses immediately according to these requests, stores data or sends data to the devices registered to it.

Finally, the end user, caregiver who interacts with the mobile devices to check the latest status of the patients. When there is an urgent situation such as heart disease detected by the TPS, it will give warning to the caregiver on the mobile devices through Cloud model and caregiver can supervise the patient in real-time on the other side.

4 System Design Detail

This section describes more idea about system design. The previous step is the basic of the next step.

4.1 Data Preprocessing

To deal with huge data size, feature selection technique has the potential to identify the most useful information from the data and reduce the dimensionality in such a way that the most significant aspects of the data are represented by the selected features. The stored data is thus transformed into new space such that the resultant data becomes easier to be separated into different classes.

As it has been mentioned before, the system collects three kinds of data: structured, semi-structured and unstructured data. In order to obtain high quality structured data set, database processing, Natural Language Processing (NLP) and image processing techniques combined with DM data preprocessing techniques will be used by TPS to process these different kinds of the data and transform it into computerized patient record (CPR), or structured data record. The result will be stored in the HBase.

4.2 Risk Factors Selection

Risk factor (RF) is something that increases a person's chances of developing a disease. For example, cigarette smoking is a risk factor for lung cancer, and obesity is a risk factor for heart disease. Actually all the patient's attributes can be treated as the RF like age, gender, ECG result, blood sugar, number of major vessels etc. The purpose of the RF selection is to find the key RF that may have greater chance than the other factors of developing a certain disease. In data mining area information Gain, GainRatio or Gini index are the basic common used methods for attributes selection.

Before RF selection, TPS needs to get all the diseases related information from the whole data set because the original data contains information of healthy people as well. Only patient data will be selected for further analysis. All the RF will be generated based on this data. In order to enhance the precision of the system, domain experts will be surveyed to provide risk factors for a certain kind of the disease.

4.3 Disease Rule Generation

Disease rule has the format like IF THEN rule, for example: IF (age>46.5, Fat_intake>42.28mg/day, married) THEN (hypertension = yes). The system will mine all these diseases related rules from the training data set.

After that, basic association rule mining algorithm like Apriori, common used decision tree algorithm like C4.5, CART will be used to generate the k-risk factor rules.

Among the decision tree algorithms, the result accuracy of random forest algorithm is superior to others in most conditions. So for the next stage of research, these algorithms will be tested to find one which fits for the most of the data set with high accuracy.

The most important thing we need to consider about is to improve the accuracy of the disease rules. Combining Hadoop techniques with DM techniques, it is pretty confident that the system will get the high accurate rules from the big amount of the data set [19].

Finally, the rules generated will be stored in the HBase for further processing.

4.4 Disease Detection

Since the disease rules have been detected and stored in the HBase, it is easy to detect diseases.

For a certain disease, usually there are more than one rules related to this disease. Compared with these rules, if the matching rate > =β (expert defined threshold, eg: 80%), the TPS will treat the object as the patient. For example, there are 5 rules for heart disease, when there are 4 rules matching above 5 rules, the heart disease will be detected with its expectation as 80%. An alert or report will be sent to healthcare giver by cloud service module. The procedure is given in Figure 3.

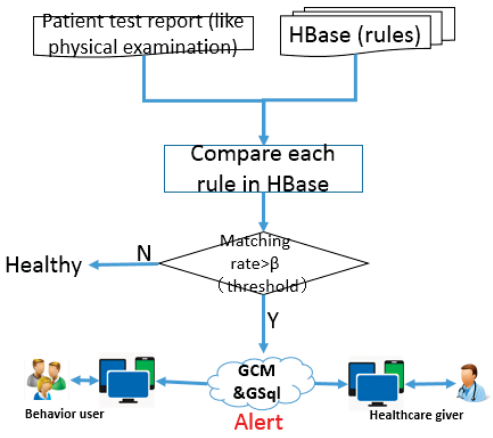


Figure 3.Disease Detection Procedure

4.5 Disease Prediction

In order to predict the disease, the prediction model will be integrated into the system. While the MARS [19] is an easy, wildly used model called Multivariate Adaptive Regression Splines with high accuracy. The general MARS model is given as below:

$$y_i = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{Km} [S_{km}(x_{v(k,m)} - t_m)] \quad (1)$$

Where $\prod_{k=1}^{Km} [S_{km}(x_{v(k,m)} - t_m)]$ is the Basis Function (BF) which relates to the domain knowledge. a_0 and a_m are both parameters that function like the coefficients in linear regression. M is the number of BF, as calculated by evaluating rules. K_m is the number of truncated linear functions multiplied in the *m*th basis function. The quantities S_{km} take on values of ± 1 to indicate the (right/left) sense of the associated step function. The v(k,m) term labels the predictor variables, and t_m represents the threshold values of each BF.

This model will be used to generate the regression model for disease prediction based on the big data set. The actual user data will be used to compare with prediction model for a certain disease risk factor. If the trend is very similar, there is great chance that this user will have the same disease. A warning will be sent to the user or doctor for further processing. At the same time, the suggestion will be given to the user about the ways to prevent this kind of disease. The procedure is given in Figure 4.

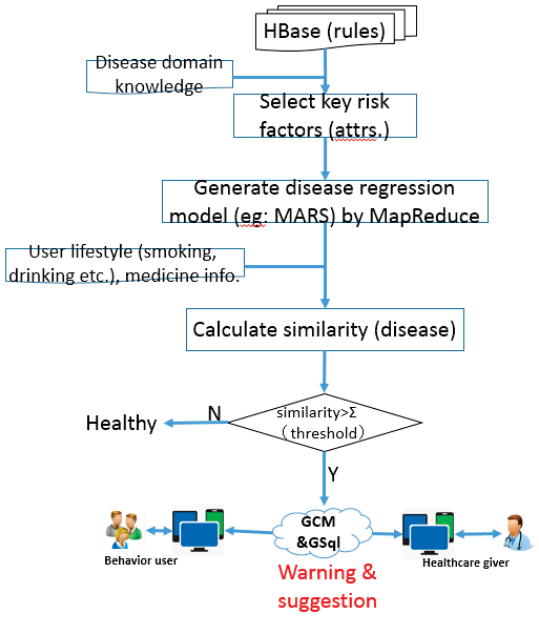


Figure 4.Disease Prediction Procedure

5 Experiment

Figure 5 shows 4-node Hadoop develop environment we used to analyze data. Ubuntu 15.04 has been installed on each server and Hadoop 2.7.2 has been configured and setup on these servers as a cluster. Figure 6 shows the app we are developing to collect user input data including normal life activity data, clinical data and history data. Meanwhile, it can be used to display statistical analysis result about the data set, like disease related information such as age, region, occupation etc. We combined simulated data and small number of real data as testing data set. Right now the TPS can interact with the android devices through GCM of cloud module. Several devices have been used for the purpose of testing including Nexus 5, Nexus 7, LG G pad 8.3, etc.

Figure 6 indicates several interfaces of the app, including login, user information (basic and clinic), food intake, and statistical analysis result interfaces.

Figure 5.Multi-node Hadoop Environment

Figure 6.App Interface

We have used public data downloaded from the Korea National Healthcare Center (KNHC) as the testing data which contains more than 690,000 patients' information including personal basic information, disease information, clinic information etc. An algorithm called disease count has been designed and implemented, the pseudo-code is shown in Algorithm 1. The result of disease count (algorithm 1) is given in Table 1.

Algorithm 1. Disease count

```
1.class Mapper
2   method map (HBase table)
3       for each instance row in table
4           write ((diseasei, patientID), 1)
5
6. class Reducer
7.     method reduce ((diseasei, patientID), ones[1,1,1,…n])
8.       sum=0
9.       for each one in ones do
10.          sum+=1
11.      return ((diseasei, patientID),sum)
```

Table 1 Disease count computing result

Disease	Hypertension	Dyslipidemia	Ostarthritis	Diabetes	Asthm
Count	9,385	3,213	5,223	3,500	923

6 Conclusion and Future Work

In the first stage of research, some basic ideas of the system design have been given in our paper. We designed and partially implemented a comprehensive healthcare system for disease prediction and detection. The system is designed and developed according to the unobtrusive, easy to deploy, effective, low-cost and real-time principles. We explored the possibility of combining Hadoop and DM techniques to handling big healthcare data. Due to the time limit, the whole system has only been partially implemented, but the whole design work has been finished, the 4-node Hadoop experiment environment has been setup in the lab to do some preparation experiments for further analysis and the experiment result is promising.

For the next stage of research, first thing is to collect big amount of the data for rule mining work. The more the data, the more accurate the mining result. Then implement the whole system, do in-depth simulations to validate the system performance, in particular, simulate multiple scenarios to confirm its scalability so as to apply our system to real environment. We also plan to extend our system to iOS platform in future.

Acknowledgement

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No.2013R1A2A2A01068923) and the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2016-H8501-16-1013) supervised by the IITP(Institute for Information & communication Technology Promotion).

7 References

[1] H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, 2005.

[2] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.H. Katz, A.Konwinski, G. Lee, D.A. Patterson, A. Rabkin, I. Stoica, M.Zaharia, "Above the Clouds: A Berkeley View of CloudComputing", UCB/EECS-2009-28, 2009 Feb 10.

[3] Dangare C S, Apte S S. "Improved study of heart disease prediction system using data mining classification techniques"[J]. International Journal of Computer Applications, 47(10): 44-48, 2012.

[4] M. Kumari and S. Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", IJCST, ISSN: 2229- 4333, vol. 2, no. 2, 2011.

[5] A. A. Aljumah, M. G.Ahamad and M. K. Siddiqui, "Predictive Analysis on Hypertension Treatment Using Data Mining Approach in Saudi Arabia", Intelligent Information Management, vol. 3, pp. 252-261, 2011.

[6] M.-J. Huang, M.-Y. Chen and S.-C. Lee, "Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis", Expert Systems with Applications, vol. 32, pp. 856-867, 2007.

[7] S. H. Ha and S. H. Joo, "A Hybrid Data Mining Method for the Medical Classification of Chest Pain", International Journal of Computer and Information Engineering, vol. 4, no. 1, pp. 33-38, 2010.

[8] Amendola S, Lodato R, Manzari S, et al. "RFID technology for IoT-based personal healthcare in smart spaces"[J]. Internet of Things Journal, IEEE, 1(2): 144-152, 2014.

[9] Jung E Y, Kim J, Chung K Y, et al. "Mobile healthcare application with EMR interoperability for diabetes patients"[J]. Cluster Computing, 17(3): 871-880, 2014.

[10] Kaur P D, Chana I. "Cloud based intelligent system for delivering health care as a service"[J]. Computer methods and programs in biomedicine, 113(1): 346-359, 2014.

[11] Horiguchi H, Yasunaga H, Hashimoto H, et al. "A user-friendly tool to transform large scale administrative data into wide table format using a mapreduce program with a pig latin based script"[J]. BMC medical informatics and decision making, 2012.

[12] Taylor R C. "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics" [J]. BMC bioinformatics, 2010, 11.

[13] Schatz M C. "CloudBurst: highly sensitive read mapping with MapReduce"[J]. Bioinformatics, 25(11): 1363-1369, 2009.

[14] Cummings Jr D D. "All care health management system", U.S. Patent 5,301,105[P]. 1994-4-5.

[15] Li D, Park H W, Piao M, et al. "The Design and Partial Implementation of the Dementia-Aid Monitoring System Based on Sensor Network and Cloud Computing Platform" [M]//Applied Computing & Information Technology. Springer International Publishing, pp: 85-100, 2016.

[16] Welcome to Apache™ Hadoop, http://hadoop.apache.org/

[17] Sayantan Sur, Hao Wang, Jian Huang, Xiangyong Ouyang and Dhabaleswar K. Panda, "Can High-Performance Interconnects Benefit Hadoop Distributed File System? ", http://doczine.com/bigdata/2/1371884970_6c99485db9/sur-asvdc10.pdf

[18] I. Hwang, K. Jung, K. Im, and J. Lee, "Improving the Map/Reduce Model through Data Distribution and Task Progress Scheduling", Journal of the Korea Contents Association, vol.10, no.10, pp.78-85, 2010.

[19] Chang C D, Wang C C, Jiang B C. "Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors" [J]. Expert systems with applications, 38(5):5507-5513,2011.

Telemedicine Aware Video Coding Under Very-Low Bitrates

Zain Ul-Abdin[1,4], Muhammad Shafique[2], and Muhammad Abdul Qadir[3]
[1] TeleSehat Private Limited, Islamabad, Pakistan
[2] Karlsruhe Institute of Technology, Karlsruhe, Germany
[3] Capital University of Science & Technology, Islamabad, Pakistan
[4] Halmstad University, Halmstad, Sweden

E-mail: zain-ul-abdin@telesehat.com, muhammad.shafique@kit.edu, aqadir@cust.edu.pk

Abstract— **Implementing a real-time telemedicine solution for remote areas using available technologies is a challenge due to limited or non-availability of high-bandwidth mediums at these areas. In this paper, we present a telemedicine aware video coding system based on the standard High Efficiency Video Coding (HEVC) codec. The HEVC codec provides much better coding efficiency at the expense of increased computational complexity, which we have dealt with by incorporating adaptive interpolation and selective quality enhancement techniques to achieve real-time performance. With our proposed telemedicine customized video coding system, we have been able to demonstrate sufficient diagnostic quality for two sets of diagnostic video sequences encoded at much lower bitrates i.e., 100 kb/sec.**

Keywords; video coding; telemedicine; medical diagnostic videos; HEVC; H.264

I. INTRODUCTION

Telemedicine has been recognized as a mean to bridge the healthcare services delivery to the remote or rural areas. Video coding has been an integral part of any real-time telemedicine system that has been used to compress a diagnostic video stream and to deliver it to physician. Since standard off-the-shelf video coding systems treat all video regions equally and the texture/motion properties of medical/diagnostics videos are very different from that of the traditional videos, there is a need of a telemedicine customized video conferencing system, such that the regions under investigation (RUI) in the diagnostics videos can be encoded with high video quality. This challenge aggravates when considering the low available communication bandwidths (i.e., restricted communication facilities) in remote areas.

We propose a novel telemedicine customized video conferencing system, which encoded the regions under investigation (RUI) and diagnostics videos with relatively high video quality. Though applicable to a wide-range of communication channels with different bandwidths, our proposed system is especially beneficial for low available communication bandwidths (i.e., restricted communication facilities) in remote areas. To enable this, we integrate an important technique in our system, i.e., selectively enhancing the video quality of RUIs at the expense of video quality of other non-interested regions.

A telemedicine-aware video coding system is based on a standard video codec specification. The currently widely used video codec in the market is H.264/AVC (Advanced Video Coding) [1], which was developed roughly a decade ago in order to achieve a bit rate reduction of 50% as compared to MPEG-2 with similar subjective visual quality [2]. To achieve this, H.264 introduced a complex prediction, variable-sized block Motion Estimation (ME) and Rate Distortion Optimized Mode Decision (RDO-MD) algorithms. However, these algorithms add to the computational complexity of video coding in the order of ~10x relative to MPEG-4 advance simple profile encoding and ~2x for decoding [3]. To reduce this complexity, a large body of research explored fast and efficient algorithms/hardware accelerators for the different components of the H.264 codec, e.g., fast RDO-MD [1], fast ME [1], fast intra prediction [1] etc.

However, in 2013, a joint venture of MPEG and ITU-T standardization bodies called the joint collaborative team on video coding (JCT-VC), have released the next-generation video coding standards, i.e. High Efficiency Video Coding (HEVC) [4][5]. The HEVC provides 2x higher compression efficiency compared to H.264/AVC while providing the same video quality at the cost of significant computational complexity [6]. Since we aim at providing selective video quality enhancement for diagnostics videos, the adaptation of the new HEVC standard is considered as a potential mean for achieving high video quality for diagnostic videos.

The rest of the paper is organized as follows: Section II reviews the literature in customized video coding for telemedicine. Section III provides the telemedicine application scenario and the requirements from diagnostic video perspective are outlined in Section IV. Section V describes the proposed system architecture for the telemedicine customized video coding. The proposed system is evaluated in Section VI, and the paper is concluded with some remarks in Section VII.

II. Related work

There have been a number of interesting approaches for improving the diagnostic quality of the compressed video. These compressed videos can then be used for improving healthcare delivery ranging from emergency services to remote monitoring of patients. Most of these techniques attempt to achieve enhanced diagnostic quality at the expense of higher bitrate usage and the ones focusing on lower bitrate tend to require quite high computational resources. Our approach aims to achieve sufficient diagnostic quality at lower bitrates and at reduced computational complexity so that it could be executed on general-purpose computers.

A. Telemedicine driven Video Coding

In the following, we provide a detailed literature review of different state-of-the-art techniques for telemedicine-driven video coding and selective quality enhancement in advanced video codecs.

Panayides et al. [7][8] proposed a video communication framework to transmit of H.264/AVC medical ultrasound video over mobile WiMAX networks. The proposed method utilizes the clinical criteria to highlight region-of-interest (ROI) in the ultrasound video sequences in the form of video slices that are based on adjusting the values of the quantization parameter (QP). Thus, significant bitrate reduction is achieved by compressing diagnostically unimportant regions. On the other hand Chinnusamy et al. [9] described a method that includes encoding of high-resolution clinical video at the clinical acquisition resolution, where the medically significant regions are coded with better quantization levels and then decoded with low delay.

Zoha [10] examined the applicability of H.264/AVC in telemedicine reference model using the Digitized Medical Information Interface (DM) standardization. It was concluded from the tests performed using CT scan and Echocardiography video sequences, that H.264/AVC provides much higher PSNR and is therefore suitable to be used in standardization of the DM interface for the telemedicine reference model. Yu et al. [11] studied the H.264/AVC for the compression of medical videos and 3-D medical data set. A new motion complexity (MC) measure is proposed that forms the basis for the new rate control scheme for H.264. The experiments using CT scan and Echocardiography video sequences reveal that the proposed rate control scheme can achieve better perceptual video quality, with an average PSNR gain of up to 0.19 dB.

A framework for efficient encoding and transmission, of atherosclerotic plaque ultrasound video sequence has been proposed by Panayides et al. [12]. The approach considers a spatially varying encoding scheme, where an automated segmentation algorithm is used to identify diagnostic ROIs. Extensive simulations incorporating the proposed encoding methods and using different quantization parameters, when transmitted over 3G (and beyond) wireless networks led to very good mean opinion scores (MOSs) that were computed using two medical experts.

Similarly, Rao and Jayant [13] presented optimized algorithms for region-of-interest (ROI) coding, that's could be applied to mobile telehealth. They focused on rate control method that computes bit allocation on individual regions within the frame rather than the frame level. The experimental results of allowing medical experts to select extended region of interest (EROI) proved that the proposed method performs better than VM8 of MPEG4 for the pediatrics video sequences.

To summarize, most of the above-mentioned techniques attempt to achieve enhanced diagnostic quality at the expense of higher bitrate usage and the ones focusing on lower bitrate tend to require quite high computational resources. Standard regions of interest (ROI) based techniques typically prioritize moving blocks considering the human visual system properties. On the contrary, in our case, human visual system properties are not the focus for quality enhancement, rather these are patient body parts under medical investigation (BPMI) or regions under investigation (RUI) along with different types of diagnostic videos. Similarly, the state-of-the-art ROI based techniques either use compute-intensive 3D-morphological operation in segmentation or they use old generation of video codec standards that offer limited potential for optimizations and quality enhancements. Furthermore, previous techniques primarily use only one control knob (e.g. allocated bit rate) for the ROI.

Our main goal is to achieve sufficient diagnostic quality at lower bitrates and at reduced computational complexity so that it could be executed in real-time on general-purpose computers. This could be achieved by employing an adaptive scheme that reacts to the changing bandwidth scenarios, while ensuring a certain quality of service for the diagnostic videos.

III. Telemedicine Application Scenario

Consider a scenario where a physician is remotely examining a patient with a wound, situated in some remote location. Two specific regions of interest are marked in the diagnostic video that the physician wants to observe at a higher quality than the other non-important regions in the video. But the commonly used video coding techniques employed in telemedicine systems will treat the complete video frame uniformly which will not give the desired video quality at the given bandwidth available at such remote rural center. A consequence of this could be that the physician will not be able to reach the correct diagnosis. The situation gets exacerbated in other cases of medical imaging such as Iris examination, ultrasound imaging, and radiology scans.

A possible solution to the above-illustrated problem is that we customize the video coding of diagnostic video

stream based on the feedback of the physician such that the marked region of interest is encoded differently with respect to the background, to produce better quality for the marked region.

Realizing such a telemedicine aware video conferencing system for such an application using the available technologies requires (1) automatic detection of regions under investigation (RUIs); (2) allowing doctors to specify regions under investigations and transferring this information to the remote patient end; (3) selectively enhancing the video quality of these regions under investigations in order to facilitate doctors to perform a better and high-quality medical investigation; and (4) selective quality enhancement of RUIs under low-bandwidth and low-computational scenarios.

IV. Requirements on Diagnostic Imaging

The most fundamental requirement to deliver the quality telemedicine services that meets the acceptance of the physicians and patients is to support real-time telemedicine mode of delivery using more affordable mediums of communication. For medical disciplines such as cardiology and general medicine, conducting off-line video conferencing sessions can fulfill the acceptance criterion of the physician. But, in order to meet the diagnostic requirements of dermatology, ophthalmology, radiology, ENT, and gynecology, the diagnostic imaging sessions need to be conducted in real-time scenario with sufficient imaging quality.

For instance, let us consider the case of a dermatology examination, the physician should be able to get the real-time video of skin exam camera from the remote patient where the physician can mark the region of interest on the video frames so that the marked region is communicated at higher quality. Thus, in a real-time examination session the physician can ask the paramedic at the patient-end to focus on a particular body part with more precision rather than the store and forward mode where the physician is totally dependent on the expertise of the paramedic to provide the diagnostic images.

A. Video Quality Constraints

Subjective and Objective Video Quality in terms of PSNR: The important regions in the ultrasound need to be preserved along with the motion. For instance, we will start with the ultrasound of womb. Moreover, the important regions (RUIs / BPMIs) need to be preserved. Doctor may provide feedback to further increase the video quality. A good quality range for such parts is between 30-40 dB. In order to meet the bandwidth constraints, some regions will suffer from video quality loss such that RUIs can be enhanced. Still to provide sufficient quality for such parts, the PSNR should preferably be more than 20 dB and may be below 30 dB. The quality constraints will be subjectively evaluated by the Doctors and feedback will be provided to further refine the objective quality constraints (i.e. minimum PSNR values for different RUIs). Therefore, for some diagnostics videos, fixed constraints on quality may be imposed at runtime. Such constraints can be imposed by the doctor by marking a particular region on the diagnostics video.

Frame Rate: For normal videos, the frame rate would be 15fps. The ultrasound videos can be sent from 10 fps to 15 fps. The frame rate for critical diagnostic video stream (like wound, skin, eyes, and X-Rays) range from 1 fps to 5 fps.

B. Communication Constraints

In order to understand the communication bandwidth constraints let us consider two video sequences taken on two different communication mediums: 1) DSL supporting a sustained bandwidth of greater than 300kbps, 2) EvDO supporting a sustained bandwidth of less than 150kbps, as illustrated in Figure 1a and Figure 1b, respectively.

Now in an ordinary telemedicine session involving diagnostic imaging there are a minimum of two video streams being played concurrently. One of them being the regular video conferencing stream, whereas the other being the diagnostic imaging stream, which means that under such scenario the maximum bit budget available for diagnostic video is not more than 50% of the total available bandwidth, keeping in view overhead of transport protocols, packetization, retransmission of lost packets.

The degradation of the Iris scan over low bandwidth medium (EvDO) is quite obvious from Figure 1a and Figure 1b, since the complete video frame is being encoded uniformly, which highlights the significance of the RUI based encoding approach to improve the quality of Iris region. Thus, our aim is to support the complete telemedicine session including the conferencing video and diagnostic video streams to a sustained bandwidth of as low as 100kbps.

V. System Architecture Design

The system design and architecture of our telemedicine customized video conferencing system is shown in Figure 2. The system processes two video streams concurrently. The first video stream is of ordinary video conferencing between the patient and the physician, while the second video stream is for high-quality medical diagnostic videos, as shown in Figure 2. The bandwidth is budgeted among these two video streams according to the priority, i.e. the diagnostic videos gets relatively more budget. Depending upon doctor's feedback, for further quality enhancement, more bit budget can be allocated to the diagnostic video stream.

(a) (b)

Figure 1. (a) Iris Scan using DSL medium, (b) Iris Scan using EvDO medium.

Moreover, within each stream, we also target selective quality enhancement for improved medical diagnosis, where parts of patient and diagnostic videos are selectively encoded in high quality. The medical video analysis is performed in the **pre-processing stage** and based on its output diagnostic-aware encoding mode, the configuration selection unit sets initialization parameters for the **telemedicine-driven video coding.** Afterwards, packetization is performed and bit stream packets are sent over the network to the Hospital Site for diagnosis and medical examination. The video decoder decodes the bit stream packets and generates the high-quality diagnostics videos along with the patient video.

In the following, we present the details on constituting novel sub-systems with respect to the design of our system.

A. Pre-Processing for Telemedicine Driven Video Coding

In order to enable selective quality enhancement during the encoding process or after the decoding process (i.e. post-processing), we employ novel techniques for statistical analysis of medical videos, customized video coding, and selective non-linear filtering. For effective employment of such techniques, there is a need for appropriate pre-processing to provide an image on which selective quality enhancement algorithms can be applied and even optimized for. The input video streams are passed through the pre-processing stage, where the videos may be filtered for noise reduction, subsampling for low bit rate video coding scenarios, and color/brightness/contrast adjustment for improved RUI detection. The pre-processing stage also performs statistical analysis (spatial or frequency domain analysis) of the texture and motion properties of the diagnostics videos and RUIs, which guides the refinement and optimization of models for block categorization such that important frequency components or video regions of diagnostic videos and RUIs can be automatically identified/predicted at run time. The pre-processing stage extracts the medical video properties. It is equipped with models defining the relationship between these medical video properties and optimal coding configuration. The medical video properties are used for automatically detecting the regions under medical investigation that require high video quality, while the models are forwarded to the telemedicine customized video encoder.

B. Telemedicine Driven Video Coding

Figure 3 illustrates the block diagram of our novel telemedicine-driven video coding sub-system. The key novelty is to leverage telemedicine specific features and statistical properties of medical videos to provide selective high quality to important diagnostic videos and video regions with BPMIs, while tolerating video quality loss in less-important regions (like background, not-selected body parts, etc.).

Our telemedicine driven video codec incorporates several adaptive algorithms for selective video quality.

Models for block categorization: For efficient categorization, a statistical analysis of medical (diagnostic) videos is performed. For example, such a statistical analysis would require spatial or frequency domain analysis of the texture and motion properties of the diagnostic videos (like ultrasound videos, X-rays, etc.) and RUIs (like wounded hand, legs, patient face, etc.). This statistical analysis provides a foundation to devise models for block categorization such that important frequency components or video regions of diagnostic videos and RUIs can be automatically identified/predicted at run time. This block categorization with respect to statistical properties of medical videos is leveraged by the adaptive algorithms in the telemedicine-driven video encoder for selective video quality. The block characterization functions are derived to identify regions of different statistical properties, while considering medical video regions of less and more-importance such that selective video quality enhancement techniques can be applied.

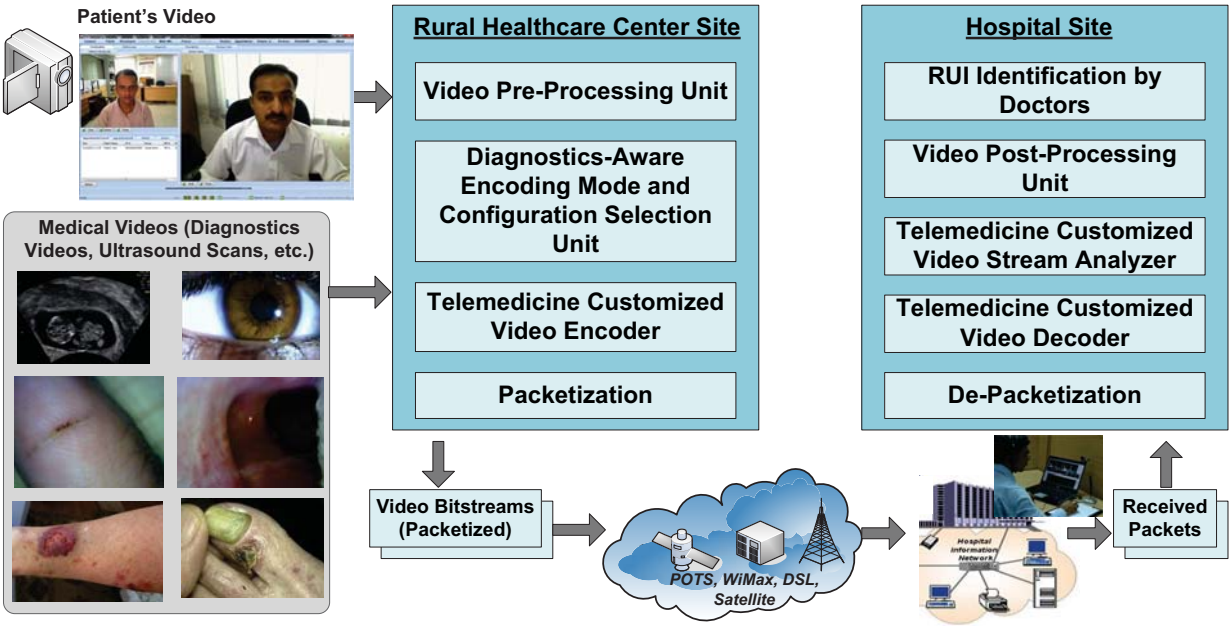


Figure 2. System Architecture of Our Telemedicine-Customized Video Conferencing System.

A relation between the statistical properties of diagnostic videos and (optimal) coding configuration and prediction mode is modeled and established. Modeling the relationship of diagnostic videos/RUIs to different coding configurations consider spatial and temporal correlation in the medical videos. The design space of different coding configurations is explored to formulate relationships between different block categories and optimal coding configuration to provide selective video quality enhancement. These models and transfer functions are used by the following sub-system components.

Bit-allocation and Rate Control Scheme: Rate control scheme is devised that performs a non-linear bit-budgeting given a low-bandwidth channel and allocates more bits to more-important block categories (i.e. selectively enhancing the video quality of diagnostic videos and regions with BMPI), while giving less bits to less-important block categories (background regions). This is different from the constant or variable bitrate controllers, as in this case, the bit budget is adaptively allocated based on the medical video properties. The algorithms for bit budgeting and bit allocation account for the above-discussed statistical analysis and block categories with respect to RUIs and diagnostics videos. The quality of RUIs and diagnostic videos is selectively enhanced by allocating more bits to them, while giving less bits to less-important block categories (e.g. background regions). For this, non-linear bit budgeting and bit-allocation for different block categories is performed. Afterwards, the control operates on the different allocated budgets on selected and non-selected regions (RUIs). Based on this information, an appropriate Quantization Parameter (QP) is selected, which is then forwarded to the telemedicine specific coding configuration selection module and the telemedicine driven adaptive motion estimation module.

Coding Configuration Selection Scheme: Our coding configuration selection algorithm adaptively explores the search space of various coding configurations (coding units/modes, block sizes, prediction units/modes, transform unit, picture prediction structures, etc.) while pruning the non-optimal configurations considering the medical video properties. For a given bit-budget, our system jointly controls these parameters is a complex multi-variable optimization problem. The key is to leverage the statistical analysis and block categories with respect to RUIs and diagnostic videos. The video quality of important blocks for RUI and diagnostic videos is selectively enhanced by assigning high quality modes (for instance, intra prediction modes). This may compromise a slight video quality loss in less important blocks. Our system employs aggressive pruning strategies and configuration exclusion algorithms for less-important blocks, such that high-quality configuration modes are allocated to more-important blocks to provide higher video quality.

Motion Estimation Scheme: Typically, the video quality is a function of prediction error. A higher prediction error corresponds to a higher residual energy that needs to be encoded by the video encoder. Therefore, a better prediction leads to a relatively high video quality. Motion estimation is one of the key functional blocks that perform prediction in the neighboring video frames to exploit the temporal correlation.

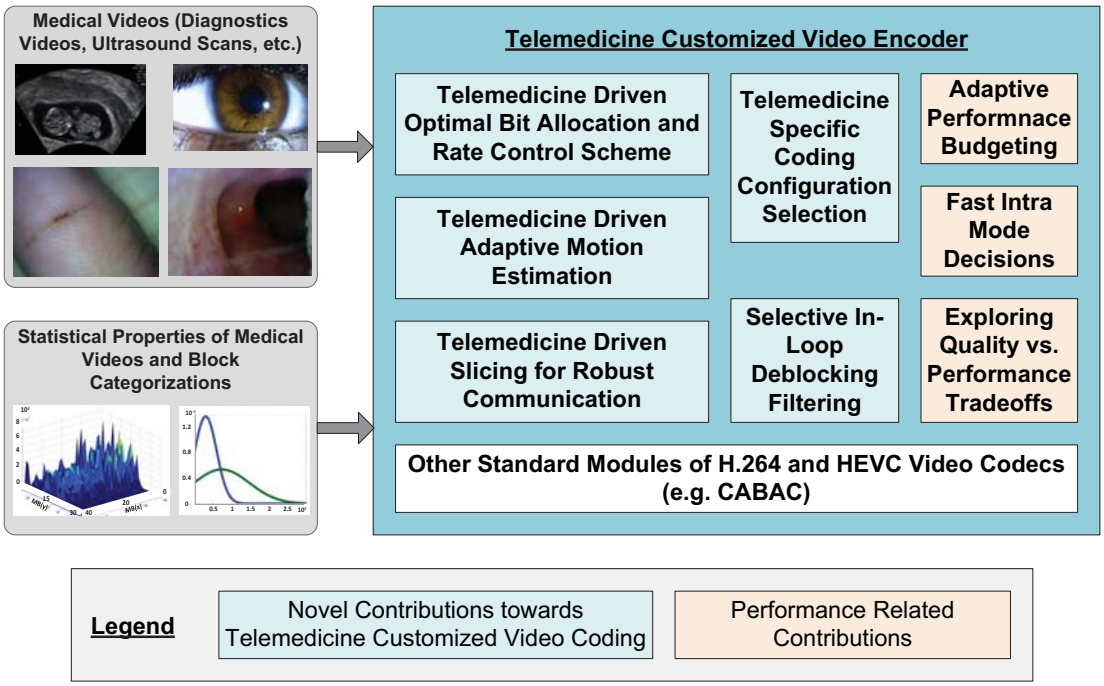


Figure 3. System Architecture of Our Telemedicine-Customized Video Encoding Sub-System.

An extensive motion search provides a better prediction, thus improved video quality at the cost of extreme computation requirements, while a limited motion search requires fewer computations but suffers from video quality loss. The motion search effort highly depends upon the texture properties and shape of blocks and the texture and motion properties in the spatial and temporal neighborhood. Therefore, a higher video quality for the diagnostic videos and RUIs is obtained by performing an extensive motion search, while compromising the motion search effort for the less-important blocks. The motion search effort is controlled using various parameters, for instance, (1) search pattern, (2) search range or search window configuration (size, shape), (3) early termination conditions, etc.

Our system employs a novel telemedicine driven adaptive motion estimation algorithm that exploits the medical image properties and block categorizes (as discussed above) in order to provide selective video quality enhancement. Our system implements an algorithm to control different knobs of the motion estimation (search pattern, search window configuration, etc.). This algorithm uses different early termination conditions that for a given bit rate and given throughput constraint determine the motion search termination criteria while avoiding trapping into local minima. We use mechanisms to borrow the computational budget from the motion estimation of less-important block categories and give it to the motion estimation of more-important block categorizes. In this way the prediction quality of more-important block categorizes is significantly improved at the cost of a slight loss in the video quality loss of less-important blocks.

VI. System Evaluation

We have evaluation of our proposed system by conducting two different diagnostic video scenarios as follows:

A. Eye (Iris) Video Sequence

The bit rate allocated for the eye examination video sequence can be controlled by passing different quantization parameter (QP) values to the customized video encoding system which is enabled by the enhanced rate controller integrated with the encoding system. We have performed the experiment, where we have set the value of QP to 32 and conducted the experiment under low bandwidth conditions corresponding to the bitrate of less than 100 kb/sec and observed the corresponding peak signal to noise ratio (PSNR) achieved by our video encoding system. It should be noted here that the same bandwidth is also being used for the conversation video sequence, which reduces the available bandwidth for the diagnostic session to one-half. It has been observed that we are able to achieve an improvement of 40.7 dB in the PSNR of the eye sequence and without adding a significant execution time for the encoding system. Thus, our customized video encoding system allows us to improve the quality of the encoded video stream according to the available bandwidth scenario, thereby making it possible to conduct the diagnostic session for Iris exam under low bitrate by encoding the region of interest of the resulting video stream at a very fine-grained level.

B. Skin Video Sequence

Next, we look at the skin examination video sequence. Again the bitrate can be controlled by passing different quantization parameter (QP) values to the customized video encoding system which is enabled by the enhanced rate controller integrated with the encoding system. We have performed the experiment where we have set the QP value to 26 and conducted the experiment under good bandwidth conditions corresponding to a bitrate of 300 kb/sec and observed the corresponding peak signal to noise ratio (PSNR) achieved by our video encoding system. In the case of skin video sequence, we were able to achieve almost 44 dB PSNR by doubling the bitrate, by adding almost 10% overhead of computing the resulting video stream. The major reason for the increased improvement in the case of skin video sequence as compared to the eye video sequence is the higher resolution of input video sequence, i.e., 640x360 for skin as compared to 448x352 for the eye. The texture properties of the skin sequence also allow the video encoding system to accurately capture texture direction. Thus, our customized video encoding system takes different (potentially contrary) coding decisions, i.e., encoding the *wound* related CUs in high quality using small partitions with angular intra modes, such that each specific texture pattern is precisely predicted and the textures are preserved for high quality diagnosis, thereby allowing us to improve the quality of the encoded video stream according to the available bandwidth scenario.

VII. Summary and Future work

To summarize implementing a real-time telemedicine solution for remote areas using available technologies is a challenge due to limited or non-availability of high bandwidth mediums at these sites. In this paper, we propose to implement a telemedicine aware video coding system based on the standard HEVC codec to leverage the coding efficiency of the HEVC codec. We have incorporated content-driven complexity reduction scheme and used adaptive interpolation filters that adapts to the workload derived from the video sequence properties. With our proposed telemedicine customized video coding system, we have been able to demonstrate sufficient diagnostic quality for two sets of diagnostic video scans encoded at much lower bitrates i.e., 100 kb/sec. However, lowering the bitrate further leads to reducing the diagnostic capabilities depending on the type of diagnostic video, which means there are still ample opportunities for improving the video compression quality by performing selective encoding of different blocks of a video frame.

Acknowledgment

We are grateful to our project team members who enthusiastically have contributed to the work. The research is funded by the National ICT R&D fund, Pakistan.

References

[1] ITU-T Rec. H.264 and ISO/IEC 14496-10:2005 (E) (MPEG-4 AVC), "Advanced video coding for generic audiovisual services", 2005.

[2] T. Wiegand, G. J. Sullivan, G. Bjntegaard, and A. Luthra, "Overview of the h.264/avc video coding standard", *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, Vol. 13, no. 7, pp. 560–576, 2003.

[3] J. Ostermann, J. Bormans, P. List, D. Marpe, M. Narroschke, F. Pereira, T. Stockhammer, and T. Wedi, "Video coding with H.264/AVC: Tools, performance, and complexity", *IEEE Circuits and Systems Magazine*, Vol. 4, no. 1, pp. 7–28, 2004.

[4] ITU-T, "SERIES H: Audiovisual and Multimedia Systems - Infrastructure of Audiovisual Services - Coding of Moving Video – High Efficiency Video Coding." April 2013.

[5] G. J. Sullivan, J. Ohm, W. Han, T. Wiegand, "Overview of the High Efficiency Video Coding," in *IEEE Transactions on Circuits in System Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[6] B. M. T. Pourazad, C. Doutre, M. Azimi, P. Nasiopoulos, "HEVC: The New Gold Standard for Video Compression: How Does HEVC Compare with H.264/AVC?," in *IEEE Consumer Electronics Magazine*, pp. 36–46, 2012.

[7] A. Panayides, Z. C. Antoniou, Y. Mylonas, Marios S. Pattichis, A. Pitsillides, and C. S. Pattichis, "High-Resolution, Low-Delay, and Error-Resilient Medical Ultrasound Video Communication Using H.264/AVC Over Mobile WiMAX Networks", *IEEE Journal of Biomedical and Health Informatics*, Vol. 17, No. 3, pp. 619-628, May 2013.

[8] A. Panayides, Marios S. Pattichis, and Constantinos S. Pattichis, " Mobile-Health Systems Use Diagnostically Driven Medical Video Technologies ", *IEEE Signal Processing Magazine*, pp. 163-172, November 2013.

[9] K. Chinnusamy and K. AbinayaPreethi, " Performance and Analysis of Video Streaming of Signals in Wireless Network Transmission ", *IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)* Vol. 9, Issue 1, PP 11-15, January 2014.

[10] M. U. Zoha, "Applicability of H.264 in Telemedicine Reference Model", *JCIT*, Vol. 1, 2010.

[11] H. Yu, Z. Lin, and F. Pan, "Applications and Improvement of H.264 in Medical Video Compression ", *IEEE Transaction on Circuits and System*, Vol. 52, No. 12, pp.2707-2716, December 2005.

[12] A. Panayides, M. S. Pattichis, Constantinos S. Pattichis, C. P. Loizou, M. Pantziaris, and Andreas Pitsillides, "Atherosclerotic Plaque Ultrasound Video Encoding, Wireless Transmission, and Quality Assessment Using H.264", *IEEE Transactions on Information Technology in Biomedicine*, Vol. 15(3), May 2011.

[13] S. Rao and N. Jayant, "Optimizing Algorithms for Region of Interest Video Compression, with Application to Mobile Telehealth ", Invited paper: *Special Session on Perceptual Visual Processing, ICME,* 2006.

A Bond Graph Approach for Modeling the Client-Therapist Relation

Abdelrhman Mahamadi and Shivakumar Sastry
Department of Electrical and Computer Engineering
The University of Akron, Akron, OH 44325 USA

Abstract*—Modeling and reasoning about human behavior is an interesting problem for a variety of emerging applications across many domains. Our particular interest is related to wellness management; specifically, we are interested to understand how to stimulate positive behaviors in individuals that lead to better management of their own wellness. The ultimate objective is to keep healthy people healthy. In this short paper, we propose a bond graph approach for modeling human behavior. We consider the interaction between a client and a therapist as an illustrative example to demonstrate the approach. We present a bond graph that represents this interaction and derive the dynamic equations for the system. Finally we illustrate how one can regulate the client-therapist interaction by applying well-known concepts from control theory.*

***Index Terms*—Bond Graph and Human behavior.**

1 Introduction

Modeling and reasoning about human behavior is an important problem in several domains. Several recent advances have been made in persuasion technologies. Our particular interest is to understand human behaviors in the context of personal wellness management [1], [2]. By developing such models, we aim to use these models to derive personalized interventions to improve the nutrition and exercise behaviors of the participants.

Several theories have been developed over the last few decades to explain human behavior. Notable among these are *The Theory of Planned Behavior* [3], *Social Cognitive Theory* [4], *Self-Determination Theory* [5] and the *Transtheoretical Model for Stages of Change* [6]. There are several other theories that are specialized to different domains such as the *Health Belief Model* [7]. While such models can explain aggregate behaviors at the scale of a population, these models are not actionable at the level of individuals.

Recent efforts have resulted in fluid analogy models for human behavior that aim to operationalize the above theories in a control systems framework [8], [9], [10], [11], [12]. These models provide an intuitive and easy approach to separate the state variables and system parameters that drive the models. Such models have been effectively used in socially relevant programs for smoking cessation and health management [13]. Despite their simplicity and effectiveness, there can be ambiguities in these models that limit their full exploitation in automated tools. For this reason, we are examining the utility of domain independent models that can be used to represent and reason about human behaviors.

Bond graphs were introduced in [14] as a domain independent graphical representation to reason about systems involving mechanical, chemical and electrical components in a unified framework. In this approach, a system is viewed as comprising several components; each component has ports through which energy can be exchanged with other components. Every component is identified as being one that generates energy in the system or one that consumes energy. Components are connected through bonds between corresponding ports. Every bond has a half arrow that denotes which element in the bidirectional relationship generates energy and which element consumes energy. Energy transfer between components is viewed as a bidirectional exchange of *effort* and *flow* [15], [16], [17].

To illustrate the utility of bond graphs for human behavior modeling, we present a model for the interaction between a client and a therapist that is inspired by the recent work reported in [18]. Here, the authors formulate a dynamic systems model for the interaction based on empirical observations in the state of the practice. In contrast, we show that the dynamic systems model can be derived from a bond graph model. The bond graph model is itself derived from a fluid analogy model that we constructed to capture the interactions between the client and the therapist. We demonstrate the value of causality analysis in our approach. In addition, we demonstrate that our approach also leads us to a similar conclusion about the stable states of the system as reported in [18]. Further, we show that our approach is extensible and we can incorporate the domain knowledge and parameters that were discovered in the psychology models for human behavior [3], [4], [6].

2 Background

Table I shows the commonly used elements in a bond graph. Components of the system must be mapped to one of the modeling elements shown in Figure I. Bond graph modeling approaches are well developed [19]. The topology of the system that is being modeled guides the construction of a bond graph model. The process starts by first considering qualitative relationships in the system and progresses toward more concrete details as the modeling process evolves. In the literature, bond graphs have been used extensively to model systems involving electrical, mechanical and chemical energy transfers [19].

TABLE I: Commonly used Bond Graph Elements

Element	Description	Notation
S_e S_f	Sources.	S_e S_f
R	Resistor.	R R
C	Storage element for a effort-type variable.	C
I	Storage element for a flow-type variable.	I
TF	Transformer.	n TF n TF
GY	Gyrator.	r GY r GY
0-Junction 1-Junction	0– and 1–junctions, for ideal connecting two or more sub-models.	1 0

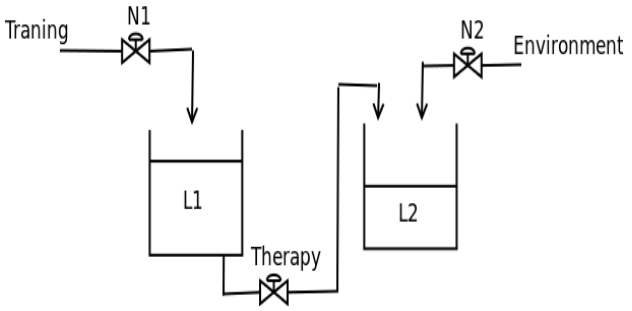


Fig. 1: Fluid Analogy representation for the interaction between a Therapist and a Client.

TABLE II: Physical Components of the System and Corresponding Bond Graph Elements

Physical Element	Type	Value
Training source	Se	S_1
Environment source	Se	S_2
Training valve	TF	N_1
Environment valve	TF	N_2
Therapist valence	I	L_1
Client valence	I	L_2
Therapy	R	R

3 Bond Graph Model for the Client-Therapist Relation

We view human behavior as a consequence of complex energy transfers across multiple domains. Our interest is to model the behavior in order to improve decision-support [1], [2] and hence, we aim to develop actionable models for human behavior. In this section, we introduce a model, that describes the interaction between a Client and a Therapist that is inspired by the work in [18].

As in our earlier work [2], we start by the fluid analogy representation for the interaction between a client and a therapist as in Figure 1. Here, there are two fluid storage tanks — the client is represented by the storage tank on the right, while the other tank representing the therapist. Following the ideas in [18], the level of fluid in each of these storage tanks corresponds to the *valence*, or *affect*, of the client (L_2) and the therapist (L_1), respectively. The valence of the therapist is a function of his or her training (S_1) and is regulated by the valve N_1. We assume that a better trained therapist, i.e., more flow in N_1, would have higher valence. The valence of the client is affected by the environmental conditions (S_2) and regulated by N_2. Through the therapy, the valence of the client and the therapist are changed because of the flows through the valve that is labeled Therapy (R).

As a first step in creating a bond graph for the subjective model, we map the physical components of the client-therapist interaction system shown in Figure 1 to the bond graph elements shown in Figure I as shown in Table II.

Using the Bond Graph construction procedure outlined in [2], we represent every distinct effort point in the system by a $0-$Junction, then we connect the physical elements between these distinct effort points using $1-$Junctions and half arrows. Finally we simplify the bond graph and we carry out the causality analysis algorithm on this model to obtain the bond graph model that is shown in Figure 2.

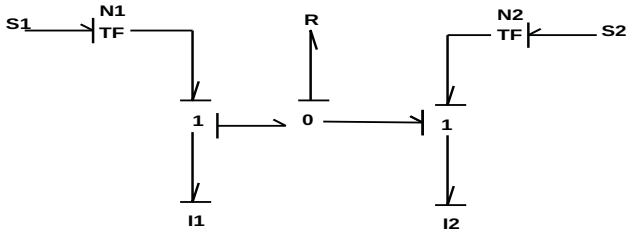


Fig. 2: Bond Graph representation for the Therapist-Client model

We then systematically derive the dynamics of the behavior of the system using the procedure outlined in [15] which yields the differential equations that describe the dynamics. After simplification, we obtain

$$\frac{d}{dt}(L_1) = N_1 \times S_1 - R \times (L_1 - L_2), \quad (1)$$

and

$$\frac{d}{dt}(L_2) = N_2 \times S_2 + R \times (L_1 - L_2). \quad (2)$$

Equation 1 shows how the valence of the therapist changes over time and Equation 2 captures the valence of the client.

Since these are the only state variables in the system, these two equations collectively describe the behavior of the client-therapist interaction system.

4 Regulating the Dynamic Behavior of the Client-Therapist Relation

The bond graph model allowed us to systematically obtain a set of dynamic equations for a subjective system. Now we can use this dynamic model in Equation 1 and Equation 2 to regulate how the client-therapist relation evolves.

To illustrate this idea, we used the models to regulate the client valence through therapy using three different control schemes and we recorded the client response under each scheme. To maintain the Client valence, we assumed that there is an optimal level, L_c, that should be maintained. A well trained therapist should be able to control the ("flow of") therapy (offered) to the client and the ("flow of") feedback from the client. To model this interaction, we designed a controllers to regulate the valve R. The integrated system is represented as a multiple inputs and multiple outputs (MIMO) system. First, we need to obtain the transfer function of the system from the dynamic equations derived using the Bond Graph. From Equation 1 and Equation 2 we get the expression for the Client's valance in terms of therapist training (S_1), training regulating valve (N_1), the environment input (S_2), environment valve (N_2) and the therapy regulating valve(R) as

$$L_2(s) = \begin{bmatrix} \frac{1}{s^2+(R^2+2R)s+(R^3+R^2)} \\ \frac{Rs}{s^2+(R^2+2R)s+(R^3+R^2)} \end{bmatrix} \times \begin{bmatrix} S_1N_1 & S_2N_2 \end{bmatrix} \quad (3)$$

From the Equation 3 we obtain the MIMO transfer function as

$$H(s) = \begin{bmatrix} \frac{1}{s^2+(R^2+2R)s+(R^3+R^2)} \\ \frac{Rs}{s^2+(R^2+2R)s+(R^3+R^2).} \end{bmatrix} \quad (4)$$

We can now explore a variety of controllers to regulate behaviors in this interaction.

4.1 Maintaining Client Valence using Proportional Controller

Under proportional Control, the controller transfer function is only one term of proportional ratio (K_p). We only focus on the first entry of the transfer function, since we can manipulate the training and the therapy valves in order to regulate the valance of the client. Environment context on the other hand is not easy to be manipulated. Next, to make the problem more realistic we choose values for the valves openings of the model, plugging these values in the transfer function we obtain a final transfer function in Equation 5.

$$H_1(s) = \frac{1}{6.25s^2 + 7.5s + 1} \quad (5)$$

After integrating the proportional controller (K_p) to the system, the response of the system to a unit set point is depicted in Figure 3.

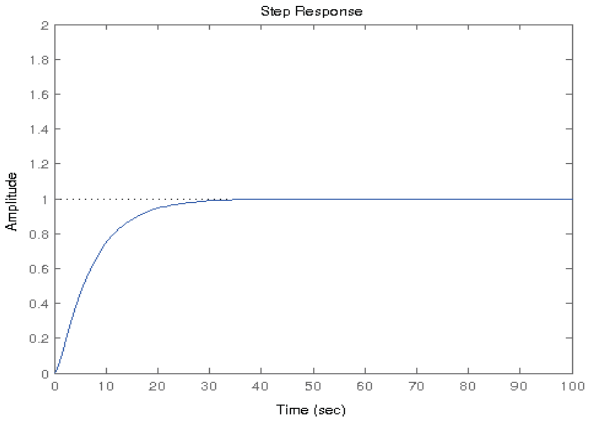


Fig. 3: The step response of the client valence under the proportional controller

4.2 Maintaining Client Valence using Proportional and Integral Controller

Because of the delay in the response of the proportional Control, we investigate adding an integral part to the controller. The new transfer function of the controller is now two terms ($K_p + \frac{K_i}{s}$). Again We only consider the transfer function between the client's valance level and the training and therapy valves, plugging same values as in the first case we obtain the transfer function as

$$H_1(s) = \frac{10s + 1}{6.25s^2 + 7.5s + 10s + 1} \quad (6)$$

After connecting the proportional integral controller ($K_p + \frac{K_i}{s}$) to the system, the response of the system to a unit set point is depicted in Figure 4.

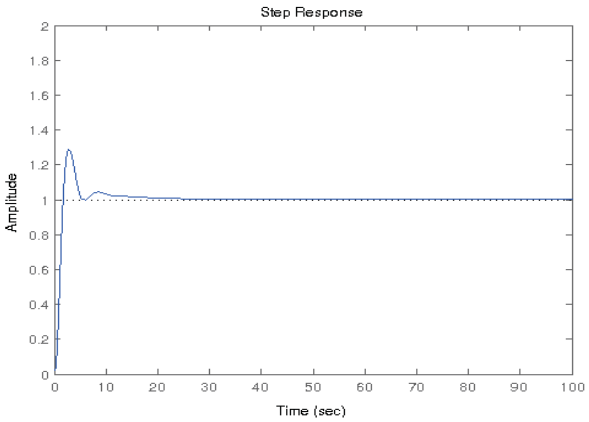


Fig. 4: The step response of the client valence under the proportional and integral controller

4.3 Maintaining Client Valence using Proportional, Integral and Derivative Controller

Adding the integral part to the controller makes the response faster and it causes an overshoot in the system response. We add a derivative part to the controller to limit the overshoot. The transfer function now becomes $K_p + \frac{K_i}{s} + K_d \times s$. Again, considering just the relation between the client's valence, and the training and therapy valves, using the same values as in the first case, we obtain the equivalent transfer function

$$H_1(s) = \frac{10s^2 + 10s + 1}{6.25s^2 + 17.5s + 10s + 1} \quad (7)$$

With this proportional, integral and derivative controller, the response of the system to a unit set point is shown in Figure 5.

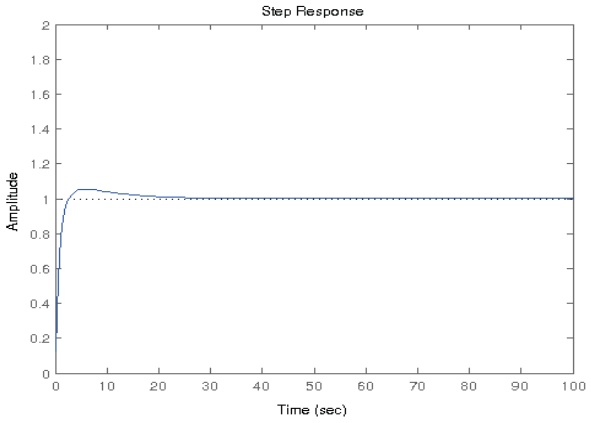


Fig. 5: The step response of the client valence under the proportional, integral and derivative controller

We can conclude from these results that the bond graph approach is useful for obtaining dynamical system model of the client-therapist interaction as illustrated in [18]. The controllers that were used to regulate the interactions offered a variety of choices. Among these controllers, the PID controller provides little delay and overshoot; it is however difficult to tune.

The next step for us is to extend the client side of the bond graph model to capture the parameters and environmental triggers in the various constructs for human behavior in the bond graph models. We believe that the bond graph approach is extensible and provides actionable models for human behaviors.

5 Conclusions

We presented a bond graph approach to modeling human behavior. Using a concrete setting of a client-therapist interaction, we presented a fluid analogy model for the interaction. Using this fluid analogy model as a starting point, we derived a bond graph model for the system. After causality analysis, we obtained the dynamic system model for the interaction using the bond graph model. We illustrated how the model could be used to analyze the behavior of the system.

In the future, this approach can be extended to incorporate the constructs of more comprehensive human behavior models that have been reported in the literature. The approach can also be extended to model collections of individuals to represent behaviors at scales larger than that of individual participants in wellness programs.

References

[1] M. k. Chippa, S. M. Whalen, F. L. Douglas, and S. Sastry, "Goal-seeking formulation for empowering personalized wellness management," in *Medical Cyber Physical Systems Workshop*, 2014.

[2] A. Mahamadi and S. Sastry, "Bond graphs models for human behavior." in *IEEE international conference for basic sciences and engineering*, 2016.

[3] I. Ajzen, "The theory of planned behavior," *Organizational Behavior and Human Decision Processes.*, vol. 50, pp. 179–211, 1991.

[4] A. Bandura, *Social foundations of thought and action: A social cognitive theory*, N. Englewood Cliffs, Ed. Prentice-Hall series in social learning theory., 1986.

[5] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being," *American Psychologist*, vol. 55, pp. 68–78, 2000.

[6] J. O. Prochaska, "Decision making in the transtheoretical model of behavior change," *Medical Decision Making: an International Journal of the Society for Medical Decision Making*, vol. 28, pp. 845–849, 2008.

[7] N. Janz and M. Becker, "The health belief model: A decade later," *Health Education Behavior*, vol. 11, pp. 1–47, 1984.

[8] J. E. Navarro-Barrientos, D. E. Rivera, and L. M. Collins, "A dynamical systems model for understanding behavioral interventions for weight loss." in *SBP*, 2010.

[9] C. A. Martin, D. E. Rivera, W. T. Riley, E. B. Hekler, M. P. Buman, M. A. Adams, and A. C. King, "A dynamical systems model of social cognitive theory," in *American Control Conference*, 2014.

[10] C. Martin, S. Deshpande, E. Hekler, and D. E. Rivera, "A system identification approach for improving behavioral interventions based on social cognitive theory," in *American Control Conference*, 2015.

[11] J. E. Navarro-Barrientos, D. E. Rivera, and L. M. Collins, "A dynamical model for describing behavioural interventions for weight loss and body composition change," *Mathematical and Computer Modelling of Dynamical Systems*, vol. 17, pp. 183–203, 2011.

[12] Y. Dong, D. Rivera, D. M. Thomas, J. E. Navarro-Barrientos, D. S. Downs, J. S. Savage, and L. M. Collins, "A dynamical systems model for improving gestational weight gain behavioral interventions," in *American Control Conference*, 2012.

[13] D. Lai, K. Cahill, Y. Qin, and J. L. Tang, "Motivational interviewing for smoking cessation," *Cochrane Database of Systematic Reviews*, vol. 1, p. CD006936, 2010.

[14] H. M. Paynter, *Analysis and design of engineering systems*. M.I.T. Press, Cambridge, 1961.

[15] P. Gawthrop, "Bond graphs: A representation for mechatronic systems," *Mechatronics*, vol. 1, pp. 127–156, 1991.

[16] J. F. Broenink, "Introduction to physical systems modeling with bond graphs," in *in the SiE whitebook on Simulation Methodologies*, 1999.

[17] P. C. Breedveld, "Modeling and simulation of dynamic systems using bond graphs," in *Control Systems, Robotics and Automation - Modeling and System Identification I.* EOLSS Publishers Co. Ltd./UNESCO, 2008.

[18] L. Liebovitch, P. Peluso, M. Norman, J. Su, and G. J.M., "Mathematical model of the dynamics of psychotherapy." *Cognitive Neurodynamics.*, vol. 3, pp. 265–275, 2011.

[19] W. Broutzky, "Bond graph modelling and simulation of multidisciplinary systems - an introduction," *Simulation Modelling Practice and Theory*, vol. 17, pp. 3–21, 2009.

A High-performance and Accurate Medicine Detection System

Atif Ullah Baig and Cao An Wang
Department of Computer Science, Memorial University, St. John's, NL, Canada

Abstract— *Wrong medication is one of the root causes of Adverse Drug Events (ADE). In this paper, we describe the design and implementation of a Medicine Detection System (MDS) that can be used in hospitals, clinics, and pharmaceutical companies. MDS first scans the barcode on medicine's packaging which is unique to the medicine for accurate detection. Then, MDS connects to Amazon cloud environment for retrieval of the desired data associated with this medicine. This allows medical person to decide whether or not the chosen medicine is a good fit for the patient, provided by this medicine's data along with the patient's medical history. It can be further automated by utilizing Natural Language Processing (NLP) in the system.*

Keywords: Medication Errors, Barcode Scanning, Cloud Storage Environment, Data Retrieval, Natural Language Processing

1. Introduction

Medication errors are the most frequent cause of adverse medical events [1]. According to the Institute of Medicine [2], each year more than a million injuries and almost 100,000 deaths are caused by medical errors. The standard procedure for nurses before using a medicine is to check the 'five rights': right patient, right medication, right dose, right route and right time [3]. However, only 34 percent of dispensing and 2 percent of administration errors are caught before the medicines reach the patient [4]. In order to avoid medication errors, MDS system enables its users to get medicine's information instantly with very high percentage accuracy. The field of clinical NLP is flourishing through the contributions of both NLP researchers and healthcare professionals who are keen on applying NLP methods for healthcare purposes [5]. Since MDS can identify and retrieve very accurate information on medicines, this information along with patient's medical history can be used to perform NLP based analysis to decide how much the selected medicine is suitable for the patient.

2. Relevant Work

There has been a lot of research on medicine identification problem using different techniques and approaches. In most of the studies, research is performed on innovative image processing techniques to identify the shape, color, and imprint of medicines. However, these image processing techniques are not high percentage accurate as indicated in many related research works.

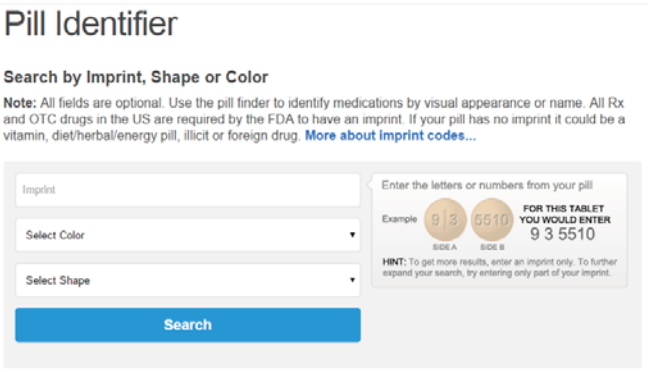


Fig. 1: An online medicine detection service provided by Drugs.com [6].

In 2011, an automatic system, called Pill-ID was developed to match drug pill images based on several features (i.e., imprint, color, and shape) of the tablet. The color and shape information was encoded as a three-dimensional histogram and invariant moments, respectively. The imprint on the pill was encoded as feature vectors derived from SIFT and MLBP descriptors. Experimental results using a database of drug pill images (1029 illicit drug pill images and 14,002 legal drug pill images) showed 73.04% rank-1 and 84.47% rank-20 retrieval accuracy [7].

In 2012, a modified shape distribution technique was used to examine the shape, color, and imprint of a pill and create an invariant descriptor that was used to recognize the same drug under different viewing conditions. The proposed technique was successfully evaluated with 568 of the most prescribed drugs in the United States with 91.13% accuracy in automatically identifying the correct medication [8].

In 2015, an imprint partition based strategy was used for automatic pill recognition. According to this strategy the imprints were partitioned on the basis of separated strokes, fragments and noise points. The results in this research demostrated 90.46% rank-1 matching accuracy and 97.16% on top five ranks when classifying 12,500 query pill images into 2500 categories [9].

In September 2003, Beloit Memorial Hospital installed a wireless, handheld barcode medication administration system in its Family Care Center (FCC) unit. The experimental results showed 67% decline in medication administration errors within the first four months of operation [10].

Different medicines can have same shape and color. Moreover, some medicines do not have any imprints on

them. Owing to these limitations, MDS system uses barcode technology to ensure very high percentage accuracy. A study conducted in 2010, found that barcode usage prevented about 90,000 serious medical errors each year and reduced mortality rate by 20% [11].

3. High level design of MDS

MDS conceptually comprises of three layers namely Data Transmission System (DTS), cloud storage environment and the mobile component. DTS is a web component of MDS. Both web and mobile components of MDS communicate with cloud storage environment to store and retrieve medicine's information. The mobile component of MDS first detects the medicine's barcode value which is unique for each medicine. The detected barcode value is then used to look up the medicine's information from the cloud storage. Mobile component of MDS is also capable of generating usage reports to keep track of the medicines that are most frequently used. Usage reports help ensure that these medicines are always in stock.

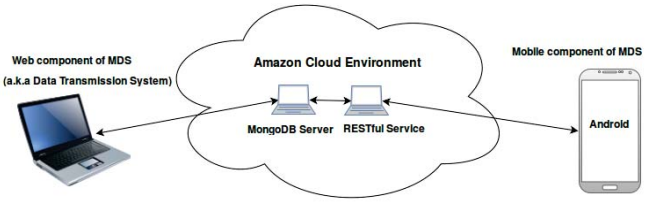


Fig. 2: High level architecture of MDS

4. Data Transmission System (DTS)

DTS is a web based application that stores medicines data into the cloud storage environment. It comprises of user interfaces that can perform CRUD operations on medicine's data inside the cloud storage. This component stores only the desired information of each medicine as it allows to add or delete fields for each medicine dynamically where each field contains specific information about the medicine. However, it is noticeable that barcode and name fields are required and can not be deleted. Each medicine can have any number of fields which can be different for different medicines as shown in figure 3.

Fig. 3: Dynamically adding or deleting medicine's fields

Following is the JSON format that is used by MDS system to dynamically add or delete medicine's fields from the database.

```
{
    "_id": { "$oid":"55622196edb013719069ee37"},
    "barcode":"91397798657",
    "name":"Propanol",
    "optionalDynamicAttributel":"attributelValue",
    "optionalDynamicAttribute2":"attribute2Value",
    "optionalDynamicAttribute3":"attribute3Value",
            .                                .
            .                                .
    "optionalDynamicAttributeN":"attributeNValue"
}
```

Please note that mock data is used in the above JSON format. The value of N can either be zero or any positive number.

5. Cloud Storage Environment

In order to permanently store medicine's data and access it from the mobile component of MDS it is important to use a storage facility that is accessible from any mobile device on which MDS is installed. For this reason, cloud storage is an ideal candidate to ensure data availability from all the running instances of mobile component of MDS. This system uses mLab platform which provides MongoDB-as-a-Service [12]. MDS uses MongoDB database deployed inside Amazon cloud environment for the storage of medicine's information. An important feature of mLab is the availability of RESTful API for the access and manipulation of data stored on cloud. This RESTful service is helpful to perform data retrieval operations from the Android mobile via HTTP which cannot be performed directly using standard MongoDB drivers. The first screen in figure 4 shows the

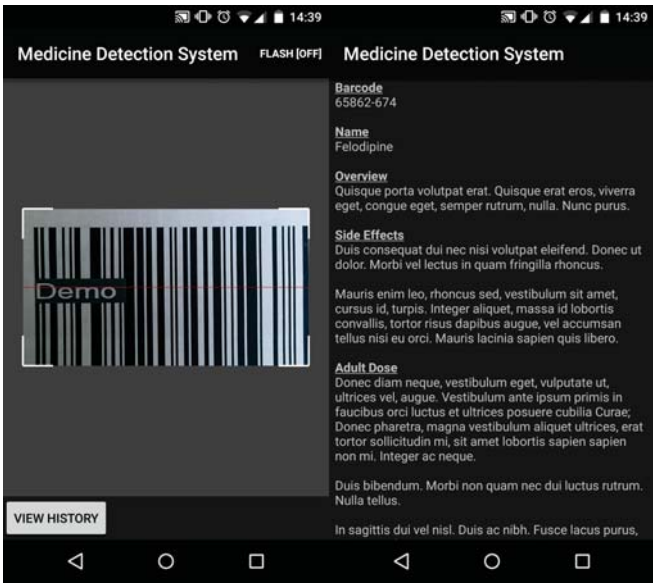


Fig. 4: Android screenshots for information retrieval from cloud

barcode scanning component of MDS that is developed using ZBar library [13]. The second screen in figure 4 shows the medicine's data that is returned by RESTful API for the given barcode value.

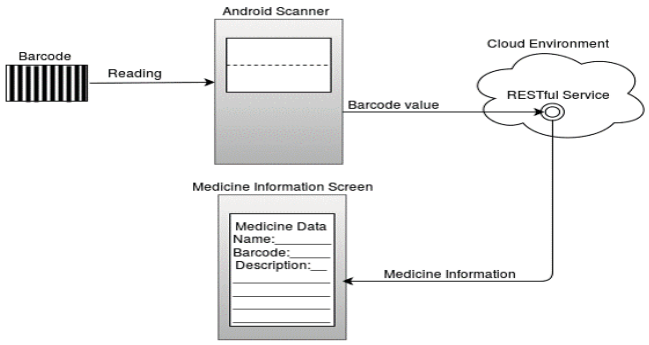


Fig. 5: Retrieval of medicine information on Android mobile

6. Mobile component of MDS

This is an Android application that has three main features. First it is capable of reading barcodes very quickly. For reliable and high-performance barcode scanning, MDS uses ZBar library. In addition to this, a button for the flash light is also added on the barcode scanning screen to make barcode visible to the barcode scanner during complete darkness. Secondly, after scanning, the barcode value is sent to the cloud environment via RESTful API. It means this Android application is capable of communicating with cloud environment to fetch the medicine's data for the given barcode value. The third important feature of this application is to keep track of medicines that are most frequently searched. This application stores medicine's name and number of times it is accessed from Amazon cloud environment on a particular date inside local SQLite database running on Android device. This locally stored information is then used to generate tabular and visual usage reports as shown in 6. These reports help ensure that these medicines are always in stock.

Fig. 6: Screenshots for medicine usage reports

7. Evaluation

In this section, we evaluate the accuracy and performance of Medicine Detection System by storing mock data of 500 medicines inside the Amazon cloud storage environment. The name and number of fields for different medicines might be different as all the data is randomly generated. Figure 7 shows the performance of Medicine Detection System recorded under different lighting conditions by using ten different barcodes that are randomly picked from the mock data.

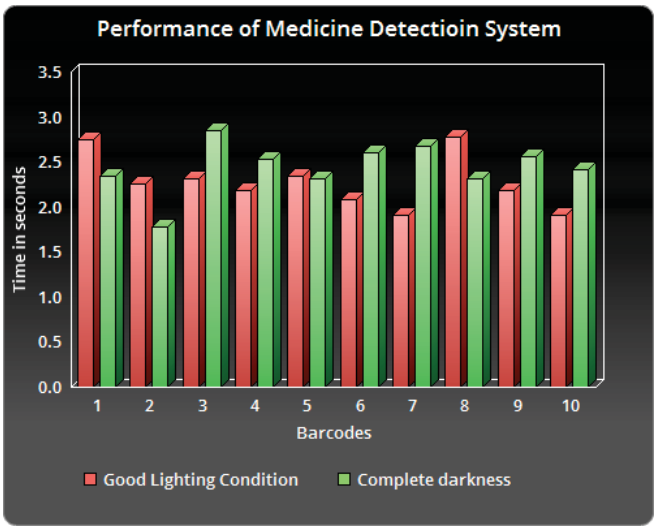


Fig. 7: Performance of MDS under different lighting conditions

As shown in figure 7, the minimum time taken by MDS is 1.78 seconds while the maximum is 2.85 seconds. The small difference in these values indicate that the performance of MDS is almost same under any lighting conditions. It is noticeable that flash light was also used while scanning barcodes under poor lighting conditions. The experimental results demonstrate that MDS takes even less than three seconds for the barcode scanning and data retrieval from the cloud storage. However, it is worth mentioning here that data retrieval time is also subject to the amount of data to be retrieved.

Table 1 clearly shows that by using barcode scanning method, MDS has improved accuracy under any lighting condition as compared to the other systems. However, the only factor that can negatively impact the usability and performance of MDS, is absence or unreadability of medicine's barcode. Since MDS is specifically designed to be used by hospitals and pharmaceutical companies, it is assumed that barcode will always be available for the detection of medicines with hundred percentage accuracy.

Proposed System	System Input	Method Used	Number of Medicines used in Tests	Average Accuracy (%)	Factors Reducing Accuracy
Automatic Drug Image Identification System (ADIIS) [14]	Image of medicine	Features Extraction (Shape, Color, Ratio, Magnitude, and Texture)	263	Rank1: 92.6; Rank2: 95.8; Rank3: 97.4; Rank4: 99.7; Rank5: 99.7; Rank6: 100	a). Poor lighting conditions. b). Broken medicines in the input image. c). Adding more medicines to the test sample.
Pill-ID [15]	Image of medicine	Features Extraction (Shape, Color, and Imprint)	1,029 illicit and 14,002 legal pills	Rank1: 73.04; Rank20: 84.47	a). All included in ADIIS. b). Medicine without imprint.
Automatic identification of prescription drugs using shape distribution models [8]	Image of medicine	Features Extraction (Shape, Color, and Imprint)	568	91.13	Same as Pill-ID system.
Accurate system for automatic pill recognition using imprint information [9]	Image of medicine	Features Extraction (Imprint Only)	12,500	Rank1: 90.46; Rank5: 97.16;	Same as Pill-ID system.
Medicine Detection System	Barcode of medicine	Barcode Scanning	500	100	Unreadable barcode on the packaging of medicine.

Table 1: Comparing accuracy of MDS with other systems

8. Conclusion and Future Work

In this paper, we propose a Medicine Detection System for instant and accurate retrieval of medicine information on mobile devices. MDS exploits the cloud storage facility to store and update the medicines information in such a way that it is instantly accessible from mobile devices whenever needed. MDS uses ZBar image processing library which is highly efficient at reading barcodes. Use of a schema-less database (i.e., MongoDB) in this system, makes it capable of dynamically adding or deleting fields of any medicine. This feature is very important as it ensures that only the desired amount of information for each medicine is stored inside the cloud storage environment. Through the preliminary evaluation, we demonstrate the advantages of utilizing cloud storage and adequacy of the Medicine Detection System.

In the future, we plan to improve Medicine Detection System by incorporating NLP based analysis on the data pertaining to both medicine and patient. As patient's data is usually stored in online healthcare systems like EMR, it can be retrieved from there. The addition of NLP based analysis would make it possible to know if the selected medicine is right for the patient even in shorter amount of time. It would also suggest the recommended dose for the patients considering their age and health conditions.

References

[1] Brennan TA, Leape LL, Laird N, et al. Incidence of adverse events and negligence in hospitalized patients: results from the Harvard Medical Practice Study I. N Engl J Med 1991;324:370-6.

[2] Kohn LT, Corrigan JM, DonaldsonMS. To err is human: building a safer health system. Washington, DC: National Academy Press; 1999.

[3] Pepper GA. Errors in drug administration by nurses. Am J Health Syst Pharm 1995 Feb 15;52(4):390-5.

[4] Leape LL, Bates DW, Cullen DJ, et al; ADE Prevention Group. Systems analysis of adverse drug events. JAMA 1995;274:35-43.

[5] Névéol, A., P. Zweigenbaum, and Section Editors for the IMIA Yearbook Section on Clinical Natural Language Processing. "Clinical Natural Language Processing in 2014: Foundational Methods Supporting Efficient Healthcare." Yearbook of Medical Informatics 10.1 (2015): 194-198. PMC. Web. 1 Jan. 2016.

[6] Drugs.com - http://www.drugs.com. Last access: May 2016.

[7] Lee YB, Park U, Anil K. Jain. PILL-ID Matching and Retrieval of Drug Pill Imprint Images. In: 20th International Conference on Pattern Recognition (ICPR) ;2010. p. 2632 - 2635.

[8] Caban, J.J.; Rosebrock, A.; Yoo, T.S., "Automatic identification of prescription drugs using shape distribution models," in Image Processing (ICIP), 2012 19th IEEE International Conference on , vol., no., pp.1005-1008, Sept. 30 2012-Oct. 3 2012.

[9] Jiye Yu; Zhiyuan Chen; Kamata, S.-I.; Jie Yang, "Accurate system for automatic pill recognition using imprint information," in Image Processing, IET , vol.9, no.12, pp.1039-1047, 12 2015.

[10] Work, M., Improving Medication Safety with a Wireless, Mobile Barcode System in a Community Hospital, A case study, May 2005 - June 2005, Available at https://www.psqh.com/mayjun05/casestudy.html Last access: May 2016.

[11] Poon EG, Cina JL, Churchill WW, et al. Effect of bar-code technology on the incidence of medication dispensing errors and potential adverse drug events in a hospital pharmacy. AMIA Annu Symp Proc 2005:1085.

[12] mLab MongoDB Hosting - https://mlab.com/. Last access: May 2016.

[13] ZBar library - http://zbar.sourceforge.net/. Last access: May 2016.

[14] Chen, Rung-Ching and Pao, Cho-Tsan and Chen, Ying-Hao and Jian, Jeng-Chih: Automatic Drug Image Identification System Based on Multiple Image Features (2010).

[15] Lee YB, Park U, Jain A K, Lee SW. Pill-ID: Matching and retrieval of drug pill images. Pattern Recognition Letters 2011; 33: 904 - 910.

Surgical Capacity Sharing and Cooperation in an Integrated Hospital System

LUO Min
Department of SEEM,
The Chinese University of Hong Kong,
Shatin, N.T, Hong Kong.
mluo@se.cuhk.edu.hk

CAI Xiaoqiang
Department of SEEM,
The Chinese University of Hong Kong,
Shatin, N.T, Hong Kong.
xqcai@se.cuhk.edu.hk

Abstract

In this paper, we address a surgery capacity sharing problem with multiple hospitals, and formulate a model where hospitals who have their own capacity (the Operating Room) locally. Each hospital has several surgery teams facing random demands and seeks appropriate capacities to accommodate the demands, so that its total profit is maximized. We first study the allocation of available surgery capacities in the systems and exploit its resemblance with the newsvendor model. Then we expect to present a approach to study the cooperative games among the hospitals, and formulate it as a stochastic linear programming problem.

Index Terms

healthcare; stochastic; linear programming; game theory; duality.

Contact author: LUO Min
Short Research Paper

1. Introduction

The hospital is experiencing increasing demand and fierce competition while technology is changing the way hospitals structure and operate assignment network to serve patient needs. Long waiting time for hospital care is a problem of great concern in many countries. Consider a typical scenario faced by a patient seeking to book a specific OR for brain surgery on a given day. If the right room, or an acceptable substitute is not available on her schedule. The hospital may give two possible suggestions: i) Go another hospital, please. ii) Do not worry, I will arrange the operating room for you. Normally, she needs access to an information system that provides her with timely data on the availability of unoccupied ORs of other hospitals in an integrated system. Based on this information, the hospital attempts to book an OR time lot which is acceptable to the patient and could be used with acceptable cost. If the operating room can be identified, both the two hospitals must also agree on the appropriate compensation to each party. If all of this can be accomplished, the two hospitals can enjoy the benefits of cooperative and pooling of capacity.

The concept of sharing of inventory is not new, especially in centralized newsvendor model. Which can be conjectured that when demands are uncertain, centralization of capacity helps the hospital reduce resource investment and improve patient satisfaction. However, how to analytically model the inter-hospital collaboration problem? How to derive solutions when capacities are available over multiple hospitals? How to secure the willingness of the various parties to cooperate with each other? This represents a new research effort and many interesting but challenging questions to be addressed.

We formulate a model allows hospitals who could use OR time locally as well as at some other hospitals. The revenues included net profit for lending the room, service fee of surgery team, as well as the costs, are assumed linear and the parameters could vary across the hospitals. Each hospital is faced with her own stochastic demand, which may be correlated with other hospitals. Naturally, we think the decisions can be made at two stages according the realization time of demand. In this complicated situation, one hospital gets the right of use of the OR time lot by paying for it before demand is realized. After the demand is realized, she owns the right to determine how the time lot is to be used, by her team or other hospitals. We analyze the cooperative allocation decision using the notion of a core. In general, not every game has a

ISBN: 1-60132-437-5, CSREA Press ©

nonempty core. However, games with cores' existence are much more conducive to cooperation. We will demonstrate that the core of this game is not empty, and suggest an allocation mechanism to support it.

2. Literature Review

Smith-Daniels et al. [1988] define capacity planning in health care for the first time as "decisions concerning the acquisition and allocation of three types of resources: work force, equipment, and facilities". Batun et al. [2011] think splitting resources can simplify the planning process but may lead to inefficiencies, and they use a two-stage model to quantify the potential benefit of sharing ORs. While inter-hospital collaboration (either within the public or private sector, or the combined system) is commonly recognized as a possible approach to tackle the waiting problems, so far its applications are limited. Could an integrated planning be developed, to provide solutions to optimally utilize the capacities available in an alliance of multiple hospitals of complementary capacities and demands? This is the question we aim to study.

Hospital is a service provider which try to satisfy customer demands with finite capacity. Its objective is to offer the best healthcare at the lowest cost (or the highest revenue). A main idea for us is to extend the techniques which have been established for the relevant newsvendor problems to our model. Anupindi and Bassok [1999] analyzed a centralization problem with n independent retailers of an identical product and each of the retailer faces a stochastic demand. Zhang et al. [2009] propose a binary solution algorithm for multi-product newsboy problem with budget constraint by analyzing properties of its optimal solution. However, all of them deal with one supplier (or retailer) or identical product, and we want to extend the method to multiple suppliers/retailers with various items.

In today's environment it is not sufficient to maximize the expected profit (or minimize the expected cost) simply, proper incentives for cooperation is also essential. The notion of core has been used to analyze cost/revenue allocation problem in other inventory management models as well. Hartman et al. [2000] show that under special assumption on demand distributions the newsvendor game has a nonempty core. This result then be generalized by Müller et al. [2002] who show that the core is always nonempty regardless of the demand distributions. More general inventory centralization games have been studied by Özen et al. [2008] and by Chen and Zhang [2009]. We have seen no research that considers the problem of inter-hospital cooperation for surgical capacity sharing as we propose to investigate in this paper.

3. Capacity Allocation in a Centralized Hospital Systems

We first present a basic model that allow us to discuss the management of capacity. Consider a set $N = \{1, ..., n\}$ of hospitals, each of them has a given capacity, namely the operating room time and a set of surgery teams M_i, ($M_i = \{1, ..., m\}$). Each surgery team $j \in M_i$ faces a random demand, D_i. We assume that the demand is distributed according to a continuous cumulative distribution function $F_i(\cdot)$. Specifically, let h_i be the unit revenue of hospital i for renting out operating room, r_j be the unit revenue surgery team j generated for hospital i, and $c_{j,i}$ be the unit cost of surgery team j performing operations in hospital i. Let the parameters depend on the hospital i, reflecting any possible differences among the hospitals in terms of size, public trust etc. Each hospital i has a total available capacity T_i, and orders quantities of OR time before the actual demand is realized, denoted by $q_{j,i}$ the units of OR time assigned from hospital i to team j.

In the integrated system with pooling, we seek a capacity deployment decision that the performance (total expected profit) of the entire system is optimized. Suppose there is a central planner to determine the solution for the cooperative alliance. The expected total profit of the integrated system can be expressed as follows:

$$\begin{aligned} \Pi_N = & \sum_{i\in N} \{h_i(\sum_{j\in M} q_{j,i}) + E[\sum_{j\in M_i} r_j \min\{D_j, \sum_{i\in N} q_{j,i}\}] \\ & - \sum_{j\in M_i}\sum_{i\in N} c_{j,i}q_{j,i}\} \\ s.t. \quad & \sum_{j\in M} q_{j,i} \le T_i, \quad i \in N \\ & q_{j,i} \ge 0, \quad i \in N, j \in M \end{aligned} \tag{1}$$

where $M = \bigcup_{i\in N} M_i$.

The objective function [1] aims at maximize the total excepted profit of the integrated system. where the 1st term on the RHS is the profit for hospital due to reservation of its capacity, the 2nd and 3rd terms are, respectively, the expected revenue (with respect to distributions of the random demands) and the cost generated by its surgery teams in meeting their demands. Constraints require that the variable q_{ij} can only take non-negative values within respective budgets.

We have the following properties with the objective function:

Proposition 1: The expected profit function Π is concave in $q_{j,i}$, $\forall j \in M, i \in M$.

Proof:

$$\frac{\partial^2 \Pi}{\partial q_{j,i}^2} = -f_j(\hat{q}_j) \leq 0 \quad and$$
$$\frac{\partial^2 \Pi}{\partial q_{i,k} \partial q_{j,l}} = 0 \quad for \quad i \neq j, k \neq l.$$

where $\hat{q}_j = \sum_{i \in N} q_{j,i}$. Thus, the Hessian Matrix of (1) is negative semi-definite. □

Since Π is concave and the feasible domain of problem (1) is convex, the KKT conditions are necessary and sufficient for optimality. Then $(q_{j,i})_{j \in M, i \in N}$ is optimal if and only if there exist non-negative dual variables $\lambda_i, u_{j,i}$ satisfy that

$$h_j - c_{j,i} + r_j - r_j F_j(\hat{q}_j) - \lambda_i + \mu_{j,i} = 0 \quad (2)$$
$$\lambda_i (\sum_{j \in M} q_{j,i} - T_i) = 0 \quad (3)$$
$$\mu_{j,i} q_{j,i} = 0 \quad (4)$$

We first investigate how to solve the above equations with any given $\lambda_i \geq 0$, and then we show how to find the optimal value. Denote by $(\tilde{q}_{j,i}, \lambda_i, \tilde{u}_{j,i})$ an solution of the equations, then we have:

Proposition 2: For any given $\lambda_i \geq 0$, $\forall j \in M, i \in N$,

$$\sum_i \tilde{q}_{j,i} = F_j^{-1}(\frac{h_i - c_{j,i} + r_j - \lambda_i}{r_j}) \quad (5)$$

Proof: If $h_i - c_{j,i} + r_j < \lambda_n$, then we have $F_j(\hat{q}_j) < u_{j,i}$ from Eq.(3). $F_j(\hat{q}_j) < u_{j,i}$ plus $u_{j,i} q_{j,i} = 0$ implies $\tilde{q}_{j,i} = 0$. If $h_i - c_{j,i} + r_j \geq \lambda_i$, then we have $F_j(\hat{q}_j) \geq u_{j,i}$ from Eq.(3). $F_j(\hat{q}_j) \geq u_{j,i}$ and $u_{j,i} q_{j,i} = 0$ implies $u_{j,i} = 0$. Then we have $\sum_i \tilde{q}_{j,i} = F_j^{-1}(\frac{h_i - c_{j,i} + r_j - \lambda_i}{r_j})$.

□

This proposition characterized the optimal total capacity of team j in all hospitals with any given $\lambda_i \geq 0$, and also indicates the optimal solution to the problem without budget constraints by setting $\lambda_i = 0$. Since $\lambda_i \geq 0$, we know that the total optimal unconstrained order of OR time is an upper bound for the total optimal order in problem (1).

Proposition 3:

$$(a) If \sum_j \tilde{q}_{j,i} \leq T_i, then \sum_j q_{j,i}^* = \sum_j \tilde{q}_{j,i};$$
$$(b) If \sum_j \tilde{q}_{j,i} > T_i, then \sum_j q_{j,i}^* = T_i;$$

Proof: (a) This property is obvious since the budget constraint is not active.

(b) If $\sum_j q_{j,i}^* < T_i$, as $\sum_j \tilde{q}_{j,i} > T_i$, there must exist at least one $k \in 1, ..., n$ such that $\sum_j q_{j,k}^* < \sum_j \tilde{q}_{j,k}$. However, since $\sum_j q_{j,i}^* < T_i$, the slackness condition in Eq.(4) implies $\lambda_i^* = 0$, and this further means $\sum_j q_{j,i}^* = \sum_j \tilde{q}_{j,i}$, which violates $\sum_j q_{j,k}^* < \sum_j \tilde{q}_{j,k}$. Thus, we have $\sum_j q_{j,i}^* = T_i$. □

Before developing the solution procedure, we first prove the following result:

Proposition 4: $\sum_j \tilde{q}_{j,i}$ is non-increasing in λ_i.

Proof: The proof is obvious from proposition 2. □

Thus we can determine λ_i^* by a binary search over the interval $\lambda_i \in [0, \max_{i=1,...n}(h_i - c_{j,i} + r_j)]$ and the optimal total capacity $\sum_i q_{j,i}^*$. Note that $(q_{j,i}^*)$ now satisfies a set of linear (in)equalities, thus we can solve it by linear programming.

Algorithm 1 Main steps of algorithm

Step 0: Calculate y_j, $j = 1, 2, ..., m$, from Eq(5) by setting $\lambda_i = 0$.
Step 1: If $x_i \leq T_i$, then $x_i^* = x_i$,$i = 1, ..., n$, goto Step 6.
Step 2: Let $\lambda_i^L = 0, \lambda_i^U = \max_{j,i}\{r_j + h_i - c_{j,i}\}$,.
Step 3: Let $\lambda_i = (\lambda_i^L + \lambda_i^U/2)$;
Calculatey_j^λ, $j = 1, ..., m$.
Step 4: If $x_i^\lambda \leq T_i$, then $\lambda_j^U = \lambda_i$, goto Step 3;
If $x_i^\lambda > T_i$, then $\lambda_i^L = \lambda_i$, goto Step 3.
Step 5: Let $\sum_i q_{j,i}^* = y_j^\lambda$ and $\sum_j q_{j,i}^* = x_i^\lambda$, $i = 1, ..., n, j = 1, ..., m$.
Calculate the linear system.
Step 6: Output $q_{j,i}^*$, stop.

4. Cooperation Between the Hospitals

4.1. Preliminaries

In the setting of this paper, the hospitals are planning jointly in order to reach a overall service level and achieve the maximum revenue. These hospitals are referred as *players*, denoted with set N $(N = 1, 2, ..., n)$. A subset of N is called a *coalition* and is denoted by S. $v(S)$ is defined as the the maximum total revenue the coalition S can obtain. The function v that assigns to every coalition $S \in N$ its value $v(S)$, with $v(\emptyset) = 0$, is commonly referred to as the *characteristic function*. The *allocation* is represented as a payoff vector $x = (x_i)_{(i \in N)}$, which is the profit that the grand coalition allots to hospitals i if he joins the coalition. We seek a so-called balanced distribution $x_1, x_2, ..., x_n$ such that

$$\sum_{i \in N} x_i = \Pi_N, \quad and \quad \sum_{i \in S} x_i \geq \Pi_S \quad \forall S \in N. \quad (6)$$

These (in)equalities completely describe the cooperative game and are referred to as core (in)equalities. In order to ensure that the grand coalition is stable , it is imperative to devise a distribution scheme that meets the core inequalities of the game.

4.2. Deterministic situation

We start with examining the situation where each surgery team has such a high demand that is definitely larger than the total capacity $\hat{q}_j$ the team j may be allocated. The problem becomes a deterministic linear program:

$$\begin{aligned} \Pi_N = & \sum_{i\in N}[h_i(\sum_{j\in M} q_{j,i}) + \sum_{j\in M_i}\sum_{i\in N}(r_j - c_{j,i})q_{j,i}] \\ s.t. \quad & \sum_{j\in M^N} q_{j,i} \le T_i, \quad \forall i \\ & q_{j,i} \ge 0, \quad \forall j,i \end{aligned} \tag{7}$$

According to Owen [1975], this linear game is balanced and has a non-empty core. However, tt is important to find points in the core and we consider the dual to linear program (7):

$$\begin{aligned} & \min\,(T_1, T_2, ...T_{\mathrm{N}})^T \cdot y \\ & s.t. \quad y_i \ge \max_j(h_i + r_j - c_{j,i}, 0) \end{aligned} \tag{8}$$

where $y = (y_1, y_2, ..., y_n)$.

As it is well known, $v(N)$ will be equal to the maximum of the program. We note (8) has a unique solution. Let y_i^* be the solution vector for (8). Then,

$$v(N) = T_1 y_1^* + T_2 y_2^* + ... + T_N y_N^* \tag{9}$$

and for any S

$$v(S) \le T_1 y_1^* + T_2 y_2^* + ... + T_s y_s^* \tag{10}$$

Define payoff vector $u = (u_1, u_2, ..., u_N)$ as $u_i = T_i y_i^*$. so we have $\sum_{i\in N} u_i = v(N)$ and $\sum_{i\in S} u_i \ge v(S)$, therefore u is an imputation in the core. Now we can simply compute the vector y^* by solving linear program of reasonable size; this then gives us an imputation u.

Based on the insight from the special case above, we will consider a general situation with a stochastic demand in the future.

5. Conclusion

In this paper we study allocation of surgery capacities available in an integrated system of multiple hospitals. We propose a new method based on the linear programming to compute the optimal solutions. We also study a cooperative game amongst the hospitals and show that in the case of a high demand situation there exists an allocation in the core, which can be computed by using the linear programming duality theory. The special case inspires us to develop a stochastic programming duality approach in analyzing the cooperative games with multiple hospitals where random demand are faced.

Acknowledgment

The work was partially supported by Hong Kong RGC/TRS Project No. T32-102/14N.

References

Ravi Anupindi and Yehuda Bassok. Centralization of stocks: Retailers vs. manufacturer. *Management Science*, 45(2):178–191, 1999.

Sakine Batun, Brian T Denton, Todd R Huschka, and Andrew J Schaefer. Operating room pooling and parallel surgery processing under uncertainty. *INFORMS journal on Computing*, 23(2):220–237, 2011.

Xin Chen and Jiawei Zhang. A stochastic programming duality approach to inventory centralization games. *Operations Research*, 57(4):840–851, 2009.

Bruce C Hartman, Moshe Dror, and Moshe Shaked. Cores of inventory centralization games. *Games and Economic Behavior*, 31(1):26–49, 2000.

Alfred Müller, Marco Scarsini, and Moshe Shaked. The newsvendor game has a nonempty core. *Games and Economic Behavior*, 38(1):118–126, 2002.

Guillermo Owen. On the core of linear production games. *Mathematical programming*, 9(1):358–370, 1975.

Ulas Özen, Jan Fransoo, Henk Norde, and Marco Slikker. Cooperation between multiple newsvendors with warehouses. *Manufacturing & Service Operations Management*, 10(2):311–324, 2008.

Vicki L Smith-Daniels, Sharon B Schweikhart, and Dwight E Smith-Daniels. Capacity management in health care services: Review and future research directions*. *Decision Sciences*, 19(4):889–919, 1988.

Bin Zhang, Xiaoyan Xu, and Zhongsheng Hua. A binary solution method for the multi-product newsboy problem with budget constraint. *International Journal of Production Economics*, 117(1):136–141, 2009.

SESSION

POSTER PAPERS

Chair(s)

TBA

ISBN: 1-60132-437-5, CSREA Press ©

ISBN: 1-60132-437-5, CSREA Press ©

MedInternet: Application of Artificial Intelligence for Medical Data Collection and Analysis

Babək Murad-Kəngərli [1],
[1] MedEffect LTD, Baku, Azerbaijan

Abstract **-** MedInternet is a program of general medical information character and can embrace all areas of medicine. It allows you to register via the Internet, organize and analyze the processes taking place in the world of medicine. MedInternet creates a strong basis on which to develop the entire computer medicine. The program will have its own think tank composed of top doctors and programmers, as well as scientists and experts in other similar areas. MedInternet has a potential to unite and centralize the whole world of medicine accentuating the application of artificial intelligence on information features and systems.

Keywords: *Information Technologies, Health Data, Diagnostics, Decision Support, Data Mining and Machine Learning*

1 Introduction

Around the world, medical practitioners make diagnosis and treatment based on their available medical knowledge, technology, equipment and tools. The whole process of diagnosis and treatment, and the results the doctors reflect on paper or computer files. This is based on local servers and local data sources. More recently there has been an ever-growing interest in computer based medical data collection and analysis, as more and more people use computers, smart phones and other information technologies. Meanwhile more doctors in various countries use computers to record patient's data and analyze and diagnose the illnesses. This has resulted in variety of sources for medical data and alike, such as BIG DATA or Internet of Things. Doctors, researchers, scientific and medical centers, pharmaceutical companies, the official state medical institutions, insurance companies collect, systematize and analyze the medical data and create statistics of them, write articles and papers and make inventions and innovations of existing treatments and diagnostics. On the basis of all this is the development of medicine. The process, as we know a long, fragmented and insufficiently effective given the fact that the medicine has not yet been able to become fully a science, and relies heavily on the experiences, skills and subjective opinions of doctors. Therefore it is not surprising that in recent years there has been a great interest in the computer technologies in medicine.

1.1 The need for a universal medical software

In recent years, there have been many attempts to create both simple programs (for example, an appointment to the doctor via the Internet) and more complex, such as genetic and biomedical programs. These developments are actively engaged in thousands of small and medium-sized companies around the world. However, no major changes in the world of computerized medicine has happened yet.

There is a strong need to unite and centralize this collection of medical knowledge. My proposed project "MedInternet" is based on a completely new methodology which will be the space for recording, collecting, storage, and analysis of all possible medical scientific information and data. The program directs and controls all (or almost all) of the processes of scientific and informational nature, occurring in the world of medicine. The program allows to transfer the information of the above-mentioned processes from the traditional manual operation to a computerized one. The program is based on artificial computer intelligence that develops and updates its own knowledge and creates new knowledge.

2 MedInternet at work

Here is how it looks:

- Doctors engage in practice of their daily work - diagnose and treat patients by use of an efficient computerized tool. They use a conventional computer where in the course of their work they enter the relevant information about the condition of the patient. This allows them to determine an exact diagnosis of the patient and its treatment.
- This information, in turn, is recorded in a special form in a cloud that has the ability to analyze it in every way, to develop statistics and provide data for scientific and medical activities, and develop new medical knowledge.
- Doctors, researchers, scientific and medical centers, pharmaceutical, and insurance companies use this information to create new knowledge and new medical technology, drugs and medical equipment.
- And all this happens in online mode, continuously covering the entire Internet space.
- The amount of information, which is colossal is collected while medical practitioners are engaged in their daily activities.
- Most of the research work is carried out by specially trained medical programs that runs within the cloud.

2.1 Use of MedInternet

The program consists of five parts. The first part is Diagnostics and Treatment which is for usage of practitioner doctors within clinics. This part is already developed and tested. Diagnostics is the program's main core and fuel. In this part using a special software, doctors are able to diagnose and identify existing patient's disease, choose the right treatment protocols, to assign the treatment and state the treatment

results. Currently it has been programmed with more than 500 most common diseases. This part is planned to be extended to include all diseases in MCB-10 database and exposed on the Internet which will benefit medical providers around the world. Doctors and clinics are offered a help to determine the diagnoses and the treatment. In the process of diagnosis and treatment, a wide possibility of online consultations with colleagues around the world, the use of medical knowledge, and other tips will be created. The program will also create an opportunity to diagnose the patient's existing latent diseases.

The second part consists of a SuperCloud. In this SuperCloud, all the work of doctors in the diagnosis and treatment of patients is stored in a special, unified, standardized, computer–formulized form. It reflects and stores all of the processes of diagnostics, choice of protocols, treatments and results of the treatment. These data, with the help of special programs with algorithms of Case-Based Reasoning, Data mining and artificial intelligence of the SuperCloud, is permanently systematized, processed, analyzed. Thus updated medical statistics and new medical knowledge are created continuously.

The third part consists of Medical Knowledge. Existing medical knowledge and new information created by artificial intelligence of MedInternet are continuously added into this part of the program. This part of the program allows doctors to immediately obtain the necessary professional and scientific medical knowledge, statistics, case summaries both in the process of working with the patient, and outside it. The knowledge is created via the operations in the Super Cloud and feeds back to medical providers and researchers.

The fourth part consists of electronic-medical card (passport) of a patient. It is formulated during the process of communication between the patient and doctor, medical screening, medical testing and collection of anamnesis. The data is recorded as a medical process in a dynamic environment. Maximum display of the information about the state of the patient, his history is reflected in special and standard form convenient for analysis. This part of the program is also available to the patient and he will complement it with additional information independently. With the help of special micro programs, data is constantly analyzed and medical advice for practitioners is created.

The fifth part is the Center of Control, which consists of administrative managers, providers, programmers, other researchers and scientists, and the brain centre. This part of the program deals with the constant analysis of information received and generates new scientific knowledge and new ways of working. It creates statistics, materials and knowledge for doctors and researchers. Managers control the current activities of the program.

2.2 The Beneficiaries

The beneficiaries of the project are as following:

- Doctors – who receive an effective tool for the diagnosis of disease and a choice of ways to treat it. At the same time, they get new knowledge and information about their work and can get an opportunity of online consultations. At the same time, doctors find themselves in a constant environment of medical research which boosts their confidence.
- Patients – who get more quality medical help and can control the quality of their treatment and get new ways of improving their health.
- Researchers - who get a chance of receiving trusted medical statistics and medical information and get updates in the shortest time periods. They can conduct and monitor a variety of experiments and studies of medical and informational character in online mode from all over the web space. I note that this does not create any additional new horizontal units of local networks and does not need any volunteers to conduct research. They can receive and continuously exchange information with the world's scientific and medical community, medical practices, as well as carry out joint development in the virtual world.
- Insurance companies - who will have an ability to control the medical processes directly from their workstation via a computer in any mode. They can affect the workflow of clinics and practitioners, encouraging them to work through MedInternet. It also guarantees that doctors do not overcharge their patients and do not assign treatment that is costly and not necessary. This way efficient mutual control and influence is created among insurance companies, providers and patients.
- National Health Ministries - who will study, control, manage and direct world's medical processes online. World Health Organization will gain access to regulating the entire world of medicine.

3 Conclusions

I acknowledge that there a lot of barriers with cross-country, political and economic issues, different standards in approaches to confidentiality that are argued to stand in front of this project being applied into real world. However these issues can be solved with an expertise of IT giants and the will of the medical community, standardizing the medical practices around the whole world. Moreover, I propose a novel methodological tool of Matrix application, which I have developed. The tool will speed up the process and reduce the cost of creating global MedInternet without which it may take billions of dollars. A connection of artificial intelligence for the diagnosis and analysis of medical data offered by me will leave the project of global MedInternet without an alternative. I ask all the experts to respond to this article and give their opinions, and the IT giants and venture companies to consider my idea in detail, remembering what of a socio-economic impact for the future it may carry. MedInternet can be the foundation that can combine and centralize in a single source the entire global medical processes while guiding and developing it.

An ontology-based data warehouse for diagnosis and communication in intensive care settings

Jeroen S. de Bruin[1], **Mohamed Mouhieddine**[2], **Christian Schuh**[3], **Michael Hiesmayr**[2]
[1]Institute for Artificial Intelligence and Decision Support, Medical University of Vienna, Vienna, Austria
[2]Division of Cardiac- Thoracic- Vascular Anaesthesia and Intensive Care, Vienna General Hospital, Vienna, Austria
[3]IT-Systems & Communications, Medical University of Vienna, Vienna, Austria

Abstract - *In the intensive care unit (ICU), a timely supply of all needed information is of the utmost importance. To facilitate this demand, we plan for a data warehouse for the ICU that employs data from multiple clinical sources as well as clinical decision support systems for analysis. This clinically integrated system enables the automated generation of concise and accurate reports, the automated classification of patient symptoms and diagnoses, and evidence-based treatment planning. Three pilot projects are planned to implement and showcase the aforementioned capabilities of the data warehouse, all using knowledge-based methods and the International Classification of Diseases, 10th version as their foundation. The final solution is deemed feasible, expected to be interoperable with existing hospital information systems due to the extensive use of standards, and likely to support and impact existing clinical workflows in intensive care medicine.*

Keywords: Knowledge Bases, Clinical Decision Support, Electronic Data Processing, Critical Care.

1 Introduction

In a stressful environment such as the intensive care unit (ICU), it is paramount that the necessary information is offered to the right persons in a timely manner. Information and reports need to be concise but complete. Too little relevant information will result in suboptimal care, while too much information will confound important facts and distract. Therefore, information about a patient has to be individualized, thereby providing data on relevant chronic and acute patient symptoms, diagnoses, and treatments. Furthermore, diagnoses need to be determined fast and accurately as a patient's health is already severely compromised, and treatment or stabilization needs to proceed as fast as possible. Given these requirements, the medical team responsible for a patient's treatment would benefit from a clinical decision support system (CDSS) that could rapidly determine (or confirm) a patient's diagnosis, and streamline communication by providing information relevant to the patient's diagnosis, characteristics, or treatment.

2 Proposed solution

Our goal is to design a data warehouse for the ICU that, using data available in the ICU's patient data management systems as well as in the hospital information system, provides expert-systems for 1) the generation of accurate, role-specific reports with high-information density that contain all relevant information while leaving out unnecessary details, 2) classification of patients based on their symptoms, past diagnoses and treatments, for the determination of optimal treatments or prediction of disease progression, and 3) the presentation of data and information from the patient data management system directly in the hospital information system, as part of an integrated solution for patient care.

For the generation of reports, a knowledge base was planned that defines for each medical role/profession involved in the treatment of a patient, which data should be included in the report, to what detail it should be included, and how it should be presented. For the classification of patient cases, the CDSS employs the patient data management system to assign values to higher-level, semantically rich and clinically relevant concept such as symptoms and diagnoses. To support the selection of treatment for these patients, a comparison with past cases can also improve healthcare quality, while speeding up the decision process; the effectiveness of different therapies in different cases may serve as a guideline or tiebreaker for the medical team in the choice of therapy.

To keep nomenclature uniform, we employ the International Classification of Diseases, 10th version (ICD-10) ontology for both classification and communication. Based on these classifications and knowledge on disease progressions, potential diagnoses might be predicted together with likelihood and options for intervention and treatment.

3 Pilot projects

To support a wide range of ICU protocols and workflows, three pilot projects are currently under development. The first project pertains a CDSS for the diagnosis of systemic inflammatory response syndrome (SIRS), which is the body's response to an infectious or noninfectious affliction. It employs an ICD-10-based knowledge base for the accurate determination of SIRS symptoms. Symptoms are thereby determined in a complex, comprehensive fashion, rather than simplistic rules. Fever, for example, is not determined by the

standard rule stating that a patient's body temperature needs to be 38°C or greater, but is rather more individualized, as recorded normal body temperatures vary depending on many factors, including age, sex, time of day, ambient temperature, activity level, and method of measurement.

The second pilot application is meant for both the diagnosis and treatment of the acute respiratory distress syndrome (ARDS). ARDS is a severe, life-threatening medical condition characterized by widespread inflammation in the lungs. While ARDS may be triggered by a trauma or lung infection, it is usually the result of sepsis. For this pilot project, an ICD-10 knowledge base is planned that can detect ARDS, determine the symptoms, and based on those symptoms, propose interventions or treatments. Based on patient demographics, ontological annotation(s) of a patient case, as well as symptom and disease progression learnt through sequential analysis of data over time, a patient case can be classified with respect to past patient cases using a propensity score analysis. After a patient has been classified, the resulting group of similar patient cases can be presented, including their treatment details and outcome; this will provide the medical team with evidence-based background information that may help them in the determination of treatment, especially in complex cases.

The final pilot project pertains report generation for communication. In the ICU-environment, time is a luxury, and it is therefore important to have concise but complete reports on patient for quick communication. By communicating all the necessary information, and only the necessary information, quality of healthcare can improve, while time can be saved. To this end, a knowledge base is designed that, depending on the role of the user, generates all information on a patient or set of patients in a ward, so that their information can be quickly communicated between healthcare professionals, or used to provide optimal care and monitoring.

4 Discussion

We discussed a data warehouse currently under development with (semi-)automated patient classification and diagnosis capabilities, as well as the ability to generate high-quality communication reports. For the collection of data, we use patient data management systems commonly available in the ICU setting, as well as demographic and clinical data from the hospital information system. For data representation and communication, we use a widely accepted medical ontology called ICD-10 to keep nomenclature and communication uniform.

Currently, aforementioned systems are still under development, but early trials have already proven this system to be feasible. Furthermore, it's integrated directly in clinical routine without the need for separate clients or processing, which enables the medical team to work in a familiar digital environment, and removes the need to familiarize themselves with another application.

However, there are also some limitations worthy of note. First, the choice of ontology determines the expressiveness of the system, as well as its ability to relate clinical linguistic concepts to raw data. While ICD-10 might be a widely accepted medical ontology, it is not the only one. The Systematized Nomenclature of Medicine--Clinical Terms (SNOMED-CT), for example, is also a medical ontology, and even more comprehensive than ICD-10. However, its use is not free. Furthermore, the system is currently only designed for functions involving retrospective data analysis. For the system to be more supportive to the ICU medical team and the ICU setting in general, real-time data analysis to support on-the-fly protocols and workflows should also be supported. This will be a future stage of the project.

Estimating Energy Expenditure by Using Radial Basis Function Network for Health Monitoring

Meina Li[1] , Seokeun Park[+] and Youn Tae Kim[+]
[1]College of Instrumentation and Electrical Engineering, Jilin University, Jilin, China
[+]IT Fusion Technology Research Center, Chosun University, Gwangju, Korea
[+]*petruskim@chosun.ac.kr*

***Abstract*— The main purpose of this study is to effectively estimate energy expenditure (EE) of walking and running in young healthy adults by using radial basis function networks (RBFN) method. 30 participants were recruited from college. The participants performed from walking to running by the submaximal treadmill protocol. Heart rate (HR) and movement index (MI) were monitored real-time and recorded by the patched type sensor. The estimated values were compared to a portable indirect calorimeter (Cosmed K4b^2). Results of the study illustrated the test based on treadmill has the high accurate estimation than the traditional method.**

Keywords— Energy Expenditure, Heart rate, Movement intensity, RBFN, Health monitoring.

Type of the Submission – Extended Abstract/Poster Paper

I. INTRODUCTION

3.5 percent of adult population in Korea is classified as obese, even if the ratio is lower than for other countries [1]. Excess nutrient and energy imbalance are considered major causes of chronic disease, such as diabetes [2], and obesity. Energy is needed to power muscular activity but excessive energy storage can lead to obesity and other metabolic disorders; it is important to understand how energy expenditure can be assessed and quantified. The most common activities of adults in modern society are walking and running. The assessment on walking and running can estimate EE and evaluate body fitness [3]). Therefore, the accurate assessment of physical activities can be helpful for Asian human health as the reference of metabolism. Traditional method need gas system to measure the oxygen consumption and carbon oxygen production. This method is high accuracy but need the participants wear mask during walking and running. Therefore, in this study we present the high accuracy method by using radial basis function network (RBFN).

II. RADIAL BASIS FUNCTION NETWORKS

RBFN have very attractive properties such as functional approximation, localization, interpolation and cluster modeling [4]. These properties made them attractive in many applications. The structure of RBFN has three layers: input layer feed the feature vectors into the network; hidden layer calculate the outcome of the basic functions; and output layer calculate a linear combination of the basic functions. Different numbers of hidden layer neurons were tried in this study as shown in Fig. 1. The two input parameters are HR and MI, and the output value is EE. The activation level of the i th receptive hidden unit is

$$\omega_i = F_i(\chi) = F_i(\| \chi - c_i \| / \sigma_i), i = 1,2,...,H, \quad (1)$$

where χ is multidimensional input vector, c_i is a vector with the same dimension as χ , H is the number of radial basis functions, and $F(\bullet)$ is i th radial basis function with a single maximum at the origin.

Consider a Gaussian basis function centered at c_i with a width parameter σ:

$$\omega_i = F_i(\| \chi - c_i \|) = \exp\left[-\frac{(\chi - c_i)^2}{2\sigma_i^2} \right]. \quad (2)$$

Each training input χ_i server as a center for the basis function, F_i . Thus, a Gaussian interpolation RBFN:

$$d(\chi) = \sum_{i=1}^{n} u_i \exp\left[-\frac{(\chi - c_i)^2}{2\sigma_i^2} \right]. \quad (3)$$

Writing them in matrix form,

$$\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{bmatrix} = \begin{bmatrix} \exp\left[-\frac{\| \chi_1 - \chi_2 \|^2}{2\sigma_1^2} \right] & \cdots & \exp\left[-\frac{\| \chi_1 - \chi_n \|^2}{2\sigma_n^2} \right] \\ \exp\left[-\frac{\| \chi_2 - \chi_1 \|^2}{2\sigma_1^2} \right] & \cdots & \exp\left[-\frac{\| \chi_2 - \chi_n \|^2}{2\sigma_n^2} \right] \\ \vdots & \vdots & \vdots \\ \exp\left[-\frac{\| \chi_n - \chi_1 \|^2}{2\sigma_1^2} \right] & \cdots & \exp\left[-\frac{\| \chi_n - \chi_n \|^2}{2\sigma_n^2} \right] \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}. \quad (4)$$

Rewriting the preceding in a compact form,

$$D = GC, \quad (5)$$

where

$$D=\left[d_1, d_2, \cdots, d_n\right]^T,$$

$$C=\left[u_1, u_2, \cdots, u_n\right]^T. \quad (6)$$

When the matrix G is nonsingular, the unique solution as:

$$C = G^{-1}D. \quad (7)$$

where the matrix G^{-1} denotes the inverse matrix of G.

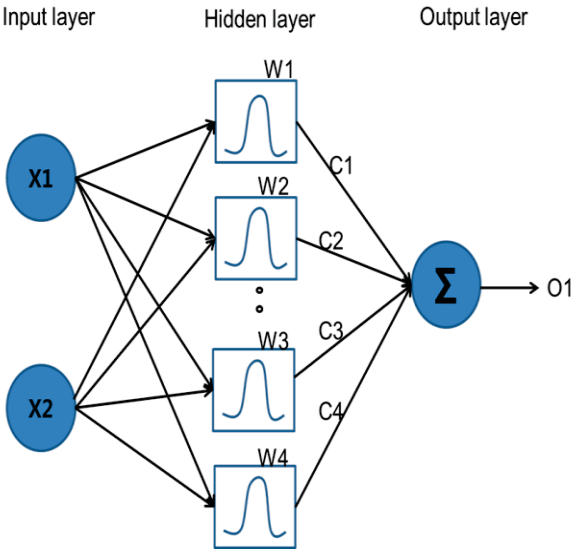


Fig.1 Structure of a radial basis function neural network.

III. Results

The participants were tested on the treadmill base on the sub-maximal Bruce protocols. The treadmill was started at 2.7 km/h and at a gradient of 10%. The incline of the treadmill increases by 2% at three minute intervals. HR and MI were monitored in real-time by AirBeat system [5]. VO_2 and EE were measured by indirect calorimeter.

The progress of finding the RBFN weights is called network training. The set of input-output pairs is called training set. In order to fit the network EE output to the given input (HR and MI) the network parameters are optimized. The training and optimized results of estimating the EE compared with measured value by the RBFN method as shown in Fig. 2. Various hidden layer node number were tested for the RBFN algorithm. Walking involves 20 input nodes and the optima number of node is 12. The root mean square error (RMSE) values between measured and estimated EE were used to evaluate the performance of RBFN method. RMSE is 0.722 for training data and 0.961 for checking data.

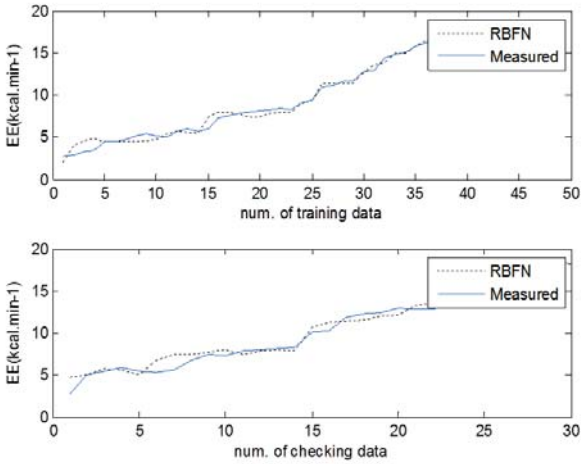


Fig. 2 Estimation of energy expenditure by the RBFN method.

IV. Conclusion

This is the first study to apply the RBFN in estimating energy expenditure for Asian people. More research is needed to develop the entire system, including the shape of the patch type sensor and the software for the calculation of the energy expenditure with an easy to use application interface. The primary aim of this study was to assess the energy expenditure of walking and running for young health adults by RBFN method. The EE were accurately estimated based on the treadmill under submaximal protocols. The heart-rate and movement combined sensor module demonstrated its ability to estimate EE during walking and running.

Acknowledgment

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2015-R0992-15-1021) supervised by the IITP(Institute for Information & communications Technology Promotion).

References

[1] Moon GL. Korea rate of obesity ranks lowest among OECD nations. JoonagAng Daily Social Affairs. 2009.

[2] Hustvedt BE, Svendsen M, Ellegard AL, Hallen J, Tonstad S. Validation of ActiReg to measure physical activity and energy expenditure against doubly labeled water in obese persons. Bri J Nutr. 2008;100:219-226.

[3] Armstrong N, Welsman J. Aerobic fitness: What are we measuring? Med Sport Sci. 2007;50:5-25.

[4] Jyh-Shing Roger Jang, Chuen-Tsai Sun, Eiji Mizutani. Neuro-Fuzzy and soft computing. Prentice Hall. 1997; 238-246.

[5] M Li, and Y Kim, Development of patch-type sensor module for wireless monitoring of movement index. Sens. Actuators A: Phys., 2012, vol. 173, pp. 277-283.

SOLUTION TO HIT CONNECTIVITY & EFFECTIVE HEALTHCARE DATA USE

Barry D. Silbermann, BSIE, JD, Project Director
Roger D. Blake[1] Director Systems Engineering
Stephen C. Berens, MD, Director Medical & Research Affairs
Foundation for Advanced Philanthropy Fund for Healthcare Engineering and Economics
%Pergamum Analytics & Technologies
3896 Carpenter Avenue,
Studio City, CA 91604
Tel: (310) 702-6699

1. The Connectivity Problem

> *U. S. medical society organizations are currently telling the Centers for Medicare & Medicaid Services (CMS) that physicians' use of Electronic Health Records (EHR) under proposed regulations* for the Medicare Access and CHIP[2] Reauthorization Act of 2015 (MACRA) "*needs new performance measures, not old recycled ones, when it comes to using EHRs* to improve patient care and hold down costs. ***And demonstrating that different EHRs talk to one another* should not pose an administrative burden to physicians, according to these groups**."[3]

The 2006 Rand Health Research Brief (*Rand Report*) was the genesis of The Clinical Health (HITECH Act), of the American Recovery and Reinvestment Act (ARRA), appropriating billions of dollars **for healthcare providers to adopt and use** Healthcare Information Technology (HIT ***and remains unresolved in 2016[4]. Thus there is no mechanism for all EMR vendor systems to connect to one another for information exchange[5]***. The *Rand Report* defined "connectivity," summarized as follows:

> "***Connectivity***—the ability to share information from system to system—**is poor (sic)**. "

[1] Roger Blake holds a Black Belt Six Sigma

[2] CHIP is the abbreviation for "Children's Health Insurance Program"

[3] http://www.medscape.com/viewarticle/864493?src=wnl_edit_tpal, Robert Lowes
June 08, 2016

[4] The 2006 Rand Health Research Brief at pg. 3 and 2015 *Enterprise HIE Survey Black Book Rankings*

[5] *Id.* fn. 3

ISBN: 1-60132-437-5, CSREA Press ©

2. The Connectivity Systems *Solution*

An excellent but limited example of the clinical application and utility for use of EMR, was published by Stanford pediatricians in 2011[6]. To solve an unusual clinical problem, they applied a word-based search engine among all of their own Stanford providers' medical records sites. (A word or dialogue based search analogous to a Google type search is more precise than a **diagnostic code** search which can contain coding errors introduced by coders or variances of code used for similar or the same diagnosis at different sites). They uncovered more cases to apply a solution to a difficult clinical problem. The physicians stated that,[7] "*We will . . . know that we made the decision on the basis of the best data available . . . In the practice of medicine, one can't do better than that.*" **Our designed approach** to solve these problems starts with:

Our R & D group's HIT information interchange (connectivity) for private and public medical records system connects the disparate, participating provider, payer and research sites' data hardware and software systems. This solution will have user-friendly data selection for clinical, research and financial use.

The solution we propose uses existing technology and proven software design currently in use. The software is tailored to the existing EMR (vendor computer and software systems) in use by any sites that have patient records. The sites' computers are then interconnected. Information flows to the medical providers', facilities' or payers' processing sites via a supercomputer HUB on a dedicated, non-Internet, secure data transport highway.

This would seamlessly overcome industry *fear of the unknown* resistance to an HIT connectivity solution with the HUBS at a ***cost of less than 1/10th*** of many of the provider and payer IT systems. HIPPA compliance is met with end-to-end encryption and confirmed provider-patient consents or de-identified data and Supercomputer/Data Highway HUBS housed in secure environments & personnel subject to HIPPA compliance policies

3. The Projects Value

The big picture of hospital connectivity in the U.S. today is "profoundly negative," says cardiologist Eric Topol, MD author . . . and chief academic officer at San Diego-based Scripps Health. "There's been tremendous resources put into this and little to show for it," he says. "We have a country characterized by information-blocking, where there is a lack of connectivity from one health system to another, and patients are the ones who are collateral damage because of all this Tower of Babel."[8]

[6] Evidence Based Medicine in the EMR Era, Jennifer Frankovich, M.D., Christopher A. Longhurst, M.D., and Scott M. Sutherland, M.D., N Engl J Med 2011; 365:1758-1759November 10, 2011

[7] *Id*

[8] *Id*,

Algorithm and Method for Automated Acquisition of Medical History

Howard Schneider [1], **Xie Li**[2]
[1]Sheppard Clinic North, Toronto, Ontario, Canada
[2]DocPod Corp, Toronto, Ontario, Canada

Abstract - *A successfully implemented algorithm and method of obtaining a medical history automatically from a patient is described. The main algorithm consists of a probe question/questionnaire, a coarse prediction algorithm, a fine prediction algorithm and a medical history generating system. The coarse prediction algorithm uses simple pattern matching against the medical knowledge database. The fine prediction algorithm uses a combination of pattern matching against the medical knowledge database, Bayesian inference where probability values so allow, and deep learning (both unsupervised and supervised as there is always human physician oversight of the system) where the number of patients is high enough to reasonably allow.*

Keywords: Automated Medical History, Patient Computer Interview

In all areas of medicine, a medical history is the first step in arriving at a diagnosis and subsequently providing treatment to the patient. Traditionally, healthcare providers ask patients questions and in documenting the encounter generate a medical history. There are many problems with generating a medical history that is accurate, comprehensive and economical. Ramsey and colleagues [1] showed that of 134 primary care physicians studied, the physicians only asked 59% of what would be considered essential questions in order to obtain an accurate history from the patient. Tang and colleagues [2] showed that physicians in ambulatory practices spent one-fifth of their day writing. Thus there is a high cost in relatively expensive physician time being spent on charting.

As early as the 1960's physicians were trying to use computers to solve the problems of generating a medical history that is accurate, comprehensive and economical [3]. Nonetheless, despite the advancements and improvements in computer technology over the decades, at the time of this writing, few physicians or other health care providers make use of automated medical history systems. Bachman in 2003[4] considers why physicians may not want to use computer-based interviewing, and in particular notes, "A computer program does not necessarily distinguish background symptoms from those leading to a visit to a physician. The physician, the patient, or the software needs to determine what is relevant."

Much of the research in artificial intelligence in medical diagnosis assumes that there is an existing data set of input data. However, to the clinician seeing patients day after day the main issue is not making a diagnosis but the large amount of time and effort and skill that is required to obtain this input data from the patient.

We describe a successful algorithm and method of obtaining this input data, ie, obtaining a medical history from a patient, that has emerged from our pilot project in automated medical history generation. In 2015 110 patients in a general psychiatry clinic used our pilot automated system which in turn generated semi-automated and automated medical psychiatric histories which were compared to medical psychiatric histories generated conventionally, and the algorithm and method incrementally improved with each cohort of patients.

The main algorithm the system used was as follows:

1. Obtain previous medical history
2. Obtain probe question answer from patient "Why are you here?"
3. Obtain medical knowledge database related to #1 and #2

4. From coarse prediction algorithm and #1, #2 and #3, ask patient next question
5. If response to question from #4 was not appropriate try #4 again
6. From fine prediction algorithm ask patient questions and obtain patient data (weight, blood pressure, lab values, etc) related to likely diagnoses.
7. If answers to #6 indicate another diagnosis repeat again from #4
8. Generate structured medical history appropriate for visit and likely diagnosis

The coarse prediction algorithm uses simple pattern matching against the medical knowledge database. The fine prediction algorithm uses a combination of pattern matching against the medical knowledge database, Bayesian inference where probability values so allow, and deep learning (both unsupervised and supervised as there is always human physician oversight of the system) where n (number of patients) is high enough to reasonably allow.

The system was implemented with a cloud architecture (Google App Engine) with physician portals and patient entry programs all running in web browsers independent of the underlying computer hardware. Due to local regulatory concerns no medical devices were directly interfaced to the system, but relied on the patient or the physician to input physical exam and lab values.

Future work includes comparing the time the physician must spend obtaining a medical history standardized to a certain level of quality and other time with the patient versus the time the physician needs to spend with the patient where an automated medical history is obtained, and comparing such comparisons against n (number of patients) for that diagnosis, with the expectation that time savings will increase as the system sees more patients of a given diagnosis.

Ethics Review Board Approval: Canadian SHIELD Ethics Review Board #14-03-002

References

[1] **Ramsey PG, Curtis JR, Paauw DS, Carline JD, Wenrich MD.** History-taking and preventive medicine skills among primary care physicians: an assessment using standardized patients. *Am J Med.* 1998;104:152-158.

[2] **Tang PC, Jaworski MA, Fellencer CA, et al**. Methods for Assessing Information Needs of Clinicians in Ambulatory Care. *Proc Annu Symp Comput Appl Med Care.* 1995; 630-634.

[3] **Mayne JG, Weksel W, Sholtz, PN.** Toward automating the medical history. *Mayo Clin Proc.* 1968 Jan;43(1):1-25.

[4] **Bachman JW.** "The Patient-Computer Interview: A Neglected Tool That Can Aid the Clinician. *Mayo Clin Proc.* 2003;78:67-78.

Algorithm and Method for Automated Processing of Medical E-mails

Howard Schneider [1], **Xie Li**[2]
[1]Sheppard Clinic North, Toronto, Ontario, Canada
[2]DocPod Corp, Toronto, Ontario, Canada

Abstract - *We describe an algorithm and method for an Intelligent Personal Assistant (IPA) for a health care practitioner such that the IPA can automatically schedule appointments with patients, automatically schedule appointments with colleagues and for other third party events, and automatically respond to routine communications from patients, from colleagues and from other third parties. The algorithm was implemented using a commercially available cloud architecture, commercial scheduling system, and commercial voicemail to e-mail conversion system. This algorithm and method has the potential to automate many of the personal assistant tasks required in a medical practice and to correspondingly lower the cost of health care.*

Keywords: Intelligent Personal Assistant, AI Assistant

E-mails are increasingly becoming a routine feature of medical practice. Medical e-mails may be received from patients [1], from other colleagues, from assistants, from laboratories, from medical marketing, as well as from non-medical sources. Time spent by the health care practitioner in responding to e-mails, is time that is in short supply and must be rationed with patients [2].

Intelligent Personal Assistants ('IPA's) are software agents that can provide some or many of the tasks an individual requires in managing one's life, or in the case of a health care practitioner, in managing a medical practice, such as scheduling time with patients, scheduling time with colleagues and third party contacts (eg, educational events), and responding to routine phone calls and e-mails. Work has been done on IPAs for the last several decades, with the resultant systems becoming more interactive and arguably useful as time has progressed [3,4,5,6]. However, at the time of this writing, no IPA commercially available can actually meet the 'reasonable personal assistant' needs of a health care practitioner, which we can define as follows:

1. Automatically schedule appointments with patients
2. Automatically schedule appointments with colleagues and other third party events
3. Automatically respond to routine communications from patients
4. Automatically respond to routine communications from colleagues and other third parties

We describe a successful algorithm and method of providing the above defined 'reasonable personal assistant' needs of a health care practitioner. Steps 5 and 6 of the algorithm below make use of another algorithm [7] which allows automated acquisition of medical histories.

The main algorithm the system uses is as follows:

1. Obtain Input Data: Receive e-mails or convert voicemails/phone calls to e-mail.
2. Context Awareness: An attempt to make assumptions about a sender's message (rather than the usual environmental interpretation of context, eg, Yan *et al* [8] versus Dey [9]) : Parse e-mail **k** times (initial value adjusted by a supervised learning algorithm depending on the health care practitioner's satisfaction of the result in the overall handling of a communication) and select the parsed version which maximizes a score with regard to the version being in the context of:
   -a patient requesting an appointment/change
   -or a colleague/third party requesting a meeting/change
   -or a communication from a patient requiring a medical response
   -or a communication from a colleague/third party requiring a medical response
   -or an 'other' e-mail (which includes e-mails explicitly stating that a human response is required)

The algorithm has access to the e-mail contact list of the health care practitioner as well as the medical records of the health care practitioner in order to best calculate this contextual score.

3. If #2 'Context Awareness'== 'patient requesting an appointment/change' then the following will occur:

i. If 'no urgency' then: An appointment will be given at the next available normal empty appointment slot for the health care practitioner. Feedback will be asked of the patient – if patient responds that the appointment is ok or does not respond then the algorithm terminates, or if patient responds that a different time or date is required, then an alternative appointment will be given to the patient and feedback is again asked of the patient.

ii. If 'urgency' detected in #2 then: The original e-mail will be forwarded to the health care practitioner, and the algorithm terminates.

4. If #2 'Context Awareness'== 'colleagues/third parties requesting an appointment/change' then the following will occur:

i. If 'no urgency' then: Similar to 3i.

ii. If 'urgency' detected in #2 then: Similar to 3ii.

5. If #2 'Context Awareness'== 'routine communications from patients' then the following will occur:

i. If the patient does not exist in the health care practitioner's medical records database then a message is sent to the patient that it is not possible to directly respond until the sender becomes an official patient of the health care practitioner and instructions how to do so.

ii. If a patient is an existing patient then the same probe questionnaire that is used in clinic by the automated medical acquisition history system (ie, 2016 Schneider and Xie [7]) is sent to the patient, and in response to the patient's responses follow-up questionnaires are sent until either an 'action plan/learning questionnaire' is sent to the patient or the questionnaire data and e-mail are forwarded to the health care practitioner for human response, and the algorithm terminates.

6. If #2 'Context Awareness'== 'routine communications from colleagues/third parties' then the following will occur: Similar to 5ii but a colleague/third party probe questionnaire is used to better determine the exact request of the colleague/third party.

7. If #2 'Context Awareness'== 'other e-mail' then the following will occur:

i. If an alternative external IPA system is being used, then the original e-mail is submitted to the alternative external IPA system, and this algorithm terminates.

ii. If no alternative external IPA system is being used, then the original e-mail is submitted to the health care practitioner for manual processing, and this algorithm terminates.

The system was implemented with a cloud architecture (Google App Engine) with a commercially available scheduling system (Google Calendar). A commercially available voicemail to e-mail conversion system was used (Vonage). At the time of this writing the system is being used in a general psychiatry clinic with most but not all modules above implemented. For example, there is no external IPA being used. At the time of this writing the system is able to replace greater than 70% of the labor spent by a human assistant in a full-time position in a similar role in a medical practice. This algorithm and method has the potential to automate many of the personal assistant tasks required in a medical practice and to correspondingly lower the cost of health care.

References

[1] **Skerrett PJ. "**Doctors debate use of email for communicating with their patients", *Harvard Medical School Harvard Health Blog*. Jan 26 2012. Retrieved from : http://www.health.harvard.edu/blog/doctors-debate-use-of-email-for-communicating-with-their-patients-201201264155

[2] **Dugdale DC, Epstein R, and Pantilat SZ.** "Time and the Patient-Physician Relationship", *J Gen Intern Med.* 14(Suppl 1): S34-S40 Jan 1999.

[3] **Erol K, Hendler J and Nau D.** "Semantics for Hierarchical Task-Network Planning", *Technical Report CS-TR-3239, Computer Science Department, University of Maryland* 1994.

[4] **Myers K, Berry P, Blythe J *et al*.** "An Intelligent Personal Assistant for Task and Time Management", *AI Magazine* v28(2) 2007.

[5] **Magee C.** "Meet Genee, Your Artifically Intelligent Personal Assistant", *TechCrunch*, Aug 12 2015. Retrieved from : http://techcrunch.com/2015/08/12/meet-genee-your-artificially-intelligent-personal-assistant/

[6] **Corbyn Z.** "Meet Viv: the AI that wants to read your mind and run your life", *The Guardian* Jan 31 2016. Retrieved from : https://www.theguardian.com/technology/2016/jan/31/viv-artificial-intelligence-wants-to-run-your-life-siri-personal-assistants

[7] **Schneider H and Xie L.** "Algorithm and Method for Automated Acquisition of Medical History", *2nd International Conference on Health Informatics and Medical Systems* 2016 (in press).

[8] **Yan C, Fan Z and Huang L**, "DRWS: A Model for Learning Distributed Representations for Words and Sentences", in Pham D and Park S (eds) *PRICAI 2014: Proceedings 13th Pacific Rim International Conference on Artificial Intelligence* pp 196-207, Springer, Switzerland, 2014.

[9] **Dey, AK.** "Understanding and Using Context", *Personal and Ubiquitous Computing* Vol 5(1):4-7, Feb 2001.

Utilizing the Google Project Tango Tablet Development Kit and the Unity Engine for Image and Infrared Data-Based Obstacle Detection for the Visually Impaired

Rabia Jafri[1], Rodrigo Louzada Campos[2], Syed Abid Ali[3] and Hamid R. Arabnia[2]
[1]Department of Information Technology, King Saud University, Riyadh, Saudi Arabia
[2]Department of Computer Science, University of Georgia, Athens, Georgia, U.S.A.
[3]Araware LLC, Wilmington, Delaware, U.S.A

Abstract: *A novel image and infrared data-based application to assist visually impaired (VI) users in detecting and avoiding obstacles in their path while independently navigating indoors is proposed. The application will be developed for the recently introduced Google Project Tango Tablet Development Kit equipped with a powerful graphics processor and several sensors which allow it to track its motion and orientation in 3D space in real-time. It will exploit the inbuilt functionalities of the Unity engine in the Tango SDK to create a 3D reconstruction of the surrounding environment and to detect obstacles. The user will be provided with audio feedback consisting of obstacle warnings and navigation instructions for avoiding the detected obstacles. Our motivation is to increase the autonomy of VI users by providing them with a real-time mobile stand-alone application on a cutting–edge device, utilizing its inbuilt functions, which allows them to micro-navigate independently in possibly unfamiliar indoor surroundings.*

Keywords: Obstacle detection, obstacle avoidance, Unity, Project Tango, blind, visually impaired.

1. Introduction

One of the major challenges faced by visually impaired (VI) individuals during navigation is detecting and avoiding obstacles or drop-offs in their path. RGB image and infrared data-based systems have emerged as some of the most promising solutions for addressing this issue; however, currently such systems fall short in terms of accurately localizing the user and providing real-time feedback about obstacles in his path.

The Project Tango Tablet Development Kit [1], recently introduced by Google Inc., is an Android device, equipped with a powerful graphics processor (NVIDIA Tegra K1 with 192 CUDA cores) and several sensors (motion tracking camera, 3D depth sensor, accelerometer, ambient light sensor, barometer, compass, GPS, gyroscope), which allow it not only to track its own movement and orientation through 3D space in real time using computer vision techniques but also enable it to remember areas that it has travelled through and localize the user within those areas to up to an accuracy of within a few centimeters. Its integrated infra-red based depth sensors also allow it to measure the distance from the device to objects in the real world providing depth data about the objects in the form of point clouds. The depth sensors and the visual sensors are synchronized, facilitating the integration of the data from these two modalities.

Our project aims to utilize the Project Tango tablet to develop a system to assist VI users in detecting and avoiding obstacles in their path during navigation in an indoors environment. The system will exploit the inbuilt functionalities of the Unity engine in the Project Tango SDK to create a 3D reconstruction of the surrounding environment and to detect obstacles in real-time. The user will be provided with audio feedback consisting of obstacle warnings and navigation instructions for avoiding the detected obstacles. Our motivation is to increase the autonomy of VI users by providing them with a real-time mobile assistive stand-alone application on a cutting–edge device, utilizing its inbuilt functions, which allows them to micro-navigate independently in possibly unfamiliar indoor surroundings.

The rest of the paper is organized as follows: Section 2 provides a brief overview of existing image and infrared data-based obstacle detection and avoidance systems for the VI. Section 3 describes the application and outlines the plan for its development and evaluation. Section 4 concludes the paper.

2. Related work

Newly emerging sensor technologies, such as Microsoft's Kinect, Occipital's Structure Sensor, and, most recently, Google's Project Tango Tablet Development Kit [1], are making it possible to exploit infrared light to extract 3D information about the environment without the need to install any equipment in the surroundings. Recent development work on obstacle detection for the VI has specially focused on Kinect, either utilizing the data from its depth sensor alone or from both its RGB and depth sensors. However, since Kinect is not designed to be a wearable or handheld device, affixing it to the body or clothing results in bulky and aesthetically unappealing contraptions; furthermore, the Kinect sensor needs to be connected to a

backend server making systems based on it vulnerable to communication and speed performance issues. The Project Tango tablet appears to have a distinct advantage over Kinect in that it is an aesthetically appealing, handheld, mobile device equipped with a powerful processor enabling it to execute computationally intensive code in real-time without the need to connect to a backend server. Moreover, it has several additional embedded sensors and in-built functionalities, which can be utilized for extending and improving the obstacle detection application in the future. A few preliminary applications for the Tango tablet have already been proposed for obstacle detection and avoidance for the VI: The system presented by Anderson [2] collects depth information about the environment, saves it in a chunk-based voxel representation, and generates 3D audio for sonification which is relayed to the VI user via headphones to alert him to the presence of obstacles. Wang et al. [3] cluster depth readings of the immediate physical space around the users into different sectors and then analyze the relative and absolute depth of different sectors to establish thresholds to differentiate among obstacles, walls and corners, and ascending and descending staircases. Users are given navigation directions and information about objects using Android's text-to-speech feature. However, both these applications need further development and are yet to be tested with the target users [4].

3. Application Development

The application is being developed for Google's Project Tango Tablet Development Kit which is a 7" Android-based tablet. It will utilize the Project Tango Unity SDK [21] to acquire a 3D reconstruction of the surrounding environment in the form of a mesh which is created and updated in real-time. A character object in the 3D reconstruction will represent the user's position in the real world. Obstacle warnings will be issued if the distance between the character object and any object in the 3D reconstruction becomes less than a certain threshold (initially, this will be set to 0.5 meters; however, this value may be modified or a user-controlled customization option may be provided based on the results of our interviews with the target users).

Feedback to the users will be provided by playing pre-recorded audio files via an open-ear bone conduction Bluetooth headset. The feedback will include warnings about approaching obstacles - potentially including some details about their sizes and their positions relative to the user - and navigation instructions to avoid the obstacles (bear right, bear left, etc.). For users with some residual vision, a visual display option may also be provided in addition to the audio output.

Adopting a user-centered design approach, we are planning to conduct semi-structured interviews with VI individuals in Athens, Georgia, USA in order to gain some insight into their preferences for the user interface of the application and to procure their suggestions for what features should be included. The results of the interviews will inform the design of the system during its development.

Once an initial prototype has been developed, the two main parts of the system - the obstacle detection component and the user interface – will be empirically evaluated. The performance of the obstacle detection component will be tested for obstacles of various sizes at various positions with respect to the user and under different lighting conditions. The user interface will be designed in accordance with the users' preferences (acquired via the interviews) and will then be evaluated by conducting usability tests with VI users.

A similar system, being developed in parallel for obstacle detection and avoidance for the VI using the Project Tango Tablet Development Kit, was introduced in [4]. However, this system employs a different approach for detecting the obstacles by directly segmenting the point cloud data acquired from the scene. We aim to eventually conduct a comparative evaluation of the proposed system with this one to study any differences in terms of speed, accuracy and general usability.

4. Conclusion

A novel image and infrared data-based application to assist VI users in detecting and avoiding obstacles in their path while independently navigating indoors has been proposed in this paper. The application will utilize the functionalities of the Unity SDK of the Google Project Tango Development Kit to provide an aesthetically acceptable, cost-effective, portable, stand-alone solution for this purpose. A user centered approach would be adopted for the design and development, with semi-structured interviews being conducted with VI users at the initial stages of the development cycle to inform the interface design and usability testing with the target users being carried out at later stages with the initial prototype in order to identify any usability problems and to better adapt the system to the users' needs.

References

[1]"Google Project Tango (https://www.google.com/atap/project-tango/)."

[2]D. Anderson. (2015, Navigation for the Visually Impaired Using a Google Tango RGB-D Tablet. Available: http://www.dan.andersen.name/navigation-for-the-visually-impaired-using-a-google-tango-rgb-d-tablet/

[3]"Tiresias: An app to help visually impaired people navigate easily through unfamiliar buildings. HackDuke 2015. Available: http://devpost.com/software/tiresias-ie0vum," 2015.

[4]R. Jafri and M. M. Khan, "Obstacle Detection and Avoidance for the Visually Impaired in Indoors Environments Using Google's Project Tango Device," in *The 15th International Conference on Computers Helping People with Special Needs (ICCHP 2016)*, Linz, Austria, 2016.

Exploring new interactions for querying a tuberculosis database

Octavio Hector Juarez-Espinosa; Eric Engle; Andrei Gabrielian
Computational Biology Biosciences Branch (OCICB) National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH)
Rockville, MD
Octavio.juarez-espinosa@nih.gov, eric.engle@niaid.nih.gov, andrei.gabrielian@niaid.nih.gov

***Abstract*—this paper describes a software prototype created to improve user queries of the NIAID TB Portals database. The goal is to take advantage of medical images, such as x-rays, as unique pieces of knowledge to search information in a database. The software is not meant to act as a diagnostic tool. Instead, the software investigates the idea that a user can query a system with tuberculosis (TB) data by selecting a single x-ray from the database or uploading a new x-ray, and using the image to search for similar x-rays. These search methods take seconds to compare thousands of images and enable real-time database retrieval. This prototype enables users to then explore and analyze the selected subsets of de-identified, anonymized and curated TB patient cases related to the similar x-rays.**

Keywords—tuberculosis; x-rays; images; similarity; data exploration

I. INTRODUCTION

With the advance of computational technologies such as, image processing, databases, and network technologies the medical databases available are growing exponentially. To deal with these large databases researchers and business have adopted new technologies such as big data, data mining, and machine learning. Other groups focus their work in technologies where the user interaction combined with automatic methods improve data exploration tasks. For example, a user can start exploring a database by interacting with a graph and searching for more detail by selecting parameters or moving slide bars to filter the data sets.

Having interactive and automatic tools to explore large datasets on specific diseases might lead to wise and more informed decisions in medicine. The NIAID TB Portals database with data on hundreds of patients with tuberculosis that contain x-rays, CT-scans, treatment plan data, sequencing data, and final outcome, researchers can get a better understanding of this desease.

Tuberculosis is a global problem; although in the USA the number of registered cases is minimal. There are some countries with high mortality with this disease[1]. "In 2014, 9.6 million people fell ill with TB and 1.5 million died from the disease [2]."

This software prototype is part of the NIAID TB Portals program that seeks to empower clinicians, academic researchers, and the health care industry with advanced solutions in bioinformatics, information technology, and genomics toward improving TB patient diagnostics and treatment. This project involves several organizations in Belarus, Moldova, Georgia, Romania, and Azerbaijan, forming a virtual team with NIAID [3].

II. USING IMAGE SIMILARITY FOR QUERYING THE SYSTEM

Content-based information retrieval (CBIR) is a technology that takes advantage of image properties to query a system. This area of research has been around for many years but there is no standard way to apply it to medical problems [4]. In [4], the authors enumerate the possible causes: the lack of productive collaborations between medical and engineering experts; the lack of effective representation of medical content by low-level mathematical features; the lack of thorough evaluation of CBIR system performance and its benefit in health care; and the absence of appropriate tools for medical experts to experiment with a CBIR application.

This poster is not presenting a new algorithm for image similarity; instead it is a response to the motivation to experiment with the possible ways to query a tuberculosis database based on image content.

ISBN: 1-60132-437-5, CSREA Press ©

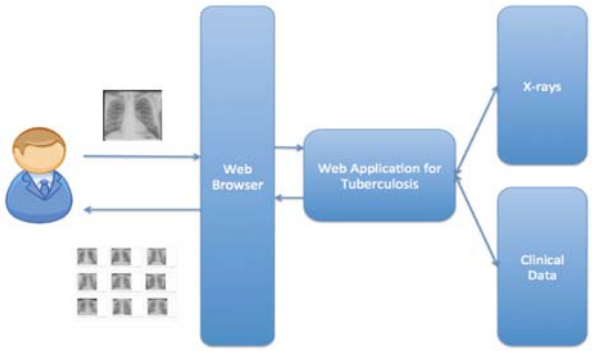


Figure1. Global view of the user interacting with the system

Two scenarios are implemented:

- The user uploads the x-ray and gets a set of similar x-rays as can be seen in Figure 1.
- The user selects an x-ray image from the database and retrieves similar images.

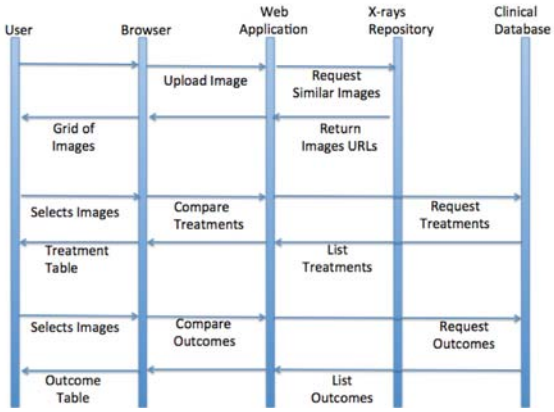


Figure 2. Example of sequences of user interactions

With this subset of similar images a user can ask more questions of the system. For example: compare the treatment plan for every one of the cases retrieved. Also the user might be interested in comparing the final outcome for the treatment as described in Figure 2. Finally, the user might want to see the evolution of tuberculosis in a specific patient by presenting the series of x-rays for specific case.

III. Our Implementation of CBIR System

The software prototype has been developed using cloud technologies and it has the following components: a web server that host the application and similarity index allocated to an EC2; a database with information about clinical data hosted on a EC2 instance; an image repository with 685 DCM files, which is hosted on a S3 instance; and a set of programs in Python and PHP that manage and respond user requests.

End-users are able to access the software using a web browser as shown in Figure 1.

A. Indexing Images

The pipeline to index the images consists of: preprocessing x-rays to improve contrast; segmenting [5]; and the extraction of features for every x-ray in the repository. The image features used for indexing consist of a histogram and the co-occurrence matrix [6].

B. Search Images

The user uploads the x-ray and the system computes the equalized histogram and segments the image. Next, the features are extracted. With the vector of features the system iterates over the index file and computes the distance to each x-ray. Finally, the software sorts the vector and returns a set of ten record identifiers and images with the best similarity scores.

IV. Future Work

This project will be improved by quantifying the impact of image-based queries in data exploration tasks.
Algorithms for indexing and searching images will also be improved using optimal features to represent the images.

Acknowledgment

This project is possible thanks to the guidance and leadership of Michael Tartakovsky, Chief Information Officer, and Alexander Rosenthal, Chief Technology Officer at NIAID.

References

[1] Stop TB Partnership. "Tuberculosis Profiles by Country." http://www.stoptb.org/countries/tbdata.asp

[2] World Health Organization. "Tuberculosis," last modified March 2016. http://www.who.int/mediacentre/factsheets/fs104/en/

[3] Belarus Tuberculosis Portal. http://tuberculosis.by

[4] Long, L. R., Antani, S., Deserno, T. M., & Thoma, G. R. (2009). "Content-Based Image Retrieval in Medicine: Retrospective Assessment, State of the Art, and Future Directions." *International Journal of Healthcare Information Systems and Informatics: Official Publication of the Information Resources Management Association*, 4, 1, 2009, pp. 1–16.

[5] Candemir, S. *et al*., "Lung Segmentation in Chest Radiographs Using Anatomical Atlases With Nonrigid Registration," *IEEE Transactions on Medical Imaging*, 33, 2, February 2014, pp. 577–590.

[6] Kovalev, V., and Petrou, M. "Multidimensional Co-occurrence Matrices for Object Recognition and Matching." *Graphical Models and Image Processing*, 58, 3, May 1996, pp. 187–197.

SESSION

LATE BREAKING PAPERS - HEALTH INFORMATICS AND HEALTH APPLICATIONS

Chair(s)

TBA

ISBN: 1-60132-437-5, CSREA Press ©

An AI model for Rapid and Accurate Identification of Chemical Agents in Mass Casualty Incidents

Nicholas Boltin[1], Daniel Vu[1], Bethany Janos[1], Alyssa Shofner[1], Joan Culley[2] and Homayoun Valafar[1*]

[1]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC

[2]College of Nursing , University of South Carolina, Columbia, SC

*Corresponding Author: homayoun@cse.sc.edu

Abstract – ***In this report we examine the effectiveness of WISER in identification of a chemical culprit during a chemical based Mass Casualty Incident (MCI). We also evaluate and compare Binary Decision Tree (BDT) and Artificial Neural Networks (ANN) using the same experimental conditions as WISER. The reverse engineered set of Signs/Symptoms from the WISER application was used as the training set and 31,100 simulated patient records were used as the testing set. Three sets of simulated patient records were generated by 5%, 10% and 15% perturbation of the Signs/Symptoms of each chemical record. While all three methods achieved a 100% training accuracy, WISER, BDT and ANN produced performances in the range of: 1.8%-0%, 65%-26%, 67%-21% respectively. A preliminary investigation of dimensional reduction using ANN illustrated a dimensional collapse from 79 variables to 40 with little loss of classification performance.***

Keywords: WISER, MCI, NLM, TOXNET, HSDB, Machine Learning.

1 Introduction

Improvement of the healthcare system in the United States is the subject of great interest and debate in the social, political, and economical arenas of our society. One obvious approach in improving the overall healthcare system is by eliminating the existing inefficiencies that impede our system[1–3]. Removal of inefficiencies impacts our system of healthcare in two inherent ways: significant improvement of the patient outcome, and a reduction in the cost of healthcare. Although in principle it is clear that removal of inefficiencies is beneficial, in practice there has been little effort in removal of the existing inefficiencies. This lack of effort is rooted in the complexity of our healthcare system that has manifested itself as a lack of consensus on the method of removing the existing inefficiencies.

Integration of technological advances in our healthcare such as utilization of mobile devices, availability of broadband systems with high throughput, and embedded clinical decision systems[4–6] can be cited as some approaches that can reduce overall inefficiencies of our healthcare system. One branch of healthcare that can benefit from better streamlining of patient-care through integration of clinical decision support is in emergency care during a mass casualty incident[7] (MCI). The rapid operational tempo of an Emergency Room (ER) serves as an ideal vehicle to study any existing inefficiencies while the resource-limited conditions of an MCI will help in clearly gauging the impact of any proposed improvements. MCI events clearly require rapid treatment of patients with minimum interruption for data collection, while optimal treatment of patients requires the hindering and cumbersome completion of detailed patient information to identify the culprit chemical. These two competing objectives have traditionally been a major impediment in optimizing the MCI treatment process with a natural priority extended to rapid treatment of patients. Therefore, there has been little advances in improving treatment of chemical MCI events. Research is needed to build a better understanding of the information and technological needs of the healthcare and public health workforce during emergency decision making[8].

A limited set of clinical decision support software have been introduced by the larger community[9]. The National Library of Medicine has created the Wireless Information System for Emergency Responders[10] (WISER), which allows emergency responders to identify a list of possible chemical substances based on the observed patient symptoms. The US Department of Health and Human Services has developed another software tool named the Chemical Hazards Emergency Medical Management-Intelligent Syndromes Tool[11] (CHEMM-IST). CHEMM-IST is a prototype that guides first-responders through a series of questions related to signs and symptoms that leads to a probabilistic diagnosis of four syndromes rather than a list of chemical hazards. Although such software make significant strides in assisting the process of emergency care, their efficacy have not been assessed during a chemical based MCI.

In this report we examine the effectiveness of WISER as the potential software for early identification of chemical material during an MCI event using simulated patient signs/symptoms (SSx) that we have reverse engineered from WISER. We also report results from Binary Decision Tree and Artificial Neural Network applications to the same set of simulated patient data. We conclude by reporting results of our initial investigation aimed at dimensional reduction of SSx space. Our final objective is to challenge the paradigm that rapid patient treatment is in contrary to data gathering that will assist in early identification of a culprit

chemical. We contest that careful design of sophisticated clinical decision support tools can satisfy both competing objectives of rapid information gathering and accurate chemical identification processes.

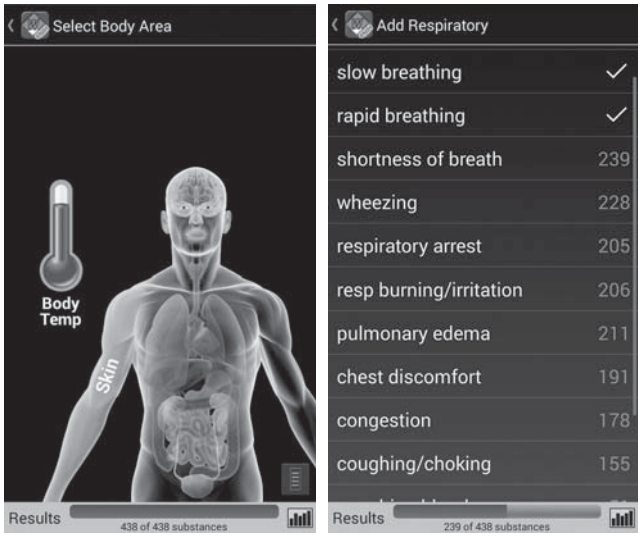
2 Materials and Methods

Our general approach consists of creating signs and symptoms (SSx) for simulated patients using a reverse-engineered table of SSx from the WISER application. Using the simulated data, we then proceed to evaluate the successful identification of a culprit chemical using WISER, Binary Decision Tree (BDT), and Artificial Neural Network (ANN) machine learning approaches.

2.1 WISER

Wireless Information System for Emergency Responders[10] (WISER) is a free application available for Android and iOS, which can also be downloaded as a standalone application on a desktop computer. Developed by the National Library of Medicine (NLM), WISER is a system designed to assist emergency responders in hazardous material incidents. It provides a wide range of information on hazardous substances, including substance identification support, physical characteristics, human health information and containment and suppression advice. Its key features include rapid access to the most important information about a hazardous substance by an intelligent synopsis engine and display called "Key Info", and access to NLM's Hazardous Substances Data Bank (HSDB), which contains detailed peer-reviewed information on hazardous substances and comprehensive decision support.

The key feature in WISER most relevant to this work is the Substance ID Support (SIDS). This allows an emergency responder to input patient SSx, from which the SIDS will identify one or more likely hazardous chemicals causing those symptoms. WISER contains a checklist of 79 SSx, which are input for selected systems of the body through an interactive tool as seen in Figure 1a. As the signs and symptoms are entered the pre-populated library of 438 hazardous substances is successively reduced (an example shown in Figure 1b). The user can view the list, select a substance and view toxicology information available in the HSDB, which contains data from the NLM Toxicology Data Network[12] (TOXNET). The HSDB data file contains information on human exposure, industrial hygiene, emergency handling procedures, environmental fate, regulatory requirements and related areas.



(a) (b)

Figure 1. Wireless Information System for Emergency Responders (WISER) for Android operating system. Panel (a) is the Interactive tool and panel (b) is the symptom selection interface. Panel (b) also shows the substance ID support in which an emergency responder can identify an unknown substance based on signs and symptoms of victims.

2.2 Reverse engineering and compression of WISER database

A thorough evaluation of WISER necessitated reverse engineering of all WISER's substances with their associated SSx. This task was performed by manually reviewing NLM's HSDB and parsing the SSx for each substance. An example of the resultant table of SSx is shown in Figure 2. Each of the 438 substances found in WISER is represented in the first column in this table, and the following 79 columns represent the corresponding SSx found in WISER for a given chemical. The presence or absence of each SSx is indicated by a 1 or a 0 respectively.

	A	B	C	D	E	F	G	H	I	J
1		abdom_di	abdom_di	agitation	arrhythmi	blistering	bloody_ni	bradycard	cardiovas	chest_
2	1,1,1-Trichloroethane	1	0	1	1	1	0	1	1	
3	1,1,2,2-Tetrachloroethane	1	0	1	1	1	0	1	1	
4	1,1-Dichloroethane	1	0	1	1	0	0	1	1	
5	1,1-Dichloroethylene	1	0	1	0	0	0	0	1	
6	1,1-Difluoroethane	0	0	1	1	0	0	1	1	
7	1,1-Difluoroethene	0	0	1	1	0	0	1	1	
8	1,1-Dimethylhydrazine	0	0	1	1	0	0	1	1	
9	1,2,4,5-Tetrachlorobenzene	0	0	1	0	0	0	0	0	
10	1,2-Dibromo-3-chloropropane	1	1	0	1	1	0	1	1	
11	1,2-Dichloroethane	1	0	1	1	0	0	0	1	
12	1,2-Dichloroethylene	1	0	1	1	0	0	1	1	
13	1,2-Dichloropropane	1	0	0	1	0	0	1	1	
14	1,2-Diphenylhydrazine	0	0	1	1	0	0	1	1	

Figure 2.WISER's reconstructed database using NLM's toxicology information stored in the Hazardous Substances Data Bank (HSDB).

Examination of the created database revealed several substances with identical SSx profiles. In such instances, a cluster of chemicals was reduced to a single representative. The list of uniquely distinguishable chemicals was then reduced from 438 substances to 311 unique substances, which serves as the reverse-engineered list of unique chemicals.

2.3 Creation of Simulated Victims (Test Set)

Simulated patient-data were generated from the ideal database of 311 unique substances by perturbation of randomly selected SSx. Although the use of experimental data is usually preferred over the use of simulated data in evaluation of any method, the use of simulated data is advantageous in some instances. For example the use of simulated data allows for evaluation of a method's performance as a precise function of the data quality and completeness. Furthermore, common conditions of limited availability of accurate and complete patient data[13,14] hinders development of clinical decision support systems, which can partially be overcome by the use of simulated data. Each substance was replicated 100 times to create a reasonably extensive testing set that consisted of 31,100 simulated victims. Three data-sets were created by random toggling of selected SSx at 5%, 10%, and 15% selection rates. To ensure the proper random selections, probability density profiles were examined for the number of perturbed SSx across each of the simulated patient-data. An overview of the perturbed data-sets (shown in Figure 3) corroborates the intended rates of perturbation.

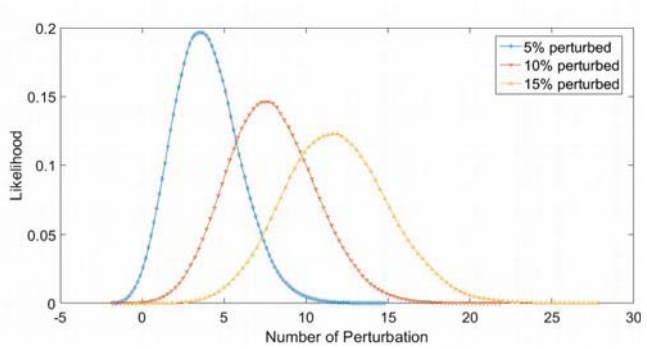


Figure 3.The Kernel Density Estimation of the 3 test data-sets. Test data were created by starting with the ideal table of symptoms from WISER and changing the symptoms by 5%, 10%, and 15%.

2.4 Overview of Machine Learning Approach

Our general work-flow for creating predictive models can be found in Figure 4. Supervised machine learning techniques were utilized in the Matlab 2015Rb environment to identify patterns and to develop predictive models. Our process began by importing the reverse-engineered database of 311 unique substances followed by training of two types of classification models: Binary Decision Trees (BDT) and Artificial Neural Networks (ANN). After successful training of a given model, the known SSx profile for all 311 substances was tested on the trained model to establish proper learning (testing for memorization versus generalization is conducted in a different step). The model with the highest accuracy during the training was chosen as the final model. Evaluation of each trained model was then assessed using the SSx profiles of the 31,100 simulated victims. A prediction accuracy was calculated as shown in Equation 2. In this equation A represents the accuracy of the model (expressed in %), N_c indicates the number of correctly identified chemicals, and N_{total} represents the total number of trials (31,100 in this case). The next sections provide a more detailed description of the training and testing for each model.

$$A = \left(N_c / N_{Total} \right) \cdot 100 \quad (1)$$

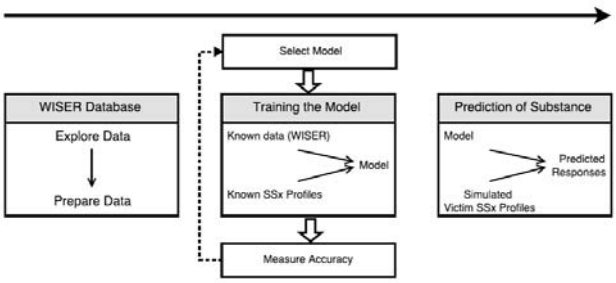


Figure 4.Work-flow for exploration of data, training models and predicting substances using supervised machine learning techniques

2.5 Training and Testing of Classification Methods

We evaluated three common classification approaches in our investigation. The classification approaches consisted of: database look-up (as implemented by WISER), Binary Decision Trees, and Artificial Neural Networks. Details for each of the three approaches are described in the following sections.

2.5.1 Database look-up (WISER)

The interactive nature of WISER was the limiting factor in automated and batch evaluation of WISER for 31,100 patients each represented by 79 SSx. This limitation served as one of our primary motivations in establishing a local database of WISER SSx. The first step in replicating a process identical to the WISER application was to understand its selection logic. WISER selects chemicals only based on the presence of a SSx and not its absence. Therefore, WISER will identify the entire library of 438 (or 311 unique) chemicals as the potential list of possible exposed chemicals for a patient exhibiting no apparent SSx. While this logic may appear questionable in our application, we proceeded with our evaluation of WISER in an exact fashion. Our initial evaluation of WISER consisted of a query-based search of our local database of chemicals using MySQL database engine housed on an Ubuntu LTS 14.04 server. This approach required a database look-up for SSx of all 31,100 simulated patients. Since the WISER approach may (and most likely will) return a list of potential chemicals, the database look-up step is followed by a search for existence of the right chemical in the list of returned chemicals. Although the time requirement of this evaluation mechanism was feasible (in the order of a week) for a list of 31,100 patients, it is an impractical approach for future investigations with larger data-sets in order to establish a

more thorough evaluation of the methods. Our most current approach consists of an in-house developed program to simulate this table look-up process. Our evolved approach returns the identical results that WISER would return while reducing the search time from months to seconds. Our testing process consisted of recording the number of times that the correct chemical was present in the list of returned chemicals similar to Equation 1.

Since WISER operates in a deterministic fashion, a statistical model of its performance can be developed. By assuming that every patient will undergo alteration of exactly *n* SSx, it can be argued that WISER's outcome should closely follow a success rate shown in Equation 2. This equation lists all of the possible perturbation of SSx that will result in removal of the correct chemical in WISER's resultant list. This equation can be simplified using the Binomial theorem as shown in Equation 2. Based on binomial distribution modeling of the WISER's outcome, a success rate of 6.25%, 0.4% and 0.02% can be expected for the cases of 5%, 10% and 15% perturbation of SSx.

$$r = 1 - \sum_{i=1}^{n} \binom{n}{i} p^{i} (1-p)^{(n-i)} = \frac{1}{2^{n}} \quad (2)$$

2.5.2 Binary Decision Tree

A Binary Decision Tree (BDT) was trained using the reverse-engineered WISER database within the Matlab 2015Rb environment. A maximum deviance reduction was used as the split criterion with 350 maximum splits. Each of the 311 chemicals was replicated 312 times to facilitate the construction of a complete tree and in consideration of Matlab's training algorithm. Under this training conditions, a classification rate of 100% was achieved.

Our adopted testing procedure consisted of observing the chemical identification accuracy of the trained network with the simulated patient-data. It is noteworthy that the trained BDT was based on ideal data while the testing was based on the perturbed data-sets (5%, 10% and 15% perturbation).

2.5.3 Artificial Neural Network

An Artificial Neural Network (ANN) was trained through the Pattern Recognition toolbox of the Matlab 2015Rb using back-propagation learning algorithm[15–17]. The unique set of 311 ideal chemical SSx were used during the training of the ANNs. The training set consisted of 5 identical replicas for each of the unique 311 chemicals (for a total of 1555 training patterns) in order to accommodate a random selection of the cross-validation and testing sets. The 1555 training patterns were randomly partitioned into 70% for training, 15% for cross-validation and 15% for testing. Numerous ANNs were trained and tested for selection of the optimal number of hidden neurons. Our investigation concluded 20 neurons as the optimal number of hidden neurons. The final trained ANN model exhibited cross-entropy results of: 4.4 for the training set, 12.7 for the cross-validation set, and 12.7 for the testing set. These outcomes correspond to: 0% error for the training set, 2.1% error for the validation set and 2.1% for the testing set.

To test the performance of the network with unknown data, the 31,100 simulated patient-data were used as inputs for the ANN trained with ideal chemical data.

3 Results and Discussion

3.1 Database look-up (WISER)

The results of WISER database look-up approach are shown in Table 1 and exhibit a reasonable correlation to the binary distribution model shown in Equation 2. The rapid decay in performance of WISER is easily expected. We use the results of WISER as the basis of comparison since it is the most prominent and existing mechanism.

Table 1: Prediction accuracy results from WISER testing using 31,100 simulated patient-data perturbed at 5%, 10% and 15%.

Data-set	Prediction Accuracy	Max	Min
5% Perturbed	1.8%	7%	0%
10% Perturbed	2.3×10^{-2}%	1%	0%
15% Perturbed	0.0%	0%	0%

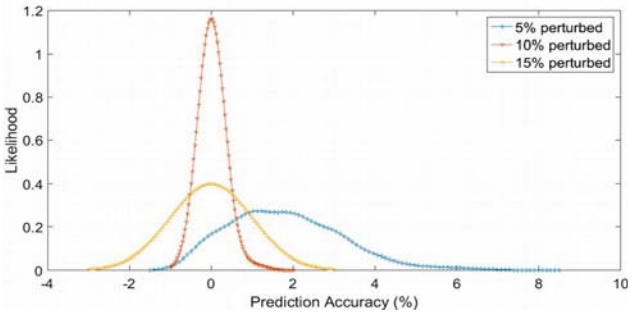


Figure 5: The Kernel Density Estimations from testing WISER with 31,100 simulated patient-data perturbed at 5%, 10% and 15%

3.2 Binary Decision Tree

Testing results for BDT are shown in Table 2. In this table the first columns represent the severity of the perturbation and the second column corresponds to the classification accuracy of the BDT. The third and fourth columns of Table 2 list the minimum and maximum performance across all of the 311 chemical substances. To better understand the performance of the BDT across the entire ensemble of 311 chemicals, a probability density function was created using the Kernel Density Estimation (KDE) technique[15,18]. Figure 6 illustrates the statistics for BDT classification behavior over the entire 100 representatives of each 311 chemicals. The nearly Gaussian distribution of the statics indicate a very well behaved system without any particular bias.

Another important factor to monitor during the construction

of a BDT is the topology of the final tree. Figure 7 illustrates the topology of the final tree (in the interest of simplicity the labels are omitted), which indicates a very well balanced tree of depth 9. This depth is in perfect theoretical agreement with the complexity of the problem, serving as another indication of a successful training session.

Table 2: Prediction accuracy results for Binary Decision Tree (BDT) testing using 31,100 simulated patient-data perturbed at 5%, 10% and 15%.

Data-set	Prediction Accuracy	Max	Min
5% Perturbed	64.9%	81%	53%
10% Perturbed	41.8%	54%	27%
15% Perturbed	25.6%	40%	13%

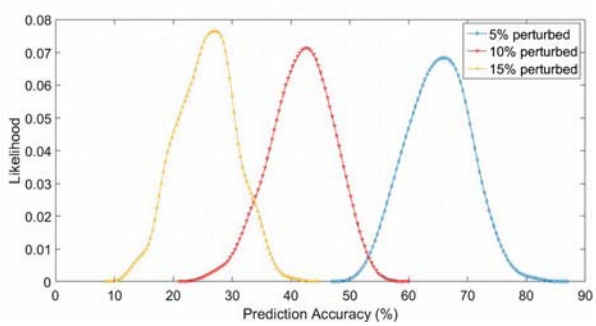


Figure 6.The Kernel Density Estimations from testing the Binary Decision Tree (BDT) model with 31,100 simulated patient-data perturbed at 5%, 10% and 15%.

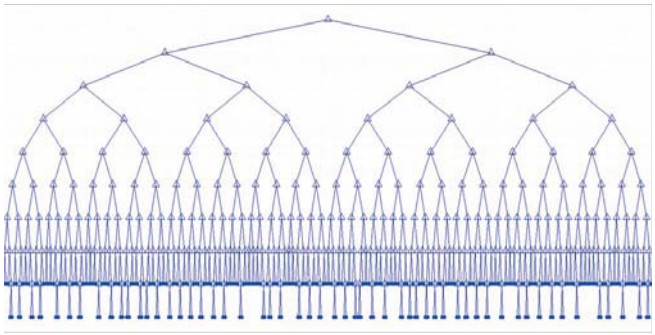


Figure 7.Static binary decision tree for 311 unique chemicals found in the National Library of Medicine's Hazardous Substances Data Bank (HSDB).

3.3 Artificial Neural Networks

The evaluation results of the ANN are shown in Table 3. Similar to the results of BDT, the first two columns of this table indicate the severity of perturbation and outcome accuracy, while columns three and four indicate the range of the outcomes across all 311 chemicals. Remarkably the accuracy of BDT and ANN appear to be similar, while the range of ANN's performance exhibit a larger variation. To better understand the statistics of ANN's results, probability density profiles were created for each of the experiments using KDE and using the exact parameters as the BDT (identical kernels). Similar to BDT, the Gaussian nature of the outcomes indicate a well behaved and unbiased system. Visual inspection of Figure 8 confirms the noted differences in variation of outcomes compared to the BDT results.

Table 3: Prediction accuracy results for the Artificial Neural Network (ANN) testing using 31,100 simulated patient-data perturbed at 5%, 10% and 15%.

Data-set	Prediction Accuracy	Max	Min
5% Perturbed	67.2%	96%	28%
10% Perturbed	38.4%	73%	10%
15% Perturbed	21.4%	49%	3%

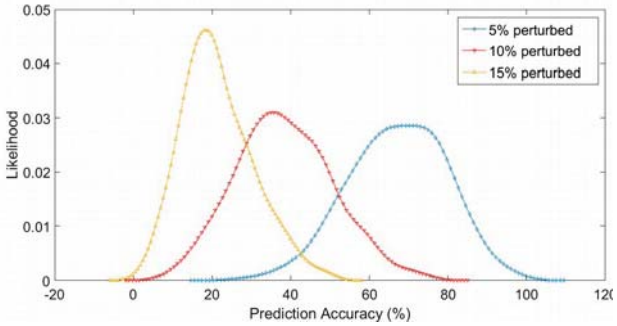


Figure 8.The Kernel Density Estimations from testing the Artificial Neural Network (ANN) model with 31,100 simulated patient-data perturbed at 5%, 10% and 15%.

3.4 Dimensional reduction

To optimize the Artificial Neural Network model, we examined the number of hidden neurons being used during the training phase of the model development. 10 models were trained, each with a different number of hidden neurons starting with 10 hidden neurons, then incrementing by 10 and the final model using 100 hidden neurons. After the model was created, additional testing was performed using the 5% perturbed data-set and the amount of error from the ANN was recorded. As seen in Figure 9, the results show that as we increase the number of hidden neurons, the amount of error from the ANN is reduced with the minimal amount of error being 15.4% at 100 hidden neurons. We then examined training the ANN with only the first 40 SSx instead of the complete database of 79 SSx. Again 10 models were trained starting with 10 hidden neurons at increments of 10 to 100 hidden neurons. After training the ANN, additional testing was also performed using the 5% perturbed data-set and recording the ANN error. It can be seen in Figure 9, the results followed the same pattern as with 79 SSx with the minimal amount of error being 25.8% at 40 hidden neurons. This indicates that using the first 40 SSx can reduce the amount of collected data with an acceptable reduction in the classification rate. This small reduction in classification can potentially be minimized through a more informed selection of SSx and analysis of the MCI over the entire cohort of victims.

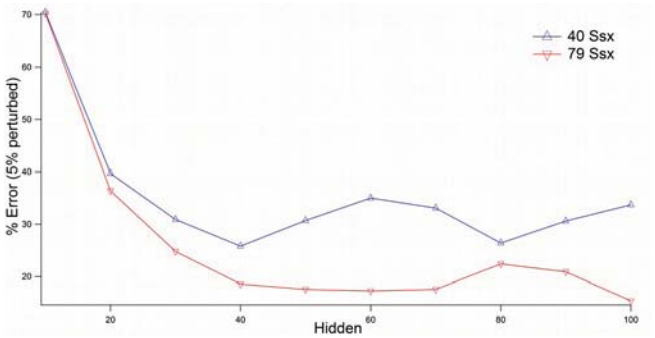


Figure 9: Optimizing the number of hidden neurons used in training the Artificial Neural Network. We used the 5% perturbed simulated patient-data for additional testing on the model.

4 Conclusion

Our overall approach consisted of evaluating WISER in application to an MCI under more realistic conditions. We have used the results of WISER as the basis of comparison to highlight the advantages and disadvantages of BDT and ANN, two common classification approaches in machine learning. The summary of results shown in Figure 10 illustrates the significant improved chemical identification performance that can be obtained from BDT or ANN compared to WISER. Results reported in section 3.1 (also summarized in Figure 10) reflect the intolerance of WISER to erroneous and imperfect data; a condition that is very likely to occur during the chaos and confusion that occurs during an MCI. Furthermore, WISER operates with a luxury of reporting a potentially long list of unrelated chemicals that share a common list of present SSx. Presenting a long list of unrelated chemicals may provide additional confusion during an MCI. However, creating a list of chemicals affords the benefit of operating with fewer SSx. Therefore, WISER exhibits the advantage of using as many or as little number of SSx as are available while BDT and ANN require a fixed number of SSx in their successful deployment.

Results for BDT and ANN evaluations reported in sections 3.2 and 3.3 highlight the significant robustness of these more sophisticated approaches compared to WISER. In summary, BDT and ANN show promise when compared to WISER for quickly and accurately identifying a culprit chemical during a chemical MCI. This gain in robustness is achieved through the use of these machine-learning techniques' ability to generalize and not simply memorize. Furthermore, BDT provides the clear advantage of arriving at a single chemical with requiring only 9 SSx (based on the depth of the tree shown in Figure 7). ANN exhibited the same degree of robustness compared to the BDT but with the apparent disadvantage of requiring all 79 SSx during the process of substance identification. However our exploration of dimensional reduction and results shown in section 3.4 support the possibility of using only 40 of the 79 SSx with little reduction in performance.

Our future investigations will focus on further reduction of data dimensionality by the use of previously established methods such and Principal Component Analysis (PCA) or Linear Discriminant Analysis[15] (LDA). AI tools employed during chemical MCIs could dramatically reduce the amount of information collected from patients resulting in increased accuracy, precision, and efficiency in identifying the chemical.

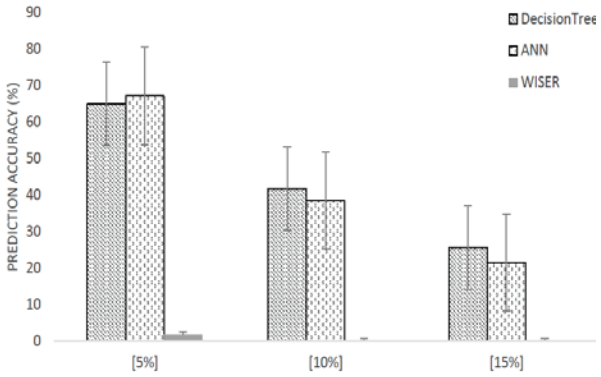


Figure 10.Overall prediction accuracy for each model tested with 31,100 simulated patient-data perturbed at 5%, 10% and 15%

5 Acknowledgments

This study was supported by the National Institutes of Health/The National Library of Medicine grant number 5R01LM011648.

6 Bibliography

1. Garber, A. & Skinner, J. *Is American Health Care Uniquely Inefficient?* (2008). doi:10.3386/w14257

2. Greene, W. Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the World Health Organization's panel data on national health care systems. *Health Econ.* **13,** 959–980 (2004).

3. Hackbarth, A. D. Eliminating Waste in US Health Care. *JAMA* **307,** 1513 (2012).

4. Garg, A. X. *et al.* Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA* **293,** 1223–1238 (2005).

5. Hunt, D. L., Haynes, R. B., Hanna, S. E. & Smith, K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. *JAMA* **280,** 1339–1346 (1998).

6. Kawamoto, K., Houlihan, C. a, Balas, E. A. & Lobach, D. F. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* **330,** 765 (2005).

7. Aylwin, C. J. *et al.* Reduction in critical mortality in urban mass casualty incidents: analysis of triage, surge, and resource use after the London bombings on July 7, 2005. *Lancet (London, England)* **368,** 2219–25 (2006).

8. Culley, J. Mass casualty information decision support. *OJNI* **15,** (2011).

9. Swain, C. WISER and REMM: Resources for Disaster Response. *J. Electron. Resour. Med. Libr.* **6,** 253–259 (2009).

10. Wireless Information System for Emergency Responders (WISER). at <https://wiser.nlm.nih.gov/>

11. Chemical Hazards Emergency Medical Management. at <https://chemm.nlm.nih.gov/chemmist.htm>

12. Toxicology Data Network (TOXNET). at <http://toxnet.nlm.nih.gov/>

13. Nie, H. *et al.* Triage during the week of the Sichuan earthquake: A review of utilized patient triage, care, and disposition procedures. *Injury* **42,** 515–520 (2011).

14. Haynes, B. E. *et al.* Medical response to catastrophic events: California's planning and the loma prieta earthquake. *Ann. Emerg. Med.* **21,** 368–374 (1992).

15. Greshenfeld, N. A. *The Nature of Mathematical Modeling*. (Cambridge University Press, 1998).

16. Holyoak, K. J. Parallel Distributed-Processing - Explorations in the Microstructure of Cognition - Rumelhart,De, Mcclelland,Jl. *Science (80-.).* **236,** 992–996 (1987).

17. Donahoe, J. W. & Palmer, D. C. Parallel Distributed-Processing - Explorations in the Microstructure of Cognition, Vol 1, Foundations - Rumelhart,De, Mcclelland,Jl, Pdp-Res-Grp. *J. Exp. Anal. Behav.* **51,** 399–416 (1989).

18. Fukunaga, K. *Introduction to Statistical Pattern Recognition*. (Academic Press, Incorporated, 1990). at <http://www.academicpress.com>

Application of Data Science to Discover the Relationship between Dental Caries and Diabetes in Dental Records

Germán H. Alférez[1], Jany Jiménez[2], Héctor Hernández Navarro[3], Merari González[4], Rusbel Domínguez[5], Alejandro Briones[6], Héctor Hernández Villalvazo[7]

harveyalferez@um.edu.mx[1], jany.jimenez@um.edu.mx[2], odontos@um.edu.mx[3], iscmera@gmail.com[4], ruxbel@gmail.com[5], alexbriones211191@gmail.com[6], hector.hdz.villalvazo@gmail.com[7]
Universidad de Montemorelos, Apartado 16-5,
Montemorelos N.L. 67500, Mexico

Abstract - *Diabetes is a chronic and metabolic disease. According to the World Health Organization (WHO), 422 million of adults suffer from diabetes worldwide. In fact, in 2012 diabetes caused 1.5 million deaths in the world. In Mexico, our country, diabetes is a highly relevant public health problem. For example, in 2015 there were 11 million cases of diabetes. Our contribution is to apply novel data science techniques to medical records at a dental clinic in Northeast Mexico to discover the relationship between diabetes and dental caries. Our research work follows IBM's data science methodology. The analysis of data was carried out with machine learning. Five experiments were performed on 193 dental records. Our findings corroborate the results in related work.*

Keywords: Diabetes, Dental Caries, Data Science

1) Introduction

Diabetes is a chronic and metabolic disease. It is characterized by high blood sugar levels that result from the deficit of the body to produce insulin. Insulin is the hormone that regulates the body glucose usage. Diabetes leads to damages in the heart, blood vessels, eyes, kidneys, and nerves.

According to the World Health Organization (WHO), 422 million of adults suffer from diabetes worldwide. In Mexico, our country, diabetes has been the first cause of death in women and the second in men since the year 2000.

Previous research work has shown that diabetes is closely related to dental caries since both of them share similar risk factors. For instance, diabetic patients that do not have control of their blood sugar levels have a higher risk of presenting systemic and oral complications [1]. One of these complications is dental caries. Dental caries is a multifactorial progressive process that can be developed on the tooth surface, inside the oral cavity, where the plaque allows it to grow on a period of time.

Diabetic patients are two or three times more susceptible to develop a periodontal disease than healthy patients [2]. Moreover, diabetes is the first cause of premature tooth loss, interrupting the physiological function of mastication, leading to a softer diet with a higher level of sugar that can cause dysglycemia [3].

Our contribution is to apply novel data science techniques to medical records to discover the relationship between diabetes and dental caries. Data science consists of analyzing data, structured and unstructured, to get knowledge [4]. Although the relationship between diabetes and dental caries has been found in related work, previous approaches are based on field studies with a limited number of patients. This may be caused by the lack of automatized techniques to analyze data in dentistry. Our approach goes a step further with the analysis of a larger number of patients' data by means of machine learning techniques.

Specifically, this work focuses on finding hidden patterns in 193 dental records of patients at the Dental Clinic "Luz y Vida" located at the campus of Universidad de Montemorelos, Montemorelos, N.L., Mexico. This study follows IBM's data science methodology [5]. In the experiments, the k-means algorithm of machine learning was executed in Weka. Seven features in the clinical records were analyzed.

This paper is organized as follows. The second section presents the description of the dataset taken from de clinical records at Dental Clinic "Luz y Vida". The third section presents how we have

applied IBM's methodology to find hidden patterns in the dataset to discover the relationship between dental caries and diabetes. The last section presents conclusions and future work.

2) Description of the Dataset

In this research work, the data was obtained from 193 clinical records from the Dental Clinic "Luz y Vida". This clinic is located at the campus of Universidad de Montemorelos, Montemorelos, N.L., Mexico. Since its foundation in 2004, this clinic has stored dental records in paper forms.

The clinical records used in this experiment belong to 193 patients located at the following cities: Montemorelos (140 patients), Monterrey (2 patients), Linares (4 patients), Solistahuacan (1 patient), Allende (9 patients), Mexico City (1 patient), Alamo (1 patient), Tampico (1 patient), Hermosillo (1 patient), Tamaulipas (1 patient), Coatzacoalcos (1 patient), Cancun (1 patient), Cd. Madero (1 patient), Tamasopo (1 patient), Altamira (1 patient), General Terán (3 patients), Anahuac (1 patient), Chihuahua (1 patient), Elkhart (1 patient), Santiago Tuxtla (1 patient), Camargo (1 patient), Cadereyta (1 patient), Chula Vista (1 patient), Mezcalapa (1 patient), Rayones (1 patient), Navojoa (1 patient), Georgia (1 patient), Riverside (1 patient), Caborca (1 patient). The location of 11 patients was not recorded in the dental records. Since this clinic has changed several times the paper forms used to record clinical records, we decided to take the sample for the experiments from the latest 193 clinical records, which use the same paper form.

3) Applying IBM's Methodology to Clinical Records

Data science requires a methodology that eases its application to industry and academy. That is why IBM offers a methodology for the application of data science [5]. This methodology is organized in ten stages structured in an iterative process. In the following paragraphs we describe the stages that were followed in this research according to IBM's data science methodology:

1. Understanding the problem: Dental caries related to diabetes is a hot topic in dentistry because there is not a clear and absolute position about the relationship between these two diseases. Moreover, related work tends to expose this relationship through field studies with a limited number of patients. In these studies, researchers take a sample of patients with diabetes and then look for the possible relationship between diabetes and dental caries. However, these studies do not propose the automatized analysis of dental records to facilitate the process.

For example, Seethalakshmi et al. [6] evaluated oral diseases that can be caused by diabetes, such as the incidence of dental caries and salivary pH in 20 patients. The results showed that not only periodontal health was affected, but also salivary pH had a decrease of 6.51. Likewise, the incidence of dental caries increased significantly in comparison with the patients without diabetes.

On one hand, Novotna et al. [7] mention that there is an increase of plaque levels and chronic gingivitis as much in adults as in children with type 1 diabetes. On the other hand, Miranda et al. [8] found out that oral health in patients with type 1 diabetes in Chile is more precarious than in healthy patients. However, they mention that this problem could be caused by poor hygiene.

Singh et al. [9] came to the conclusion that patients with type 2 diabetes have a higher risk of developing dental caries. Also, they pointed out that saliva flow and saliva calcium levels are significantly lower compared to healthy patients. Therefore a reduction in the saliva components reduces the enamels capacity to endure the remineralization and demineralization process. It creates the right environment for dental caries.

Oral bacteria are with no doubt a determinant factor in the formation of dental caries. For example, Kampoo et al. [10] found that the incidence in diabetic patients in Thailand is much higher compared to non-diabetics. Also, the number of acidogenic bacteria in diabetic patients is much higher than in healthy patients. Therefore, the high dental caries incidence in diabetic patients in Thailand is positively related to de Streptococcus and Lactobacillus bacteria.

Iqbal et al. [11] made a study to establish if there is a relationship between diabetes mellitus and dental caries by measuring glucose levels related to dental caries in different patients. These authors found out that glucose levels in diabetic patients' saliva are slightly higher than in healthy patients. Also, the levels of calcium in diabetic patients' saliva are lower.

Jawed et al. [12] found out that the level of blood sugar and glycosylated hemoglobin, and the number of decayed, missing, and filled teeth (DMF) is significantly higher in type 2 diabetics than in healthy patients. These results were obtained by a saliva sample and a DMF test (DMFT).

Similarly, Miko et al. [13] mention that the deficiency in glycemic control as well as the early

occurrence of diabetes can increase the risk of dental caries. This study was made with a DMFT applied to 259 teenagers with type 1 diabetes.

Stojanoviç et al. [14] studied the condition of type 2 diabetic patients related to metabolic control. The sample was composed by 47 type 2 diabetic patients randomly chosen and divided into two groups: poorly controlled diabetics and controlled patients. They found out that patients with a poor control of diabetes have a significant higher amount of dental caries compared to those that control the disease.

Hintao et al. [15] found that patients with type 2 diabetes, compared to healthy patients, have a higher risk of root surface dental caries. However, the prevalence and crown surface decay were not significantly different. Therefore, they concluded that type 2 diabetes is an important risk factor for root decay, but not for crown surface decay.

2. Analytic approach: In this stage, machine learning was used to analyze the data from the clinical dental records. Machine learning is a branch of artificial intelligence that consists of developing techniques that allow computers to learn by means of analyzing structured or unstructured data [16].

Weka[1] was used to analyze the dental records by means of machine learning. Weka is a data-mining software developed by the University of Waikato [17]. This software was programmed in Java and has powerful algorithms to extract information contained in datasets [18].

3. Data requirements: The clinical records at the Dental Clinic "Luz y Vida" have 60 features. In the field of machine learning, a feature is a variable that summarizes key aspects to be analyzed. In our case, the features contain data about personal information, anamnesis, and intraoral exploration (see Table 1).

4. Data collection: The clinical records analyzed in this study were in paper. Therefore, 15 students at the School of Engineering and Technology of Universidad de Montemorelos digitalized them. These students took a period of around two months in this process.

5. Data understanding: In this step, we decreased the number of features to 7. These features are the ones that we considered to be associated to dental caries and diabetes according to related work (see the related work presented in the first step). These features are as follows: 1) endocrine problems, including diabetes, family history of diabetes, thyroid gland problems, and others; 2) teeth problems, including sensitivity and bad habits (biting nails, thumb sucking, pencil biting, etc.); 3) number of decayed teeth; 4) number of missing teeth; 5) number of restored or filled teeth; 6) age; and 7) blood type.

6. Data preparation: In this stage, the data related to each feature in the clinical records was converted to numbers. This process was necessary because Weka requires numeric values to do the analysis. The dataset with the studied clinical records is available online[2].

7. Modeling: Among the machine learning algorithms that Weka provides, we chose k-means to generate clusters that model the relationship between dental caries and diabetes in the clinical records. k-means is an unsupervised-learning algorithm that allows to group data in clusters by discovering their centroids [18]. In k-means each sample inside the dataset must be included in one of the clusters [17]. We decided to use this algorithm because of the following reasons: 1) data in the clinical records was not labeled; and 2) k-means is a very popular and effective clustering algorithm [19].

8. Deployment: The experiments were executed with a different number of clusters (k from 4 to 7). We chose the results from the experiments with the minimum within cluster sum of squared errors (sum of distance functions of each point in the cluster to the k center). The results were iteratively presented to the team of dentists to get their feedback and avoid results that were not congruent with their domain of knowledge.

As a way of illustration, Figure 1 shows one of the results in the experiments. First, this figure shows the number of iterations in the experiment (11 in this case) and the within cluster sum of squared errors (21.55317619018957). This image also shows the selected features for each test (21, 30, 45, 54, 56 and 57. Table 1 shows the descriptions of each one of these features), the number of clusters (k = 6), and the number of instances or samples contained in each cluster (e.g. 28 instances in Cluster 0).

The results of the 5 experiments that were conducted on the 193 clinical records are described as follows:

[1] Weka: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

[2] Our dataset: https://docs.google.com/spreadsheets/d/1prAv_cj6nFpZejotYje4gpfqjT291vM_tpu9cLTZSGg/edit?usp=sharing

Table 1. Description of dental records

Patient Information	
No.	*description*
1	age
2	gender
3	city
4	state
5	marital status
6	occupation
7	medical service
8	religion
9	health status

Anamnesis			
No.	*description*	*No.*	*description*
10	under medical care	28	other diseases (e.g. asthma, cancer, etc.)
11	hospitalized or sick	29	heart disease
12	excessive bleeding	30	endocrine disease
13	pregnant	31	bones or muscles disease
14	if pregnant, date of birth	32	digestive disease
15	smoking	33	urinary disease
16	alcoholic beverages	34	allergies
17	more than two drinks per day	35	anesthesia reaction
18	drug use	36	problem with dental treatment
19	frequently tired	37	nerves dental treatment
20	skin disease	38	last dental visit
21	eyes disease	39	mouth disease
22	ears disease	40	teeth disease
23	nose disease	41	brushing teeth
24	throat disease	42	toothbrushing
25	nervous system disease	43	flossing
26	respiratory disease	44	fluorine usage
27	blood disease	45	current medicines (e.g. antibiotics, nitroglycerin, aspirin)

Intraoral Scan (Normal or Abnormal)	
No.	*description*
46	lips
47	tongue
48	corners of lips
49	soft palate
50	hard palate
51	mucous membrane
52	floor of the mouth
53	salivary glands
54	number decayed teeth
55	number of teeth with pain
56	number of missing teeth
57	number of restored teeth
58	restoration status
59	periodontal status
60	occlusion status

- **First experiment:** The objective of this experiment was to verify that diabetic patients also present problems in their teeth. 8 clusters were analyzed in this experiment. We found out that diabetic patients tend to present teeth loss and food accumulation in some zones. The within cluster sum of squared errors in this experiment was 2.5.

- **Second experiment:** The objective of this experiment was to analyze the number of teeth with caries in diabetic and healthy patients. We found out that diabetic patients tend to present 9 to 17 teeth with caries, whereas healthy patients tend to present between 1 and 9 teeth with caries. The within cluster sum of squared errors in this experiment was 6.8. The data was organized in 4 clusters.

- **Third experiment:** The objective of this experiment was to analyze the following features: endocrine disease, number of decayed teeth, number of missing teeth, and number of restored teeth. The data was organized in 6 clusters and the within cluster sum of squared errors in this experiment was 14.3. The analysis of these features showed that diabetic patients tend to present between 9 and 17 decayed teeth, a lower number of teeth, and a lower number of restored teeth than healthy patients.

- **Fourth experiment:** The objective of this experiment was to analyze the following features: age, blood type, if the patient presents endocrine problems, and total number of teeth decay. On one hand, we found out that there is a tendency in patients around 23 years old to not present diabetes and to have around 7 teeth decay. On the other hand, patients around 56 years old tend to easily have gum bleeding and tend to be diabetics or have a relative with diabetes. Also, these patients present around 9 decayed teeth. As the age increases, the number of decayed teeth also increases. The within cluster sum of squared errors in this experiment was 11.96. The data was organized in 7 clusters.

- **Fifth experiment:** The objective of this experiment was to analyze the following features: eyes disease (since diabetes can cause vision problems), endocrine disease, if the patient takes medicines, and the number of decayed, missing, and restored teeth. We found out that healthy patients tend to have between 7 and 9 decayed teeth. In this experiment the error range was 21.5. The data was organized in 6 clusters. This finding supports the result obtained in the second experiment: healthy patients have a lower tendency to have dental caries than diabetic patients.

Figure 1. Sample result of an experiment in Weka

kMeans
======

Number of iterations: 11
Within cluster sum of squared errors: 21.553176719018957
Missing values globally replaced with mean/mode

Cluster centroids:

		Cluster#					
Attribute	Full Data	0	1	2	3	4	5
	(193)	(28)	(52)	(42)	(38)	(18)	(15)
21	0.2487	1	0	0.0714	0.0263	0.6667	0.2667
30	1.1606	0.5357	0.0769	3	1.0263	1.2222	1.2
45	1.8152	0.3797	0.4089	1.881	0.1004	12.7222	0.442
54	9.2721	8.464	7.0242	7.8179	16.7368	8.0502	5.2
56	3.1322	3.7972	3.1648	3.0091	2.1227	3.6217	4.0931
57	3.0909	2.539	2.5087	2.1104	2.012	2.4394	12.4

9. Evaluation: The evaluation of the experiments' results was made by a team of dentists at the Dental Clinic "Luz y Vida". This team analyzed each one of the results obtained in the previous step. Based on the presented results, they concluded that the clinical record analysis of patients' data trough data science corroborates the existence of the relationship between dental caries and diabetes. This results supports the findings of related work in this area. For example, according to Seethalakshmi et al. [6] and Singh et al. [9], patients with type 2 diabetes have a higher risk to develop dental caries. Also, Kampoo et al. [10] point out that the number of acidogenic bacteria in the mouth of diabetic patients is much higher than in healthy patients.

10. Feedback: In this step, the team considered to make further experiments to include a higher number of clinical records. To this end, the team came to the conclusion of the need for counting on a software tool to record clinical records to avoid the time-consuming process of digitalizing records in paper forms.

4) Conclusions and Future Work

This paper presented the application of data science to discover the relationship between dental caries and diabetes in dental records. Our results corroborate the relationship between diabetes and dental caries found in related work. This study opens an unexplored field in dentistry: the application of data science, based on a formal methodology and machine-learning techniques, to find hidden patterns in clinical records.

As future work, we are going to build a software tool to store and manage the clinical records of the Dental Clinic "Luz y Vida". The objective of this project is to reduce time when capturing and analyzing patients' data. Moreover, the features in the dataset will be extended with more features related to patients' lifestyle (e.g. exercise and nutrition habits). Also, we are going to carry out more experiments on this clinic's data about other diseases, which could be related to other dental pathologies.

References

[1] Sutthavong S, Sangasapaviriya A. Oral Health Care in Systemic Diseases. New York: Nova Science Publishers, Inc; 2009.

[2] Navarro AB, Faria R, Bascones A. Relación entre diabetes mellitus y enfermedad periodontal. Avances en Periodoncia. 2002;14(1):9-19.

[3] Juárez RP, Chahín, JR, Vizcaya MM, Arduña EI. Salud oral en pacientes con diabetes tipo 2: caries dental, enfermedad periodontal y pérdida dentaria. Odontología Sanmarquina. 2007;10(1):10-3.

[4] Dhar V. Data science and prediction. Commun ACM. 2013;56(12):64-73.

[5] IBM. Foundational Methodology for Data Science. Available from https://public.dhe.ibm.com/common/ssi/ecm/im/en/imw14824usen/IMW14824USEN.PDF

[6] Seethalakshmi C, Reddy RCJ, Asifa N, Prabhu S. Correlation of Salivary pH, Incidence of Dental Caries and Periodontal Status in Diabetes Mellitus Patients: A Cross-sectional Study. Journal of Clinical

& Diagnostic Research. 2016;10(3):12-4.
[7] Novotna M, Podzimek S, Broukal Z, Lencova E, Duskova J. Periodontal Diseases and Dental Caries in Children with Type 1 Diabetes Mellitus. Mediators of Inflammation. 2015;2015:1-8.
[8] Miranda X, Troncoso J, Rodríguez C, Aravena P, Jiménez del R P. Caries e índice de higiene oral en niños con diabetes mellitus tipo 1. Revista Chilena de Pediatría. 2013;84(5):527-31.
[9] Singh I, Singh P, Singh A, Kour R. Diabetes an inducing factor for dental caries: A case control analysis. J Int Soc Prev Community Dent. 2016;6:125-9.
[10] Kampoo K, Teanpaisan R, Ledder RG, McBain AJ. Oral bacterial communities in individuals with type 2 diabetes who live in southern Thailand. Applied And Environmental Microbiology. 2014;80(2):662-71.
[11] Iqbal S, Kazmi F, Asad S, Mumtaz M, Khan AA. Dental caries & diabetes mellitus. Pakistan Oral & Dental Journal. 2011;31(1):58-61.
[12] Jawed M, Shahid SM, Qader SA, Azhar A. Dental caries in diabetes mellitus: role of salivary flow rate and minerals. Journal of Diabetes and its Complications. 2011;25(3):183-6.
[13] Miko S, Ambrus SJ, Sahafian S, Dinya E, Tamas G, Albrecht MG. Dental caries and adolescents with type 1 diabetes. British Dental Journal. 2010;208(6):E12-E.
[14] Stojanović N, Krunić J, Cicmil S, Vukotić O. Oral health status in patients with diabetes mellitus type 2 in relation to metabolic control of the disease. Srpski Arhiv Za Celokupno Lekarstvo. 2010;138(7-8):420-4.
[15] Hintao J, Teanpaisan R, Chongsuvivatwong V, Dahlen G, Rattarasarn C. Root surface and coronal caries in adults with type 2 diabetes mellitus. Community Dentistry And Oral Epidemiology. 2007;35(4):302-9.
[16] Camargo-Vega JJ, Camargo-Ortega, J. F., & Joyanes-Aguilar, L. Conociendo Big Data. Facultad de Ingeniería. 2015;24(38).
[17] Rivero D, Salgueiro-Silicia, Y, Domínguez, R. Valuation of several learning techniques in the Weka software. Innovación Tecnológica. 2012;18(3).
[18] Bogorny V, Avancini H, de Paula BC, Kuplich CR, Alvares LO. Weka-STPM: A Software Architecture and Prototype for Semantic Trajectory Data Mining and Visualization. Transactions in GIS. 2011;15(2):227-48.
[19] Cambronero CG MI. Algoritmos de aprendizaje: knn & kmeans. Inteligencia en Redes Comun Univ Carlos III Madrid 2006.

Security and Privacy in Mobile Applications

Azene Zenebe[1], and Karishma Thakkallapally[2]
[1]Managment Information Systems, Bowie State University, Bowie, MD, USA
[2]Managment Information Systems, Bowie State University, Bowie, MD, USA

Abstract - *The implementation of mobile technologies in every aspect of life has become phenomenal and started playing a vital role in healthcare industry. In accordance with the growth in information and communication technologies, mobile devices play a vivacious role in medical field where patient's health status is reported to the doctors on timely basis and also during emergency situations. To dodge this, it is very important that any mobile technology related to healthcare stay safe and secure while remaining HIPAA (Health Insurance Portability and Accountability Act) compliant. However, mHealth apps may contain significant risks to the privacy and security of patient's protected health information. This paper presents the results of an exploratory study by analyzing 50 top rated mHealth apps with respect to their security and privacy using the three-dimensional Model for Classifying mHealth Apps in terms of Security and Privacy Concerns proposed by Plachkinova, Andres and Chatterjee in 2015. We found that majority of the applications are prone to security and privacy threats and challenges.*

Keywords: *Security, Privacy, Mobile health apps, Threats*

1 Introduction

The World Health Organization defines mobile health (mHealth) as "the spread of mobile technologies as well as advancements in their innovative application to address health priorities." With the increase in use of smartphones, people are now interested to share their information by for healthcare purposes by using a variety of mobile health applications. Usage of mobile phones has increased significantly over the years, which lead to the development of mHealth applications. Remedy Health Media defines "Mobile health applications (mHealth apps) are software programs that offer health associated facilities for mobile phones and tablets" [1]. The use of mHealth apps exploded with the increased usage of smartphones. Mobile devices are playing a vital role in transforming health care into a more-efficient, patient-centered system of care in which individuals have instant access to their electronic medical records.

Although mHealth is innovative and beneficial to both physicians and patients, there are many challenges to overcome. Privacy and security in healthcare sector encompasses a procedure that needs to be followed in order for protecting the patients, providers, organizations, and vendors. But the process is complex because of the outside impacts, such as regulation, policies, crime, and technology. This exploratory research attempts to answer the following questions:

- To what extent the mobile health applications are prone to security challenges?
- To what extent the mobile health applications are prone to privacy threats?

The paper has five sections. Section 1 presents the introduction. Section 2 presents the literature review. Section 3 is the research methodology. Section 4 will discuss the results of the findings, followed by conclusion in Section 5.

2 Literature Review

Schulke states that, the increase of free and paid mobile health applications in market might pose health risks to individuals due to a deficiency of health professional involvement in the development of the apps. And he classified mHealth applications into two broad classes i.e., provider-focused and patient-focused [2].

David Collins [3], Senior Director of mHIMSS, mentioned that, some applications of mHealth are most inspiring as they focus on providing easy access to individuals with better care. For example, Collins references CallDR, a mobile app which is developed to streamline the consultation process by allowing physicians to securely send patient data which includes photos & video and engaging in real-time associations with specialists anywhere in the world via a mobile device [3].

Faudree stated that the mHealth applications usage amongst healthcare providers and individuals might take along major issues, such as security and privacy challenges [4]. According to Kharrazi, Chisholm, VanNasdale and Thompson, data security and privacy are major concerns for personal health records, and the consequences can be momentous if healthcare providers are not capable to offer adequate safeguards to individual privacy [5].

McCarthy [3] stated that poorly protected individual data in mHealth apps is one of the major concern that should be taken care of. She also reported the results of a study that was done on 43 health and fitness apps. Out of these, only 74% of the free apps and 60% of the paid apps had a privacy policy which are available in the app stores. On the other hand, only 48% of the paid apps and 25% of the free apps informed the individuals about the privacy policy. Moreover, not any one of the free apps and hardly any of the paid apps encrypted the data that individuals entered in the applications. Hence,

mHealth apps that do not encrypt individual data can bring a threat to data privacy [6]. In Healthcare Business Technology, Nasiri[7] reported data privacy risks in mobile health apps. In his study, he found that many individuals use mHealth applications to communicate with their healthcare providers, besides tracking and managing symptoms and other info. And he also found that the information shared between individuals and others might carry privacy risks. He reported the results of a research done on 20 mHealth apps which has been selected from the most popular free mHealth apps and found that 50% of them send data to third-party advertisers and 39% of them send data to anonymous parties without using any data encryption techniques [7].

The literature search reveals that there is lack of research in identifying the level of the security challenges and privacy threats that exist in the mHealth apps.

3 Methodology

This research adopts a taxonomy of mobile health applications shown in Figure 1 proposed by a by Plachkinova, Andres, and Chatterjee in the year 2015 [8. This taxonomy allows mHealth apps to be categorized first, and then the various types of privacy threats and security challenges of mHealth apps are studied. For each mHealth app, three variables or dimensions are consider: degree of relevance of the security challenges, degree of relevance of the privacy c threats, and degree of membership to different categories of mHealth apps based upon their functions or purposes. The possible values are full, partially, or none.

The mobile health applications are categorized into four different types (Plachkinova, Andres, and Samir Chatterjee, 2015): Wellness Applications are various types of wellness applications but formal programs typically include preventive measures, and surveillance for common diseases; Fitness Applications focus on the state or condition of being fit, especially good physical condition resulting from exercise and proper; Medication Reminder Applications organize all the medications and vitamins in one place that helps the users intended for medical purposes by proving alerts/notifications based on their schedule; and Patient Care & Monitoring Applications provide the opportunity of seeing patients via mobile devices.

The security dimension refers to the security challenges associated with the apps, and it has four categories [8]:

a. Authentication and Authorization(AA) refer to the security challenges associated with the need for proof of identity to access an apps as well as the right to carry out a certain activity using the apps.

b. Integrity and Accountability (IA) that refer to the security challenge of information may be altered when exchanged in an insecure network. Furthermore, accountability refers to the ability to identify misuse by an individual.

c. Availability and ease of use (AE) refers to the challenges associated with loss of availability and difficulty of using the apps, respectively.

d. Confidentiality and Management of security (CM) refers to the security challenges associated with ensuring that information is available only to those who are authorized to access it and not shared inappropriately, and the need to ensure the normal flow of operation and information involved, respectively.

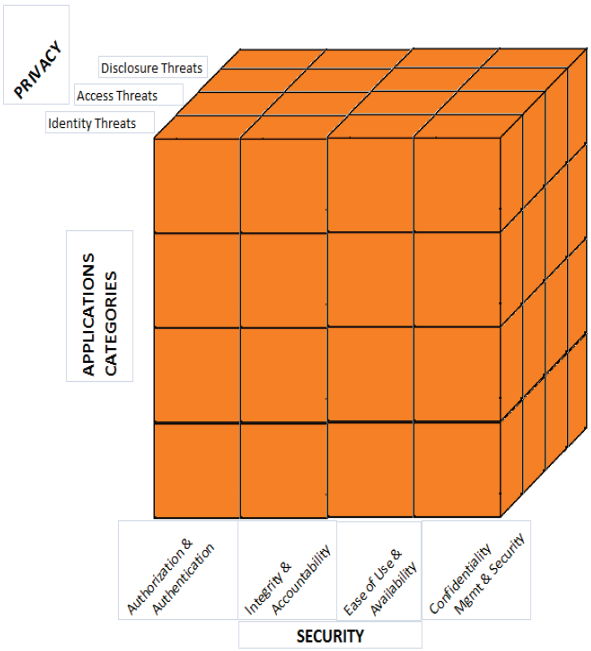


Fig 1. A Three-dimensional model for classifying mHealth Apps in terms of Security and Privacy Concerns (proposed by Plachkinova, Andres, and Chatterjee [8]*)*

The privacy dimension has three categories [8]:

a. Identity Threats (IT) are related to patients losing or sharing their identity credentials, thus enabling others to access their PHR. Also, insiders may use the credentials for medical fraud, potentially with financial or medical damage to the patient.

b. Access Threats (AT) are related to the fact that broader than intended access can be granted, insiders may share patient data leading, or health records can be modified.

c. Disclosure Threats (DT) are related to unauthorized disclosure of patient data.

Data Collection

Evaluation of the apps with respect to the relevance of the security challenges and privacy threats is done using available information and through testing them by the authors. They are rated as fully presence (1), partially presence (2), non-existence (3), or cannot be determined (7). For the security dimension, fully presence (1) means the security challenges, with respect to AA, IA, EA, and CM, are fully relevance to the mHelath apps. Partially (2) means the apps have partial relevance for the four security challenges: AA, IA, EA, and CM. No security threat (3) means the apps does not have any

relevance for the four security challenges: AA, IA, EA, and CM.

For the privacy dimension, fully relevance (1) means the apps have full relevance privacy threats with respect to the three privacy goals: DT, AT and IT. Partially (2) means the apps provides partial relevance from three type of privacy threats: DT, AT and IT. No privacy feature (3) means the apps does not have any relevance to the privacy threats: DT, AT and IT. When there is no sufficient information to make the determination of level relevance of security challenges or privacy threats that exist we use "cannot be determined" coded as 7.

In this study, secondary data for the 38 mobile health applications from an article by Plachkinova and et al. [8] has been considered, and new data for 12 mobile health applications was generated applying the taxonomy by the authors.

4 Results

Out of the 50 apps, 48% are Wellness apps (WA), 20% are Patient Care & Monitoring apps (PCMA), 18% are Fitness apps (FA), and 14% are Medication Reminder apps (MRA). *Table 1* represents the percentage of level of security challenge of the mobile applications. 52% of the apps do not have any security challenges with respect to AA; about 16% and 24% of the apps have high and partial security challenges with respect to AA, respectively. Furthermore, 60% and 64% of the mHealth apps do have high security challenges w.r.t. EA and IA respectively.

Table 1.Level of Security Challenges of mHealth apps

Rating	CM	AA	IA	EA
Fully	18.0	16.0	64.0	60.0
Partially	12.0	24.0	24.0	40.0
Does not exist	60.0	52.0	8.0	0.0
Cannot be determined	10.0	8.0	4.0	0.0
Total	100	100	100	100

Table 2 represents the percentage of privacy threats to mHealth applications for the different types of privacy threats. 46% of the mHealth apps do fully or partially susceptible to IT, about 12% and 34% of the apps have fully and partially open to AT, respectively. Furthermore, 58% of the mHealth apps are fully or partially susceptible to DT.

Table 2. Level of Privacy Threats to mHealth apps

Rating	IT	AT	DT
Fully	12.0	16.0	20.0
Partially	34.0	30.0	38.0
Does not exist	46.0	46.0	34.0
Cannot be determined	8.0	8.0	8.0
Total	100	100	100

5 Conclusion

The study focused on the important data security and privacy concerns associated with the use of mHealth apps. Even though some providers give terms of use and privacy policies when downloading the app, there is not yet an adequate way to provide level of security challenges associated with and privacy threats that exist in mHealth apps to users. Even after installing a mHealth app, it is sometimes not clear what data are collected, how data are managed and who has access to them. By using the taxonomy model, we able to conduct analysis of 50 mHelath apps, and found that very few number of mHealth apps are have no security challenges and are not susceptible to to privacy threats.

There are two practical implications of the results. First, developers of mhealth apps should find a way to indicate what and extent of the security and privacy threats that exist with the mobile health applications and the associated risks. Second, users should be aware of what and extent of the security and privacy threats, and risks associated with using the mobile health applications.

Future studies will focus on exploring if there is variation in the relevance of the privacy threats in mHelath apps and security challenges of mHealth apps by categories of mHealth apps, and by the platforms in which the apps are running on.

6 References

[1] Richards, R. Adhikari and Deborah, "Security and Privacy Issues Related to the Use of Mobile Health Apps"; 25th Australasian Conference on Information Systems, 8th - 10th Dec 2014, Auckland, New Zealand.

[2] D. F. Schulke, "The Regulatory Arms Race: Mobile-Health Applications and Agency Posturing"; Boston University Law Review, vol. 93, no. 5, p. 1699, October 2014.

[3] K. Congdon, "The Rise Of mHealth," *Health IT Outcomes Magazine,* 25 March 2013.

[4] B. Faudree and M. Ford, "Security and Privacy in Mobile Health," CIO Journal, 6 August 2013.

[5] Kharrazi, R. Chisholm , D. VanNasdale and B. Thompson , "Mobile personal health records: an evaluation of features and functionality," 17 July 2012.

[6] M. McCarthy, "Experts warn on data security in health and fitness apps," BMJ, 13 September 2013.

[7] B. S. Narisi, "Mobile health apps create privacy risk", Healthcare Treatments & Outcomes, 19 July 2013, retrieved on March 25, 2016 from http://www.healthcarebusinesstech.com/mobile-health-apps-privacy/.

[8] M. Plachkinova, S. Andres and S. Chatterjee, "A Taxonomy of mHealth Apps – Security and Privacy Concerns," HICSS '15 Proceedings of the 2015 48th

Hawaii International Conference on System Sciences,
Pages 3187-3196, 2015.

ISBN: 1-60132-437-5, CSREA Press ©