Title
Dynamic Adaptive Remote Health Monitoring for Patients with Chronic Disease

Author
Suh, Myung-kyung

Publication Date
2012

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Dynamic Adaptive Remote Health Monitoring for Patients with Chronic Disease

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Myung-kyung Suh

2012

Abstract of the Dissertation

Dynamic Adaptive Remote Health Monitoring for Patients with Chronic Disease

by

Myung-kyung Suh

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2012

Professor Majid Sarrafzadeh, Chair

Chronic diseases are the leading causes of death and disability in the United States. More than 70% of deaths among Americans are caused by chronic diseases and more than 133 million Americans have at least one chronic disease. Due to the prevalence of chronic disease-related issues, it is prudent to seek out methodologies that would facilitate the prevention, monitoring, and feedback for patients with chronic diseases.

This dissertation describes WANDA (Weight and Activity with Other Vital Signs Monitoring System); a system that leverages sensor technologies and wireless communications to monitor the health-related measurements of patients with chronic diseases. The system was developed and validated in conjunction with the Computer Science Department, the School of Nursing and Ronald Regan Medical Center at the University of California, Los Angeles to enable real-time patient monitoring, user task optimization, missing data imputation and key clinical symptom prediction. The main contributions of designing and developing the WANDA system are 1) data abstraction and integration of the server side; 2) development of the smartphone and web applications; 3) data backup and recovery; 4) algorithm design and development of missing data imputation; 5) algorithm design and development of task optimization and early adaptive alarm; and 6) system deployment for clinical trials.

The WANDA system is a three-tier architecture consisting of wireless sensors, web servers, and back-end data analytics engines. The first tier comprises sensors that measure patients' vital signals and transmit data to the web server tier. The second tier consists of web servers that receive data from the first tier and maintains data integrity. The third tier is a back-end database server that performs data backup and recovery and various data analysis including dynamic task optimization, missing data imputation and adverse event prediction.

The WANDA dynamic task optimization function applies data analytics in real-time to discretize continuous features and apply data clustering and association rule mining techniques to manage a sliding window size dynamically and to prioritize required user tasks. The developed algorithm minimizes the number of daily action items required by patients using association rules that satisfy a minimum support, confidence, confirmation and conditional probability thresholds. Each of these tasks maximizes information gain, thereby improving the overall level of patient adherence and satisfaction. Experimental results from applying EM-based clustering and Apriori and confirmation-based rule mining algorithms show that the developed algorithm can reduce the number of user tasks by up to 76.19% with higher confidence levels .

Although missing data is highly undesirable as automated alarms may fail to notify healthcare professionals of potentially dangerous patient conditions, many studies reported high missing data rates in remote health monitoring. In this dissertation, I exploit machine learning techniques including projection adjustment by contribution estimation regression (PACE), Bayesian methods, and voting feature interval (VFI) algorithms to predict both non-binomial and binomial data. The experimental results show that the aforementioned algorithms are superior to other methods with high accuracy and recall. This approach also shows an improved ability to predict missing data when training on entire populations, as opposed to training unique classifiers for each individual.

The WANDA early adaptive alarm function discretize continuous features, applying Expectation Maximization (EM) clustering and association rule mining techniques for predicting future sensor readings and system non-use dynamically. The experiment results shows that developed algorithm can predict upto 27.08% of sensor readings and system non-use within the next three days.

Additionally, the results of performed clinical trials shows that patients monitored by WANDA are less likely to have readings fall outside a healthy range and have higher adherence rate and more communications with health care professionals.

The dissertation of Myung-kyung Suh is approved.

Mario Gerla

Carlo Zaniolo

Alex Bui

Majid Sarrafzadeh, Committee Chair

University of California, Los Angeles

2012

To my dear husband, Jonathan Woodbridge and daughter, Abigail Woodbridge,

my family and friends,

and God...

who have believed and supported me to accomplish my dreams.

Especially thank you to our parents,

Junghun, Myunghyun, Robert and Rosemarie who are the greatest role models.

Table of Contents

List of Figures

List of Tables

Acknowledgments

My research has been supported by NIH/National Library of Medicine Medical Informatics Training Program Grant T15 LM07356.

I would like to gratefully acknowledge the enthusiastic supervision of Prof. Majid Sarrafzadeh during my Ph.D. I wish to express my sincere gratitude to my co-advisor, Prof. Alex Bui for his wonderful and helpful advise. I also would like to thank to Prof. Alfonso Cardenas.

I wish to express my gratitude to my collaborators in my research, Jonathan Woodbridge, Lorraine S. Evangelista, Aurelia Macabasco-O'Connell, Tannaz Moin, Ani Nahapetian, Mars Lan, Hassan Ghasemzadeh, Mahsan Rofouei, William J. Kaiser, Kyujoong Lee, Kyle Dorman, Marjan Yahyanejad, William McCarthy, Alfred Heu, Victor Chen, Wen-Sao Hong, Jamie Macbeth, Florence-Joy Figueras, Chien-An Chen, Kyungsik Han, Jinha Kang, Michael Kai Tu, Sheila Ahmadi, Lauren Samy, Nabil Alshurafa, and Jung In Kim.

Vita

2004 – 2007	Outstanding Student Scholarship, Sogang University, Korea.
2006 – Present	A Member of Alpha Sigma Nu Honor Society.
2007	B.S. (Computer Science and Engineering), Sogang University, Korea. Summa Cum Laude and Early Graduation.
2009	Teaching Assistant, Computer Science, University of California, Los Angeles. Computer Science 143 (Database Systems). Computer Science 33 (Introduction to Computer Organization).
2009 – 2012	NIH/National Library of Medicine Medical Informatics Training Program Fellowship.
2009	Best Paper Award, Conference on Mobile Computing, Applications, and Services (MobiCASE).
2010	M.S. (Computer Science), University of California, Los Angeles.
2011	Technology Intern, Zynx Health

Publications

Machine Learning-Based Adaptive Wireless Interval Training Guidance System. **Myung-kyung Suh**, Ani Nahapetian, Jonathan Woodbridge, Mahsan Rofouei, Majid Sarrafzadeh.

Mobile networks and applications (MONET), 17(2), 163-177. 2012.

A Remote Patient Monitoring System for Congestive Heart Failure. **Myung-kyung Suh**, Chien-An Chen, Jonathan Woodbridge, Michael Kai Tu, Jung In Kim, Lorraine S. Evangelista, Majid Sarrafzadeh. Journal of Medical Systems (JOMS), 35(5), 1165-1179. 2011.

Dynamic Task Optimization in Remote Diabetes Monitoring Systems. **Myung-kyung Suh**, Jonathan Woodbridge, Tannaz Moin, Mars Lan, Nabil Alshurafa, Lauren Sany, Bobak Mortazavi, Hassan Ghasemzadeh, Alex Bui, Sheila Ahmadi, majid Sarrafzadeh. Annual IEEE Healthcare Informatics, Imaging, and Systems Biology Conference (HISB). September 2012.

Dynamic Self-adaptive Remote Health Monitoring System for Diabetics. **Myung-kyung Suh**, Tannaz Moin, Jonathan Woodbridge, Mars Lan, Hassan Ghasemzadeh, Sheila Ahmadi, Alex Bui, Majid Sarrafzadeh. International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). August 2012.

Missing Data Imputation for Remote CHF Patient Monitoring. **Myung-kyung Suh**, Jonathan Woodbridge, Mars Lan, Alex Bui, Lorraine S. Evangelista, Majid Sarrafzadeh . International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). August 2011.

An Automated Vital Sign Monitoring System for Congestive Heart Failure Patients. **Myung-kyung Suh**, Lorraine S. Evangelista, Chien-An Chen, Kyungsik Han, Jinha Kang, Michael Kai Tu, Victor Chen, Ani Nahapetian, Majid Sarrafzadeh. ACM International Health Informatics Symposium (IHI). November 2010.

Bayesian Networks-Based Interval Training Guidance System for Cancer Rehabilitation. **Myung-kyung Suh**, Kyujoong Lee, Alfred Heu, Ani Nahapetian, Majid Sarrafzadeh. Conference on Mobile Computing, Applications, and Services (MobiCASE). October 2009. **(Best Paper Award)**

Interval Training Guidance System with Music and Wireless Group Exercise Motivations. **Myung-kyung Suh**, Kyujoong Lee, Ani Nahapetian, Majid Sarrafzadeh. IEEE Symposium on Industrial Embedded Systems (SIES), July 2009.

CHAPTER 1

Introduction

1.1 Chronic Disease and Remote Health Monitoring

Chronic diseases are the leading causes of death and disability in the United States. More than 70% of deaths among Americans are caused by chronic diseases and more than 133 million American (or about 1 out of 2 adults), have at least one chronic disease [KHX08],[WG00]. For example, the Center for Disease Control and Prevention(CDC) [DP12] indicates that approximately 670,000 individuals are diagnosed with congestive heart failure (CHF) every year and more than 280,000 deaths are caused by CHF each year. Diabetes is one of the leading causes of death in the United States and accompanied by complications including blindness, hypoglycemia, renal failures, cardiovascular disease, etc. during diabetes' lifetime [CG94]. In 2011, 25.8 million Americans, which are 8.3 % of the population have diabetes [Ame11]. Costs of chronic illness are also hard to neglect. Hoffman expected the direct medical costs for chronic diseases will be 685 billion dollars a year by 2020 and 906 billion dollars by 2050 [CHF96]. In the United States, heart failure costs 39.2 billion dollars in 2008 and 24% of patients must be readmitted to the hospital within twelve weeks of discharge. The largest expenditure for heart failure was inpatient care including 13.3% of emergency room visits and the average hospitalization cost of heart failure was 10,000 dollars [LAC09][Ass08]. In 2007, the cost of diabetes exceeded 174 billion dollars including 27 billion dollars to directly treat diabetes, and 58 billion dollars to treat the diabetes-related chronic complications [DMZ08]. Half of expenditures for diabetes were inpatient care and 23% were used for purchasing diabetes and its complication-related medications.

Moreover, as getting older, the probability and risk for chronic disease including CHF, diabetes and stroke goes up. In the United States, the average life expectancy at birth increased to 78.2 years in 2009, and this is 8.3 years longer than the life expectancy in 1959 [DP12]. As the average life expectancy is longer than before, the chances of having a chronic disease, the death rate, complications and costs caused by chronic disease are getting higher.

However, early diagnosis and treatment can improve quality of life and life expectancy for people with chronic diseases. Treatment usually involves life style changes such as measuring biomedical readings, taking medicines, regulating diets, and doing regular physical activity. For instance, for individuals at risk for diabetes, weight loss and regular exercise, regulating blood glucose level and blood pressure prevents adverse results. Furthermore, by working with their support network, including caregivers and families, patients can prevent disability and death by controlling their symptoms in a timely manner. As most of chronic diseases are related to their daily habits such as monitoring symptoms, eating and exercise, sharing daily readings, planning meals, providing information and emotional support are critical [WFD98]. Further details of related comorbid conditions and health-related readings are described in Section 1.1.1.

Recent advances in sensors, smart phones and wireless technology have made a new generation of health monitoring systems available. Remote health monitoring leverages sensor devices and wireless communication technologies such as Bluetooth, Zigbee, Wifi, etc. to monitor the health-related measurements without location, time and other constraints. As such, remote health monitoring systems provide a potentially feasible option for dealing with the expanding population of patients that have chronic diseases, but are unable to access clinics due to either a lack of resources, location, infirmity, or limited time and resources. Care facilitated by technology has the potential to enable early detection of key clinical symptoms indicative of related decompensation and allows health professionals to offer surveillance, advice, and continuity of care and triggers early implementation of strategies to enhance adherence behaviors.

There have been many studies that have shown that well-designed remote health monitoring improves quality of life and leads to a higher efficiency and an increased accessibility to patients with chronic disease and healthcare professionals [SPL08] [Bla08]. In order to design an effective remote health monitoring system for chronic diseases, it is important to make an automated real-time system for checking important health-related readings and transmitting data to the server layer. The system should be in real-time to ensure the timely delivery of data to physicians. As systems that are not automated or real-time have shown that remote health monitoring did not improve patient outcomes [CBM07], it is important to design the system to have user-friendly interfaces, monitor required readings and deliver data in a timely manner. Furthermore, in order for remote health monitoring to be truly successful, it should also be able to perform data analysis. This means the system should provide a wealth of analytical algorithms that can infer useful information or discover predictive patterns from the data. Armed with the information, user task optimization, missing data imputation and early interventions can be performed to improve quality of care. Unfortunately, most current remote monitoring systems lack such advanced analytics capabilities.

1.1.1 Critical Health-related Measurements for Chronic Diseases

Many studies show that lifestyle changes with careful symptom and health-related measurements monitoring, and social support can help to delay or prevent chronic diseases. Attributes to be monitored in remote health monitoring can be different depending on patients' chronic disease, complications, improvements or regressions in disease status and also economical constraints. In this section, I mainly focus on attributes related to CHF and diabetes that I have been actively researching and participating in several clinical trials.

1.1.1.1 Weight

In a study of 5881 subjects (3177 women and 2704 men) by Kenchaiah [KEL02], the risk of CHF was increased by 5% for men and a 7% for women when the Body Mass Index (BMI) was increased by only 1%. For diabetes, a study by Hamman [HWE06] of 1079 participants showed that every kilogram of weight loss reduces 16% in risk of diabetes and proved that weight loss is the dominant predictor of reducing diabetes. In addition, 5-7% weight loss and 20 minutes of exercise each day can reduce the chance of developing of Type II diabetes by 58% [The02].

1.1.1.2 Blood Pressure

A study by Vasan [HFN00] of 6859 patients establishes that increased blood pressure is correlated with an increased risk of cardiovascular disease. CHF is often caused by systolic dysfunction where the heart muscle cannot adequately pump or eject the blood out of the heart, or by a diastolic dysfunction where the atrium does not fill up. As this pumping procedure stops, blood may back up in other areas of the body, producing congestion in the lungs, liver, gastrointestinal tract, arms, or legs. Heart rate is an additional factor that predicts the risk for CHF in an elderly person. Heart rate may help identify patients at high risk for overt CHF who are candidates for aggressive blood pressure control [HLF03][RJB03]. For diabetes, Vijan's work [VH03] showed that aggressive blood pressure control with blood pressure goals of systolic blood pressure, 135 mm Hg and diastolic blood pressure, 80 mm Hg prevents adverse results in patients with diabetes.

1.1.1.3 Blood Glucose

For diabetes, blood glucose control is the most common method to prevent emergent events, as diabetes is caused by a shortage of insulin or a disability to properly use insulin in the body. Healthcare providers generally recommend diabetics to use a blood glucose meter and keep a daily log to decide health care regimens. Hanefeld's study [HFJ96] of 1139 subjects emphasized that regulating blood glucose level between 80 and 200 mg/dL is proper.

1.1.1.4 Activity

He's study [HOB01] suggests several risk factors for CHF including low physical activity, which accounts for 9.2% of risk. The results of Hambrecht's work [HFW98] suggests that long-term aerobic exercise training in patients with CHF restores function of the skeletal muscle microvasculature of the lower limb. As mentioned in Section 1.1.1.1, twenty minutes of exercise each day can reduce the chance of developing of Type II diabetes by 58% [The02].

1.1.1.5 Fall Detection

Walking disorders and frequent falls are common symptoms that can be found in elderly populations who are also more likely to experience chronic diseases. As Patel [PSS07] mentioned, walking disorders are one of the most prevalent symptoms of congestive heart failure patients. Walking problems are related to other symptoms such as dizziness, swollen legs, shortness of breath, feebleness, sore stomach, chest pains, shakiness, irritating feelings while walking, and problems with balance. As walking problems are related to imbalance, it is important to monitor patients and notify caregivers and family members when they need help.

1.1.1.6 Symptom Questionnaires

The Heart Failure Somatic Awareness Scale (HFSAS) in Table 1.1 is a 12-item Likert-type scale for the purpose of measuring awareness and perceived severity of CHF specific signs and symptoms. The 12 items of the HFSAS reflect the most common signs and symptoms of CHF. A 4-point Likert-type scale is used to address the degree of these symptoms and ascertain how much the patient is bothered by the specific symptom (0: Not at all, 1: A little, 2: A great deal, 3: Extremely). Scores range from 0 to 36, with higher score showing higher perceived somatic awareness and symptom distress [JFR06], The HFSAS is useful in studies designed to improve symptom recognition and self-management.

Talbot designed Multidimensional Diabetes Questionnaire (MDQ) to assess social and cognitive factors and proved that MDQ can help understand diabetics [TNG97]. MDQ includes questions related to social support from the patient's network, positive and misguided self-management behaviours, and the patients' expectancies. In addition, Ciechanowski [CKR03] emphasized that depressive symptoms are related to low physical activities and other indicators in diabetes patients.

1.1.2 Wireless Sensor and Communication Technologies

In remote health monitoring systems, it is critical to provide sensor devices which are light weight, accurate and secure [MOJ06]. To achieve unobtrusive frequent or continuous monitoring, the sensors should be light weight or easy to carry. However, some cases such as monitoring blood pressure or glucose at home, devices which are easy to store or carry in a house would be good enough. The readings from sensors should be accurate, as the sensor readings are used to generate alarms, feedbacks and data analytics. In addition, the sensor devices' internal storage and its communication methods should be highly secure to protect the privacy of individual's health information. The wireless communication requirements of remote health monitoring systems are reliable access and transmission of sensor readings,

Questionnaire Items
Q1. I could feel my heart beat faster
Q2. I could not breathe when I laid down
Q3. I felt pain in my chest
Q4. I had an upset stomach
Q5. I had a cough
Q6. I was tired
Q7. I could not catch my breath
Q8. My feet were swollen
Q9. I woke up at night because I could not breathe
Q10. My shoes were tighter than usual
Q11. I gained 3 or more pounds in the past week
Q12. I could not do my usual daily activities because I was short of breath

Table 1.1: Heart Failure Somatic Awareness Scale Questionnaire

location management, etc. Combinations of existing wireless networks such as cellular-oriented (2G/3G/4G), Local Area Networks(LANs) and personal area network technologies such as Bluetooth, WiFi, Zigbee, etc. could support these requirements [Var07]. However, as different age groups and economical or residential situations can affect on preferred communication methods (Figure 1.1), considering the patient population and its characteristics of the chronic disease is essential.

1.2 Uncertainty Problems in Remote Health Monitoring

Remote patient monitoring systems can suffer from uncertainty. Uncertainty is classified as either system-based uncertainty or human-based uncertainty. System-based uncertainties are derived from sensing coverage, network failure, system-based data uncertainty and connectivity issues. Issues in sensing coverage occur when sensing devices cannot detect objects to sense or transmit data when the objects or devices are out of its sensing range. Routing problems occur when links between sensing nodes or data dissemination algorithms are not clearly defined in sensor networks. Network failures occur when communication layers fails to transmit entire or partial data to the server systems. System-based data uncertainties result from physical environments and hardware defects resulting in raw readings from sensors that are inaccurate. Human user-based uncertainties are from system misuse or system non-use. System misuse is mostly from users that are not technologically savvy. System non-use is caused by a loss of interest or conflicts with the users schedule. For example, users can be under pressure to follow given instructions and wear sensing devices in their daily life and have limited amount of time to measure all their readings. Such uncertainty problems from system and human factors can result in missing data, noisy data and erroneous data that can cause problems in data quality and decision making process in remote health monitoring.
In this dissertation, the task optimization function is developed to improve the adherence rate and minimize uncertainty problems due to system non-use. The missing data imputa-

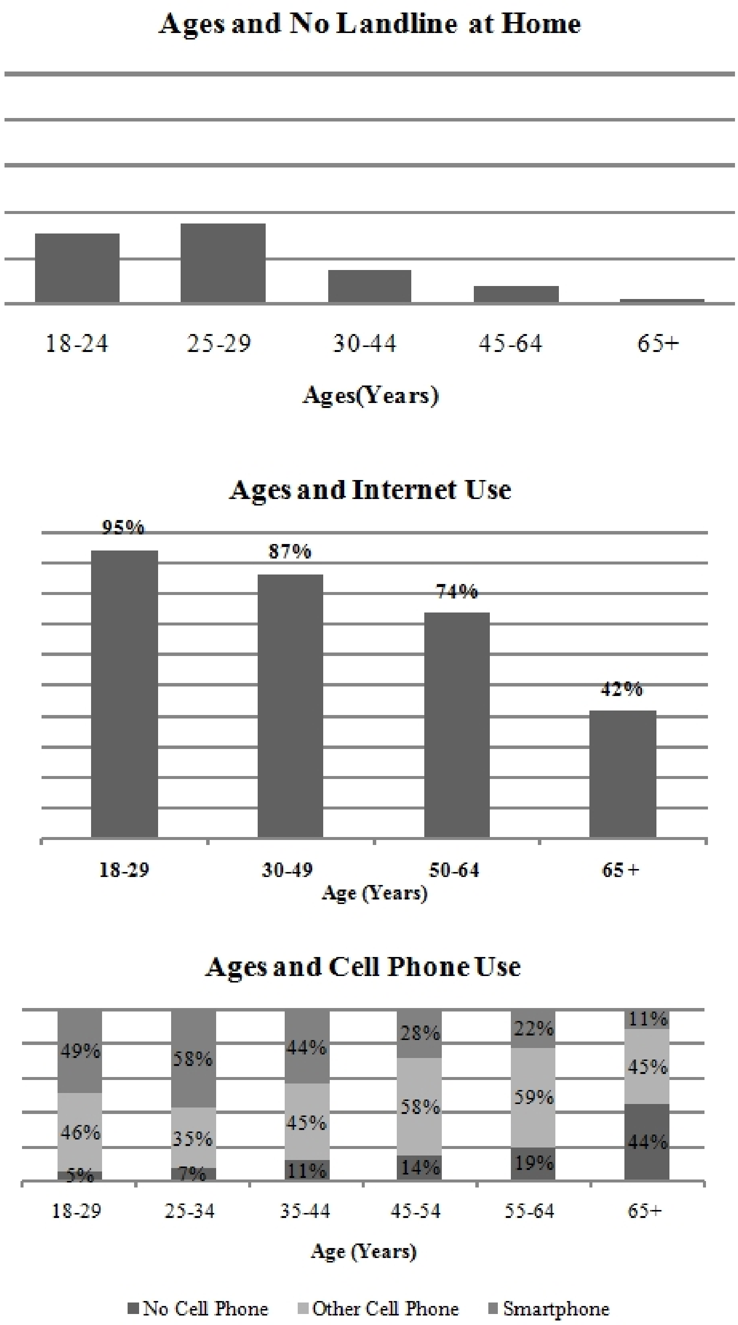


Figure 1.1: Ages and Preferred Communication Methods [Bur11]

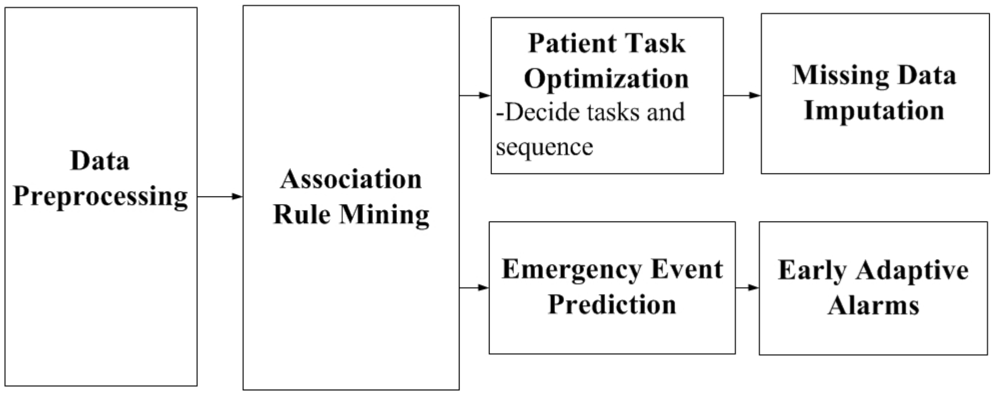


Figure 1.2: Optimization Flow in WANDA.

tion function will predict missing values of sensor readings and questionnaire answers due to both system-based and human user-based uncertainties. The outputs from aforementioned functions will be used to predict future emergency events and provide guidance to patients and healthcare professionals.

1.2.1 Task Optimization in Remote Health Monitoring

In remote health monitoring, patients are required to perform a series of daily tasks requested by their healthcare professionals. For instance, CHF patients in Chaudhry's work were required to answer 16 questions and measure and enter their weight using telephone keypads [CBM07] and the study results showed a high missing data rate. one of my studies [SEC10] required patients to measure weight, blood pressure, and 12 symptom questionnaires on a daily basis and showed frequent system non-use. Such system non-use in remote patient monitoring can severely degrade the patient participant rate and the effectiveness of designed systems. As missing data can lead to biased and dangerous conclusions, it is important to reduce the missing data rate and adequately handle missing data [WWT04]. As task complexity is a factor that highly affects user participation and satisfaction [MGW94][Mag01][OKS11][LG94], reducing the number of required tasks would reduce missing data rate.

For designing a human-centered system, it is critical to distinguish which tasks should be handled by users or automatically processed by computers [Mag01]. In remote health monitoring, analysing outputs of user tasks in real time can help dynamically schedule sequence of tasks, avoid unnecessary tasks, increase usability and effectiveness of the system.

Most remote health monitoring systems utilizes medical domain experts knowledge to determine and assign priorities and sequences of tasks. For example, Tang applied a heuristic evaluation method using expert knowledge [TJT06] and Dabbs utilized expert knowledge and patients' survey feedback for designing a health monitoring system [DMM09]. As such, most remote health monitoring systems do not apply data-driven dynamic process for designing human-centered units and yield redundant information gains.

1.2.2 Missing Data Imputation

The first randomized trial of the developed system [SCW11][SWL11] experienced a considerable amount of missing data: only 33% of the somatic questionnaires were completed; and 55.7% of data had missing values for weight, systolic and diastolic blood pressure, and heart rate data. Moreover, 22.2% of patients experienced system misuse and requested help to accustom themselves to the technologies. Missing data was further caused by system non-use and service disorder (such as a network failure, resulting in as much as 6.3% of all of the missing data) as mentioned earlier. Notably, other studies have experienced similar data loss [CMC10][DS10] and missing data problems can cause biased and erroneous results in adverse events detection. Unfortunately, there has been not enough studies on missing data imputation using data mining techniques in the remote health monitoring domain.

In this dissertation, I try to minimize missing data caused by system non-use through task optimization in Section 3.3.2 and impute further missing data via data mining techniques.

1.2.3 Early Adaptive Alarms

Adverse event indicates any harmful changes or side effects occurring during the treatment. Certain adverse events including hospitalization, death and other life-threatening events are considered as serious problems and should be prevented. There have been many studies proves prognosis of adverse events [HFK95][MYG00] including items in Section 1.1.1 and careful and close prognosis monitoring is necessary for preventive interventions. However, most of the existing studies found prognosis based on the hospital visits and lab results in a long-term monitoring.

Although statistical analysis is one of the key factors in remote health monitoring [GHS11], it is not explored enough to find prognosis of adverse events in remote health monitoring. As remote health monitoring provides continuous health-related measurements in an unsupervised real-world setting, the system should provide a wealth of analytical algorithms that can infer useful information or discover predictive patterns from the data. As data from remote health monitoring are unsupervised and continuous, proper data analytics methods are required and verified with clinical data. Armed with the information, personalized early interventions can be made to prevent harmful medical events from happening [GS11].

1.3 WANDA, Remote Health Monitoring System

In this dissertation, I present a remote health monitoring system and its applied data mining techniques. The WANDA (Weight and Activity with Other Vital Signs Monitoring System) system was developed in conjunction with the University of California, Los Angeles Computer Science Department, the School of Medicine and the School of Nursing. WANDA can monitor the various health-related readings of patients who are at risk of chronic diseases and aid in health management and preventative care. WANDA is not only useful to physicians, but can function as a tool for patients to become self-aware of their own weight, as the WANDA platform can be used for continuous daily monitoring. Additionally, WANDA's data analytics functions will enhance patient adherence rate, predict missing data and prevent adverse events.

In this dissertation, I will focus on the system architecture and data mining algorithms designed for patients with congestive heart failure (CHF) and diabetes in our clinical trials. CHF is a condition in which heart function is inadequate to supply oxygenated blood to the patient. Diabetes is by a shortage of insulin or a disability to properly use insulin in the body. I will also verify the effectiveness of the developed system and the proposed data mining algorithms.

CHAPTER 2

Related Work

2.1 Remote Health Monitoring Projects

There have been many studies of designing, developing and verifying the system effectiveness through clinical trials. The following studies are examples of remote health monitoring system developments and their results helped tailor and develop the WANDA system.

2.1.1 Congestive Heart Failure and Remote Health Monitoring

Chaudhry [CBM07][CMC10] utilized a telephone-based interactive voice response system (Pharos Tel-Assurance system [Pha12]) for CHF patients. This system collected daily information about symptoms and weight. Patients in this study were required to make daily calls to the system. During each call, patients were asked a series of questions about their general health and CHF symptoms and weight. Responses were entered into the system using the telephone keypad. Information from the telemonitoring system was downloaded daily and reviewed by nurses on every non-holiday weekday. The study suggests that telemonitoring did not improve patient outcomes.

Soran's [SFP10][SPL08] used an electronic scale and an individualized symptom response system (the Alere DayLink monitor [Ale12]) for CHF patients. System components were linked via a standard phone line to databases. If criterion values were met for weight or symptom alerts, the nurses immediately contacted the patient to check on the status of the

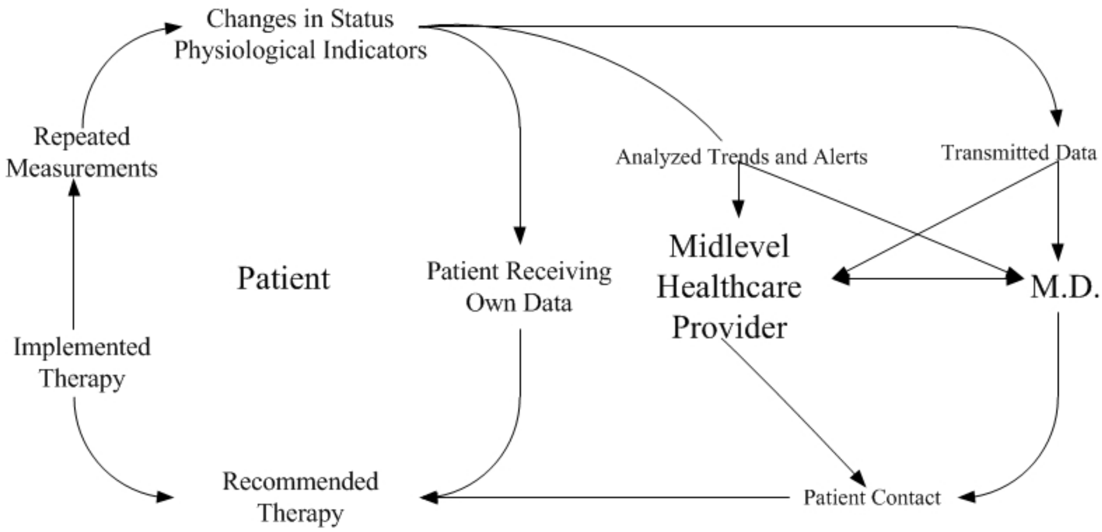


Figure 2.1: Desai's Circle of Remote Health Monitoring for Heart Failure [DS10].

patient. After nurse-patient interactions, the primary physician was notified by a fax report to adjust medications and schedule an appointment. Soran's work showed that enhanced patient education and follow-up was as successful as a home monitoring device for elderly patients receiving care from a community-based primary care practitioner.

Desai's work [DS10] attempted to explain the reason why there was no benefit seen with telemonitoring intervention in Chaudhrys study (Figure 2.1). First, the signals of weight and symptoms do not provide adequate warnings for CHF. Results from trials of CHF patients monitoring [ZBS08] suggest that only monitoring weight is inadequate, as the target dry weight changes on the basis of caloric intake. Second, the telemonitoring system was underutilized, with only 55% of patients making three calls week by week. Third, the intervention may not have been structured for timely and appropriate corrective action. With regard to timing, the requirement that coordinators review data may have caused a serious break in patient care. The team member receiving the data should have been able to contact the patient directly to discuss a treatment plan, without having to triangulate the discussion with a physician. Similar to Chaudhry's work, Soran's study also only monitored weight and symptom responses, and was limited by the need to wait on a physicians decision.

2.1.2 Examples of Diabetes and Remote Health Monitoring

Columbia University's Informatics for Diabetes Education and Telemedicine (IDEATel) project was started in 2000 and is the largest telemedicine effort ever funded by the federal government. This project provided video conferencing, blood glucose and blood pressure monitors, Web-based messaging and clinical data review, and Web-based educational materials via a standard telephone line connection. The study's clinical trials have shown long-term improvements in HbA1c, low-density lipoprotein cholesterol and blood pressure and patients who used IDEATel show high satisfaction levels [Bla08][SHS02]. However, only providing dial-up modems using a landline connection can affect patient recruitment. Younger generation prefer Ethernet and smart phone applications while older generations prefer using a landline. In addition, some people do not have any of the aforementioned options. Also, measuring more than blood pressure and glucose values can be helpful to monitor patients with diabetes.

Mougiakakou [MKI09] developed a mobile diabetes monitoring system for tracking blood glucose, blood pressure, physical activity, insulin and food/drink intake information. The main limitation of this system is the developed system is not seamless and automated. The user manually enters timestamps, units, reception places and annotations of the measurement through a user interface.

Zhou [ZYA10] developed an Android-based diabetes patient monitoring system. The system measures blood pressure, blood glucose, exercise, food intake and medication intake. The main limitation of both Mougiakakou and Zhou's work is that the system was not evaluated with real patients, as it is critical to verify the system's effectiveness and ease of use through clinical experts and real patients. As the system utilizes the Android operating system, users who are not accustomed to using new smartphone technologies can experience troubles with the system.

2.2 Patient-oriented Remote Health Monitoring and Task Optimization in Remote Health Monitoring

One way to quantify patient satisfaction with health monitoring systems is evaluating the amount of the system use [Klo09]. Huang [Hua10] designed a neural network-based remote health monitoring system adoption model to predict the behavior intention (BI) toward to using the system. Huang's model is based on attitude toward to use (ATT), perceived ease of use (PEOU), perceived usefulness and benefits (PUB), perceived disease threat (PDT), perceived barrier to take action (PBTA), external cues to action (ECUE) including helps from networks and internal signs (IS) such as disease history (Fig.2.2). The data used for evaluating the methods are survey answers in five-point Likert-Type scales. Wu's study [WWL07] showed that acceptance and satisfaction related to mobile healthcare system is relevant to compatibility, ease of use and perceived usefulness. In addition, the study results show that perceived usefulness and ease of use are highly related to the self-efficacy which is a belief that the person has an ability to execute series of required tasks. As self-efficacy is related to task complexity [SF97], it is important to make the task procedure simple.

Time and cost-effectiveness is also an important criterion related to patient satisfaction with remote health monitoring systems [HHM03]. Time-effectiveness is a factor of perceived ease of use and cost-effectiveness is one of perceived usefulness. Many studies in Seto's paper [Set08] have shown 1.6% to 68.3% cost reductions from 10 remote health monitoring studies compared to usual care. Cost reduction referenced in Seto's study is mainly because of reduced hospitalization and nurse home visits. However, initial costs of device and service purchase can be an obstacle to user satisfaction despite of long-term savings [SWT09].

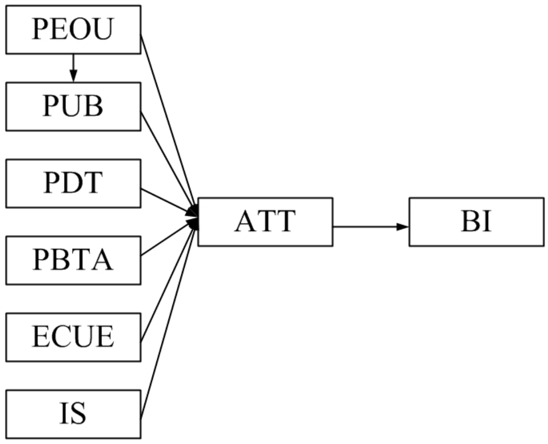


Figure 2.2: Huang's Research Structure [Hua10]. Note: perceived ease of use (PEOU), perceived usefulness and benefits (PUB), perceived disease threat (PDT), perceived barriers of taking action (PBTA), external cues to action (ECUE), internal signs(IS), attitude toward using (ATT), and behavioral intention to use (BI).

Therefore, reducing the number of required devices and daily tasks can reduce further costs of purchase and physician workload of reviewing and analyzing collected data.
In this dissertation, I mainly focus on the perceived ease of use and usefulness for enhancing the patient satisfaction and adherence rate by decreasing the number of required sensors and tasks.

2.3 Missing Data Imputation in Remote Health Monitoring

Missing data is especially unavoidable and common in clinical trials. Wood's study [WWT04] showed that 89% of 71 trials published in 2001 in well-known journals (British Medical Journal, BMJ; Journal of the American Medical Association, JAMA; Lancet; and New England Journal of Medicine, NEJM) reported having partly missing outcome values. Many studies applied ad-hoc approaches including last observation carried forward, worst case imputation, missing category indicator and complete case analysis techniques that can lead

to biased results. Moreover, the cumulative effects in missing data often exclude considerable amount of data, and results in low accuracy, precision and recall [SWC09] [VB91]. However, if missing at random (MAR) hypothesis used, which assumes that missing data can be explained by observed data, less biased and more powerful analyses are available [SWC09]. For example, if MAR is used, it is possible to build a more general model using random effect models with partially observed data on previous or next time points. Other approaches utilize maximum likelihood estimation, probability weighting, etc. under MAR [SRR99][CKV06].
Pecchia [PMB10] developed a home monitoring system with classification and regression tree algorithm (CART) for heart failure detection and severity assessment. The experimental result showed high precision and accuracies. However, authors didn't apply their algorithms to data collected from their system because the clinical trial is not completed. Also, they only applied the algorithm to ECG data. This system doesn't address any missing data problems.
Unfortunately, to date, there has been not enough studies on missing data imputation using data mining techniques in the remote health monitoring domain.

2.4 Early Adaptive Alarms in Remote Health Monitoring

Pérez-Gandia used artificial neural network methods for predicting blood glucose data at prediction horizon time, 15, 30 and 45 minutes [PFS10]. The author applied the developed methods on data collected from 15 patients with type 1 diabetes. The study results showed that the root mean square error was approximately 10, 18, and 27 mg/dL for 15, 30, and 45 min of prediction horizon time and the prediction delay was about 4, 9, and 14 min for upward trends and 5, 15, and 26 min for downward trends. The main differences in Pérez-Gandia's study with the WANDA diabetes (Section 4.1.2) are the use of continuously monitored glucose data and the short prediction horizon time in Pérez-Gandia's study.
Most of existing remote health monitoring studies have applied various data mining and

statistical analysis approaches on time-series data including activities, ECG, etc. as Pérez-Gandia's study. However, most of existing studies did not verify the developed methods on data from real patients collected under unsupervised remote health monitoring environment. As remote health monitoring requires high usability and efficiency, the use of real clinical data from patients is essential. Additionally, as the characteristics of data from patients with chronic diseases are different from non-patient data, the prediction algorithm should be verified with real patients.

Jakkula applied the Box-Jenkins methodology to predict blood pressure data in the smart home-based health care system [JCJ07]. The Box-Jenkins methodology is a time series analysis which builds a model based on autocorrelation [Geo70]. Jakkula's study utilized 61-days motion sensor data and health-related data such as blood pressure and pulse from a single person. However, this study used a small size of data and did not use data from clinical trials.

Jin used an artificial neural network for detecting abnormal cardiovascular activities, combining the patient's ECG data and patient profile [JOH09]. However, the author used MIT-BIH Arrhythmia Database which is collected under supervised settings for validating the results [JSC09]. Therefore, the method can lead different results in remote health monitoring under an unsupervised setting.

CHAPTER 3

System Architecture

WANDA is a three-tier end-to-end remote monitoring system with extensive hardware and software components designed to cover the broad spectrum of the telehealth and remote monitoring paradigm. The overall architecture is summarized in Figure 3.1.

The first tier of the architecture consists of a data collection framework, which is formed from a heterogeneous set of sensing devices that measure various bodily statistics such as blood glucose, weight, body fat, body water, blood pressure, heart rate, blood oxygen saturation and body movements. The data from these sensors are collected, processed, and transmitted via a phone-line, Ethernet or smartphone-based gateway to the cloud - the second tier of the WANDA architecture. The large amount of data are stored and indexed using a scalable database and can be accessed easily. The third tier is a back-end database server that performs data backup and recovery jobs. Additionally, data in the third tier is used for data analysis such as linear regression, missing data imputation, task optimization, early adaptive alarm, signal search and clinical data security projects.

The main contributions of this chapter are data abstraction, data integration and smartphone and web application development of the second tier; and development of the existing data backup and recovery algorithm, algorithm design and development of missing data imputation, task optimization and early adaptive alarm of the third tier.

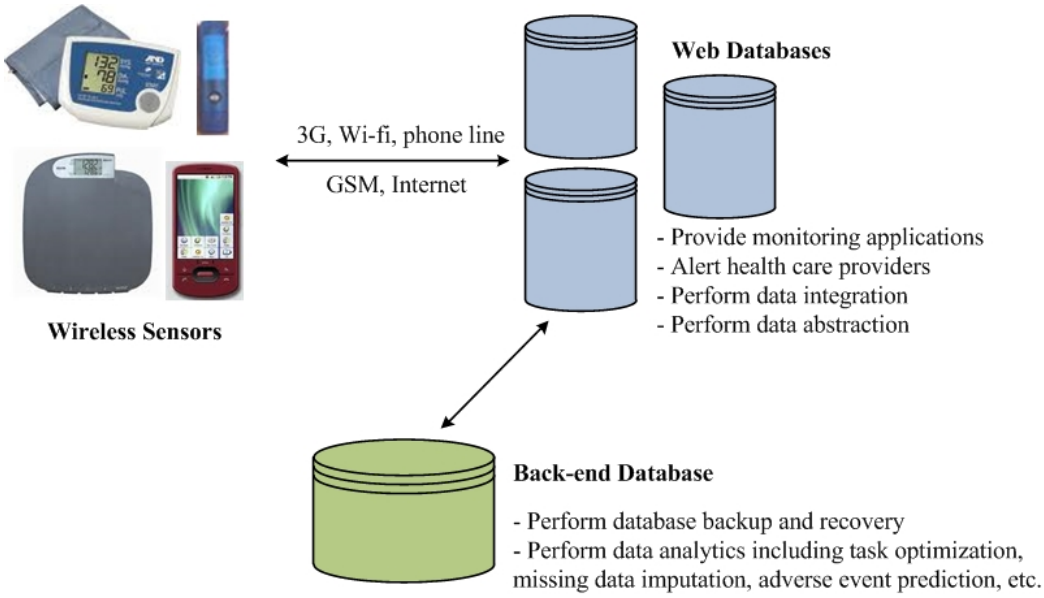


Figure 3.1: WANDA System Architecture

3.1 Sensor Tier

The first tier is comprised of wireless sensors and mobile devices. Sensors in this layer monitor patients and transfer data to web-servers. The first iteration of WANDA is designed for elderly patients who are not accustomed to smart phones or computers. Thus, WANDA only uses devices that look and function as standard weight scales, blood pressure monitors and blood glucose monitors with a standard phone line and Ethernet connection. The second version of WANDA uses a smartphone to collect and transfer data. This mobile version also allows patients to view their own health data through a smartphone interface. The second version's graphical user interface provides detailed instructions with images to make the device easy to use for patients.

The first version's sensor tier uses Bluetooth-based weight scales, blood pressure monitors, blood glucose monitors, WHI (UCLA Wireless Health Institute) [WHI12] Personal Activity Monitors (PAMs), cell phones, and an SMS message server system in order to monitor patients (Figure 3.2). As previously mentioned, patients using this system are generally

Figure 3.2: Devices Used in the First Iteration of WANDA

Figure 3.3: Devices Used in the Second Iteration of WANDA

unfamiliar with computers or smartphones, thus WANDA interfaces with the second tier through a phone line/Ethernet system in real-time. PAMs (Personal Activity Monitors) are delivered to the users via mail and are used to record patient activities. Data collected by the PAMs are uploaded to the databases every two weeks.

The second version of WANDA (Figure 3.3) utilizes a different collection of health monitoring devices from the first version in order to implement a mobile version of WANDA. The second version not only has all the functions of the original system, but also gives developers greater customizability compared to the first system. In the sensor tier, I use Bluetooth-based weight scales, blood pressure monitors, blood glucose monitors, Android-based activity monitors, fall detection monitors, and symptom questionnaire applications.

In terms of the Bluetooth device-smartphone interface, all of the devices act as masters that initiate Bluetooth communication with an Android phone. The Bluetooth protocol has a range of approximately 10 meters and provides secure data transmissions. The communication between the phone and the medical server is through Wi-Fi or 3G networks. Data measured from a sensing device is uploaded to the Android phone within 5 seconds. The Android phone transfers data to a networked server as well as stores data on a local SD card.

3.1.1 Weight Scale

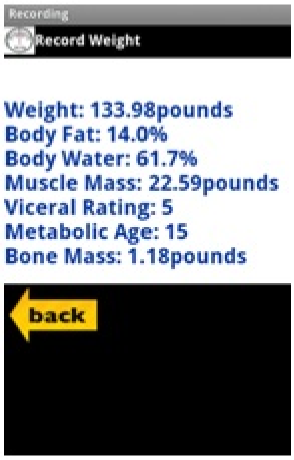
The Ideal Life system [Ide11] is a part of the first tier of the first version of WANDA. It includes the Body Manager$^{\text{TM}}$ body weight scale and the BP Manager $^{\text{TM}}$ blood pressure monitor device. The Body Manager$^{\text{TM}}$ system collects weight data and sends it to the Ideal Life Pod. As the system supports Bluetooth, the components can communicate in a range up to 300-400 feet.

The mobile version of WANDA uses a Tanita BC-590BT body composition scale, which measures body weight, body fat, body water, bone mass, muscle mass, metabolic age, and visceral rating (Figure 3.4). With the additional body data provided by the scale, health providers may be able to make an even more thorough analysis of patients' symptoms. For example, as one of the most effective means of monitoring CHF patients is monitoring one's fluid status, the weight scale features that relate to measuring the fluids in the body may help doctors diagnose patients more precisely.

3.1.2 Blood Pressure Monitor

The Ideal Life BP Manager$^{\text{TM}}$ blood pressure monitor device measures diastolic blood pressure, systolic blood pressure and heart rate in the first version of WANDA. The BP

(a) Tanita BC-590BT



(b) Body Composition Information

Figure 3.4: Weight Information in the Mobile Version of WANDA

Figure 3.5: Designed Bluetooth-enabled Blood Pressure Monitor

Manager^{TM} system collects blood pressure and heart rate data and sends it to the Ideal Life Pod^{TM} via a Bluetooth connection.

The blood pressure monitor used in the mobile version is a UA-767PBT Bluetooth blood pressure monitor from AND [A12]. This blood pressure monitors measures systolic and diastolic blood pressure, mean-arterial-pressure, and heart rate values. This version of blood pressure monitor does not natively support Bluetooth connection to the smartphone. In order to solve this lack of compatibility, we connected the blood pressure monitor with a RN-270M Bluetooth adapter from Roving Networks via a 3.5mm DB9 cable [Inc12] (Figure 3.5). A UA-767PC blood pressure monitor sends the measurement results through a serial port, and has a subsequent adapter transmit the data to the smartphone through a Bluetooth connection.

3.1.3 Blood Glucose Monitor

The Ideal Life Gluco Manager^{TM} blood glucose monitor device measures blood glucose in the first version of WANDA. The Gluco Manager^{TM} system collects blood glucose data and sends it to the Ideal Life Pod^{TM} via a Bluetooth connection.

MyGlucoHealth MGH BT1, a portable blood glucose monitor with Bluetooth capability is utilized in the mobile version of WANDA. The wireless transmitter is a class two Bluetooth device that can communicate up to 10 meters. The communication is encrypted and a PIN code is shared by the meter and the phone. To download data from the meter, two devices need a handshake procedure before the actual data can be sent. Either the phone or the meter can initiate the connection, but to be consistent with other devices, we set the meter as master always.

3.1.4 Activity Monitor

The phone-line version of WANDA uses the WHI Personal Activity monitor for monitoring patients' daily activity. The WHI Personal Activity Monitor, or PAM (shown in Figure 3.6), is a small, lightweight, triaxial accelerometer-based activity recorder. The WHI PAM's small form factor allows it to be easily carried in a patient's pocket. The sample rate as well as the minimum acceleration threshold can be adjusted to ensure that data resolution requirements are met while optimizing for longer battery life. Time-series acceleration data is stored using an on-board flash memory card. Data transfer is achieved via USB on an internet-enabled PC. Using a patient's age, gender, height, and weight, the WHI PAM system calculates daily caloric expenditure based on the metabolic equivalents (METs) associated with approximations of the patient's activity levels throughout the day.

WANDA calculates the METs value (Table 3.1) based on activity information detected by the PAM device. Calories burned by each activity are calculated by the following equation

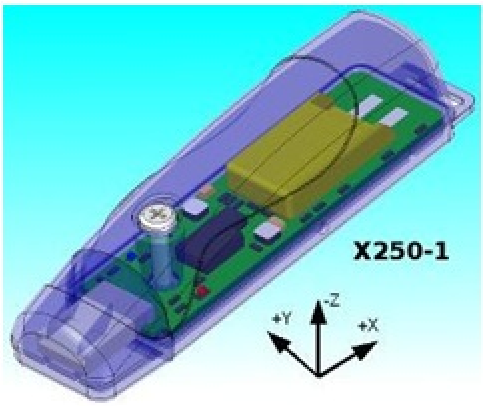


Figure 3.6: WHI PAM (Personal Activity Monitor)

(Equation 3.1) based on Jones's work [JC88].

$$Calories = ((METs \times 3.5 \times weight(kg)) \div 200) \times duration(minute) \tag{3.1}$$

The mobile version of WANDA provides an Android-based activity monitor application shown in Figure 3.7. It provides information about daily activity level, pedometer function and calorie expenditure. The method for estimating activity level is based on an algorithm proposed by Panasonic [YYN09][HMY06]. It is proven that the value calculated by this algorithm has high correlation (R^2=0.86) with the Doubly Label Water method, which is one of the most accurate methods for evaluating total energy expenditure under a free-living condition. Metabolic equivalent task (METs) level of physical activities and approximated calories burned is calculated using Equation 3.1 and Equation 3.2.

$$K_m = \sqrt{\frac{1}{n-1}[(\sum_{i=0}^{n} x_i^2 + \sum_{i=0}^{n} y_i^2 + \sum_{i=0}^{n} z_i^2) - \frac{1}{n}\{(\sum_{i=0}^{n} x_i)^2 + (\sum_{i=0}^{n} y_i)^2 + (\sum_{i=0}^{n} z_i)^2\}]} \tag{3.2}$$

K_m value in Equation 3.2 has a high correlation with the actual total energy expenditure. The METs level can be found by the first order linear regression fit. n is the number of

Physical Activity	MET
Light Intensity Activities	< 3
Sleeping	0.9
Writing, desk work, typing	1.8
Walking, less than 2.0 mph (3.2 km/h), level ground, strolling	2
Moderate Intensity Activities	**3 to 6**
Bicycling, stationary, 50 watts, very light effort	3
Sexual activity (position dependent)	3.3
Callisthenics, home exercise, light or moderate effort, general	3.5
Bicycling, < 10 mph (16 km/h), leisure, to work or for pleasure	4
Bicycling, stationary, 100 watts, light effort	5.5
Vigorous Intensity Activities	> 6
Jogging, general	7
Callisthenics, heavy, vigorous effort	8
Running jogging, in place	8

Table 3.1: Activity Levels and METs Values

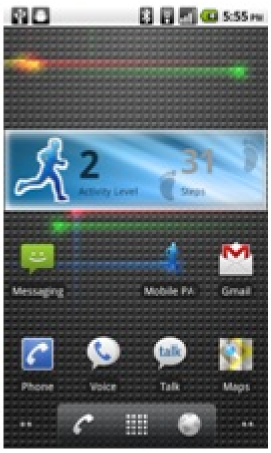


Figure 3.7: Activity Monitor in the Mobile Version of WANDA

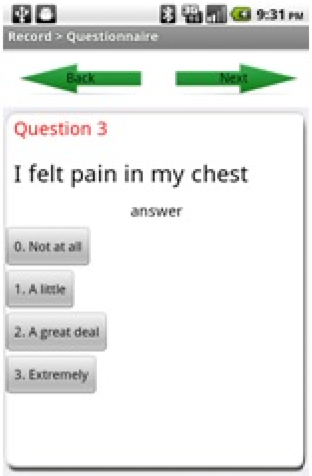


Figure 3.8: Daily Symptom Questionnaire Smartphone Application

samples in a given time window. X_i , Y_i , Z_i are accelerations in x, y, z directions at i^{th} sample. $\sum x$, $\sum y$, $\sum z$ are the summation of the accelerations within a time window. In our application, the sampling rate is about 20 Hz and the time window is one minute. Therefore, the number of samples, n, in one minute is 1200.

Each new sample contains three-axis acceleration and timestamp data measured every millisecond. A new K_m value is generated every minute, which is written onto a local SD card. The data recorded on the card is then transmitted to the database in the second tier via WiFi or 3G connections.

3.1.5 Daily Symptom Questionnaire System

Based on the schedule set by doctors, WANDA has patients answer a questionnaire (Table 1.1 for CHF and Table 3.2 for diabetes) via an SMS survey system, the Ideal Life Blood pressure or glucose monitor or an Android-based application (Figure 3.8).

The questionnaire is given for the purpose of checking for symptoms related to chronic diseases and health status. The corresponding user responses are collected and recorded

Questionnaire Items
Q1. Have you had any blood sugar readings < 80 or > 200?
Q2. Have you missed doses of your medication?
Q3. Today, is your health, good, fair or poor?
Q4. Compared to yesterday, are you feeling better, about same, or worse?

Table 3.2: Diabetes Questionnaires

by the database in the second tier. The SMS system sends a text message to the users to which they can reply to in the same way as they would reply to texts from friends. For displaying questions and collecting corresponding answers via Ideal Life BP and Gluco Manager^{TM}, healthcare providers are required to register questions on the web to make the devices programmed remotely. The Android-based questionnaire application asks questions to the users and users can answer by using the touch screen. The smartphone application has a daily reminder function. If the function is enabled, users receive a reminder notification if they forget to complete all questions every 24 hours. The answers obtained from the smartphone application are stored locally in the SD Card and uploaded to the server via WiFi or 3G.

3.1.6 Fall Detection Application

Because most of the body movements are constrained within frequency components below 20Hz, and 99% of the energy is contained below 15 Hz [AM85], a 40Hz sampling rate is sufficient for fall detection. The algorithm we used to detect falls can be categorized into three sequences.

1. Freefall detection.

2. Impact detection.

3. Orientation change detection.

In a fall, these three events happen consecutively. The algorithm detects falls if at least two consecutive signal vector magnitude (SVM) are above a defined threshold. The SVM essentially provides a measure of movement intensity [KNM06].

Theoretically, the SVM of freefall is zero. A freefall is detected if SVM is under a defined threshold. The orientation is calculated from the acceleration based on the transformation equations in Equation 3.4. The change of the orientation is computed by the equations in Equation 3.5. The thresholds of freefall, impact, and orientation change are defined based on the desired sensitivity.

When a fall is detected, the application uses light, sound, and vibrations to get the users attention. Liquid crystal display (LCD) and light-emitting diode (LED) indicators will be activated and flash, a phone will vibrate for 5 seconds, and a high-pitched ring tone will go off. The users have one minute to cancel the alarm by clicking a pop up box on the phone. If the phone is not touched in 60 seconds, it sends an emergency message to the caregivers. The message includes the users name, the alarm triggered time, and the possible location (determined by GPS data). If the user is indoors, a cellular network determines the location. By using Google Geocoding technology [Goo11], the phone can map the latitude and longitude coordinates returned by GPS to a real street address, which provides more easily useable and understandable information for caregivers. Based on initial experiments, however, the street address can be off up to 3 street blocks if the user is indoors.

$$SVM = \sqrt{x_i^2 + y_i^2 + z_i^2} \tag{3.3}$$

$$\begin{aligned} &\text{pitch angle of x-axis relative to horizon} = \text{atan2(gx, } \sqrt{gy^2 + gz^2}) \\ &\text{roll angle of y-axis relative to horizon} = \text{atan2(gy, } \sqrt{gx^2 + gz^2}) \\ &\text{yaw angle of z-axis relative to gravity} = \text{atan2(} \sqrt{gx^2 + gy^2}, \text{gz)} \end{aligned} \tag{3.4}$$

$$
\begin{aligned}
\text{Orientation Change} = \quad & \mid \text{difference of "pitch" before and after the impact} \mid \\
& + \mid \text{difference of "roll" before and after the impact} \mid \\
& + \mid \text{difference of "yaw" before and after the impact} \mid
\end{aligned} \tag{3.5}
$$

3.1.7 Reminder Function

As mentioned in Section 1.2.2, remote patient monitoring systems often suffer from system non-use [CBM07][SEC10]. In order to reduce the number of missing data, we developed a reminder function on the mobile version WANDA. If the function is activated, users receive a pop-up reminder if they forget to measure values and complete all questionnaires.

3.1.8 Communication Center

As the phone line version system supports Bluetooth, the components can communicate within a range of up to 10 meters. When the Ideal Life Pod receives data from the weight scale and blood pressure monitors, it transmits data to the database system through a standard phone line via a long-distance phone service plan or an Ethernet connection.

The software application in the smartphone version is implemented on the Motorola Droid Smartphone™. In terms of the Bluetooth device and smartphone interface, all of the devices act as masters that initiate Bluetooth communication with an Android phone. As the Android phone acts as a slave node, only one connection can be established at each time and a piconet is not possible. However, this is not an issue in the mobile version of WANDA, as patients will not be operating two sensing devices simultaneously if they are following standard usage. The communication between the phone and the medical server is through Wi-Fi or 3G networks. The smartphone transfers this data to a networked server and also stores this data on a local SD card. If an upload procedure fails due to a problem with the network, the application will retry the upload when the network is available.

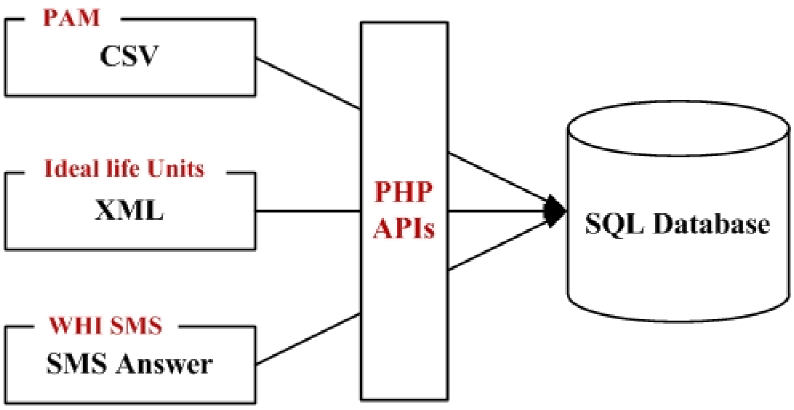


Figure 3.9: Data Abstraction in WANDA

3.2 Server Tier

Data collected from the first tier are sent to several web servers to store data and provide monitoring applications. Unlike the mobile version, the phone line version draws data from several different web servers. As Ideal Life, PAM, and the WHI SMS system use incompatible data types and different databases, analysis of data involves drawing data from different databases. Incompatible data formats in different databases are solved through usage of an abstraction of file formats and a shared ID table. The server also performs additional data processing for calculating specific variables from several different database sources. In addition, when the obtained values are out of the acceptable range, the system sends alert messages to healthcare providers via text message or e-mail.

3.2.1 Abstraction of Data Formats

The objective of abstraction in programming is to separate behavior from implementation. To allow changing implementations without affecting users, we should change the representation without having to change all programs by encapsulating the representation. If an

```
<ArrayofBPMReading>
    <BPMReading>
        <CilnetUserID>  </ClientUserID>
        <ReadingDate>   </ReadingDate>
        <ReadingTime>   </ReadingTime>
        <Systolic>      </Systolic>
        <Diastolic>     </Diastolic>
        <HeartRate>     </HeartRate>
    </BPMReading>
</ArrayofBPMReading>
```

Figure 3.10: Blood Pressure Data Format in the Ideal Life File System

implementation is encapsulated, other modules do not have to depend on its implementation details [Lis87].

A data abstraction procedure is required as the phone line version WANDA draws data from files systems that support different file formats such as CSV, XML and text (Figure 3.9). For example, in order to insert data stored in an Ideal Life file system (Figure 3.10) into a SQL server, data should be parsed and abstracted. Thus, we developed a PHP API to draw XML data from the file system and insert into the WANDA SQL server system.

Data abstraction facilities in WANDA web servers offer users the advantages of encapsulation. The representation of abstract objects can be changed without requiring corresponding changes in original files that manipulate the objects. In order to update data, the WANDA system calls other databases and file systems periodically and extracts data from them.

3.2.2 Data Integrity

The need to combine records from different systems or healthcare organizations exists for many reasons, such as patients moving or changing healthcare providers. When the number of available repositories and analyses increases, linking information between them becomes

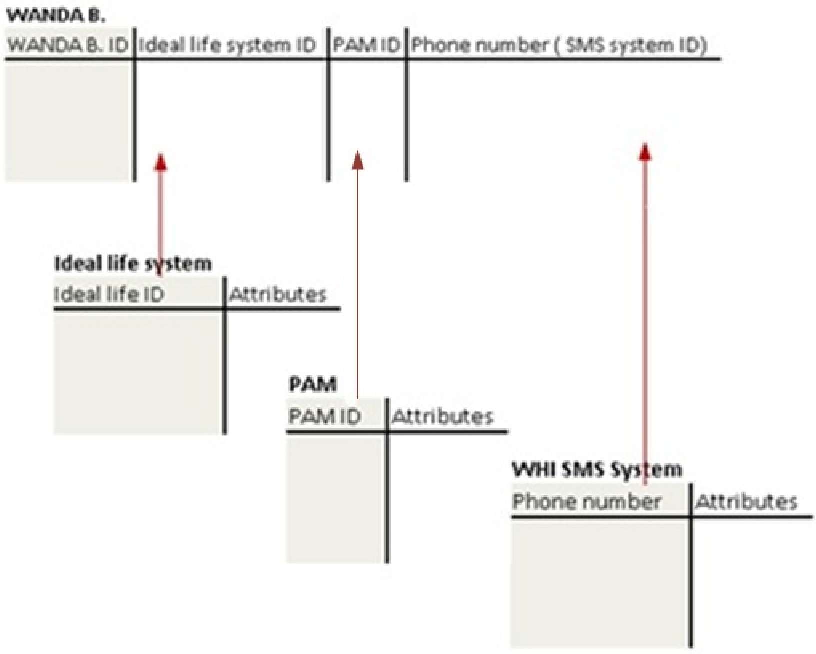


Figure 3.11: Database Structure in WANDA and a Shared ID Table

a major concern. Finding the relevant resources and making connections between their content and analysis output poses significant challenges [PTM09]. To make all patients medical records accessible to care providers, Bell [BS01] links electronic medical records together using a massively distributed Master Patient Index (MPI). An MPI is a facility that correlates and references patient identifiers and performs matches.

WANDA is an integrated architecture of Ideal Life system units, WHI PAM devices, and the WHI SMS system. Each system has its own database, patient ID, and item ID to distinguish the inputs measured by patients. As WANDA can use different system units and add more devices, finding relevant resources and making connections among them is essential.

The concept of semantic integration comes from the world of business and electronic commerce, where similar problems of legacy software and complicated data exist [Noy04]. To interpret the real mining of data from multiple sources, finding, establishing, and maintaining connections between relevant tools and most up-to-date data is significant. For solving the

```
SELECT weight
FROM WANDA, Ideallife system
WHERE WANDA.ID = \* A Given ID *\AND
WANDA.Ideallife system ID = Ideallife system. Ideallife ID
```

Figure 3.12: SQL Query for Integrating Data from Different Systems

data integration problem in distributed systems, a shared ID table is used in the WANDA system. To join keys used in the Ideal Life systems, the WHI PAM, and the WHI SMS systems, the WANDA ID is used as a primary key. Other input in different system databases are referenced via the WANDA ID table and original database tables (Figure 3.11), which enables the linkage of information located in several systems to a specific patient.

For example, when a user tries to access weight data, the WANDA system draws data from the Idea Life system using the shared ID table. Therefore, the system should execute the SQL query in Figure 3.12 to perform the given task.

3.2.3 Additional Data Processing

Based on requests, APIs were designed on the server side. The APIs perform data processing by using imported data from several different databases. For example, in order to calculate the calorie expenditure in Equation 3.1, the most recent weight value from the database was extracted. The designed API calculated the calorie expenditure using the METs and weight value with matching timestamps (Figure 3.13). Also, the duration of a certain METs value during a day can be calculated on the second tier of the system.

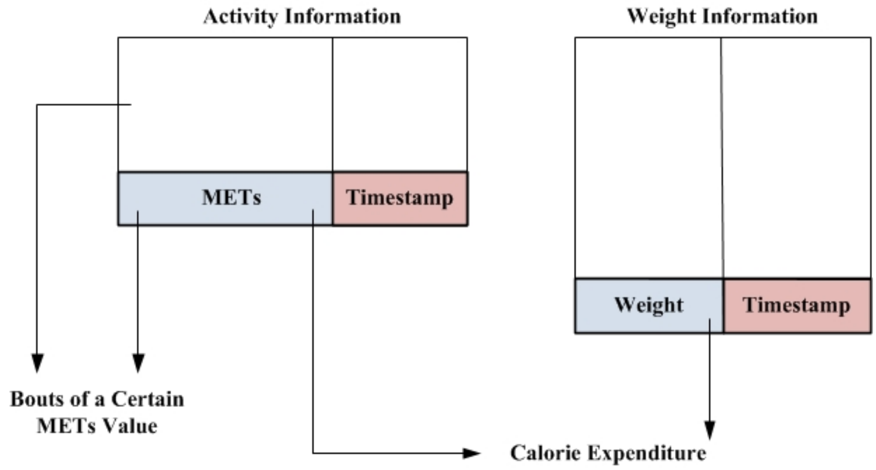


Figure 3.13: An Example of Additional Data Processing on the Web Server Tier

3.2.4 Web and Mobile Applications

WANDA provides two different monitoring applications for healthcare providers, and they operate on the web server tier. One is a web application and the other is a smartphone application. Through the monitoring applications (Figure 3.14, Figure 3.15a and Figure 3.15b), a healthcare provider can review the data and leave comments and annotation of collected data. In other words, doctors and nurses can see call and communication logs with patients and also discuss their plans and concerns via the applications. Below are samples of call log and discussion data of a patient in an intervention group.

Called at 4:20pm (Call time: 4 min) - took first steps on Christmas day. Has been checking BG mostly 100s, occasional 200s during holiday meals. Has not received shipments of new meter or supplies yet. Will email to the person in charge.
Pt was re-admit to UCLA ED had 2 stents placed - still having pain - seeing outside cards - taking lantus bid and nov tid. call time 5 min.

The interfaces of applications are clear, consistent, and communicative, so users do not need to worry about underlying file types, software architectures, and operating systems. Data

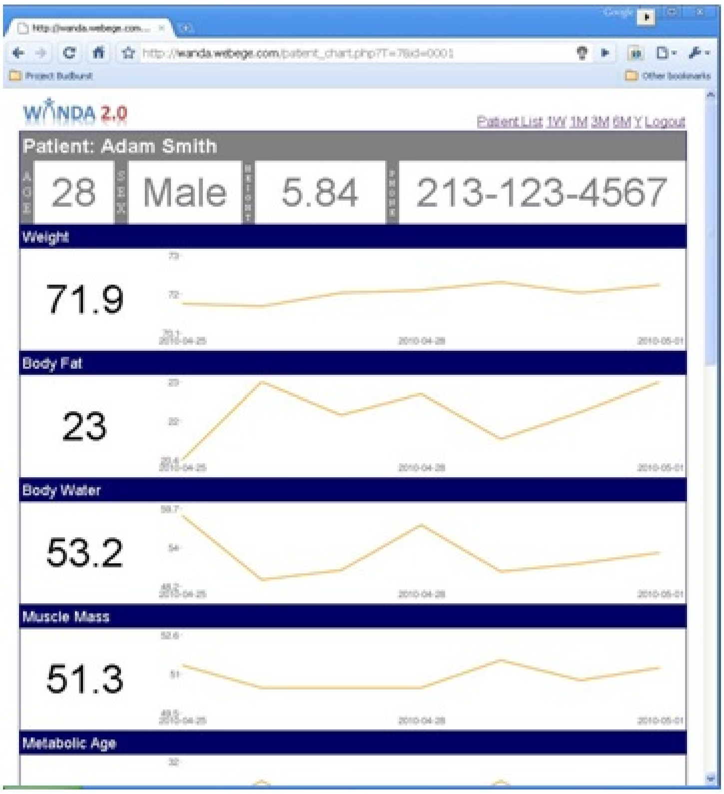


Figure 3.14: WANDA Web Application

can be viewed as graphs or tables. The graph mode shows the amount of calories burned in hourly-based format and other values in daily-based format using the length of the bar, spectrum of colors, and annotation. The structure of the website and data display modes, fonts, and graph colors were defined in the meeting with caregivers who are users of this system.

Access to the web application requires user verification in order to prohibit access by unauthorized users. The initial page asks for a user ID and a password that are provided to individuals in a private meeting. Healthcare providers can also add more patients and synchronize data via the WANDA webpage if necessary.

The iPhone's 3.5-inch multi-touch display with 480-by-320-pixel resolution lets users navigate by touching the screen. The iPhone 3G uses a technology protocol called HSDPA

(a) WANDA iPhone Application (b) WANDA Android Application

Figure 3.15: WANDA Smartphone Applications

(High-Speed Downlink Packet Access) to download data quickly over UMTS (Universal Mobile Telecommunications System) networks. In addition, its weight is only 135g, making it easy to carry. Therefore, the WANDA iPhone application helps users access the WANDA web database data. Using a WANDA application on an iPhone enables caregivers to monitor patient status whenever and wherever they want. In the table mode, healthcare providers can check their patients' daily question answers, weight, blood pressure, heart rate, blood glucose, and calorie expenditure. The table mode also provides statistical information. The graph mode displays each value as a color bar (Figure 3.15a).

We also built the WANDA system on the Droid$^{\mathrm{TM}}$, another popular smart phone platform (Figure 3.15b). The Droid One has a 3.7-inch screen with a 854-by-480-pixel resolution. The Droid One has weight similar to the iPhone's weight, 169g, and it also uses the same HSDPA protocol. Moreover, as the Droid One's Google Android platform allows users to use Bluetooth API, we can directly get data from healthcare devices. When users start the application, they choose between 'record' and 'history' options. If a user chooses the record option, they are brought to a screen where they can choose the specific measurement they wish to take. After choosing the measurement type and following the instructions on the phone, the user can see the numerical data on the screen and the recorded data will be uploaded to the server. For the questionnaire option, the user can answer registered questions and the answers will directly transmit to the web server. If the user presses the 'history' option, he is given the option of choosing to view a graph of weekly or monthly data. This function would assist users in monitoring their own personal health status.

Additionally, the WANDA web application includes a basic statistical analysis tool to verify the test result and the effectiveness of the clinical trial. This function includes Wilcoxon rank test, logrank test, t-test, etc. which are widely used in many randomized trials [PPA77]. The statistical analysis tool provides several values including p-value and z-value when the user enters values.

3.3 Back-end Database Tier

As data in the WANDA system is critical and personal, the loss of any data must be actively guarded against. If there is data loss on the server, the system cannot evaluate the patient's status accurately and demonstrate the system's effectiveness. Therefore, a back-end server performs data backup and recovery was developed. Data backups are useful for restoring state following a disaster or restoring small numbers of files after they have been accidentally deleted or corrupted.

In addition, data in the third tier is used for data analysis including task optimization, missing data imputation, adverse event prediction, etc. [SWL11] [LSA12] [SMW12]. Based on the data and annotations collected from the sensor and server tier, the analytics engine can generate multiple statistical models using various machine learning and data mining algorithms including association rule mining and classification. These models can then be used for both diagnostic and prognostic purposes. For example, in the case of heart failure patients, it is highly desirable to be able to predict the worsening of symptoms before the patient is actually hospitalized.

The third tier performs data pre-processing before performing data analytics. Once data are transmitted to the server, basic preprocessing and dimensionality reduction algorithms including data labeling and clustering are executed prior to data analytics. Data cleaning and signal transformations are the main goals of this pre-processing step. Data cleaning consists of dropping erroneous or non-necessary data attributes. Signal transformation including data discretization and quantization is utilized for efficient processing and prediction in data analytics.

The analytics process normally consists of two stages. Firstly, the data are downloaded and analyzed offline based on various hypotheses. Once a strong model has been generated and validated, it can then be uploaded to the server to perform real-time analytics. Finally, when the algorithm detects a pattern that is strongly associated with a predictable user action, a predicted outcome of missing data or an undesirable outcome, real-time feedback is dynamically provided to remote health monitoring systems, patients and healthcare professionals.

3.3.1 Data Backup and Recovery

WANDA applies an offline backup method that performs data backup when there are no current updates [MN93]. The data entry times are statistically analysed and the system sets a backup and recovery schedule. When a database backup occurs, all unfinished file backups are checked for completion before the new backup procedure is executed. The DBMS communicates with the client application to support transactional properties of data, coordinated backup and recovery. The data backup log file in the client application maintains the last modification timestamp, distributed database information, the unique ID of the last updated data, and other various components. To support coordinated backup and recovery with the DBMS, the application interfaces with the back-end server through WANDA APIs. In order to perform a backup of distributed web databases, the client application takes a

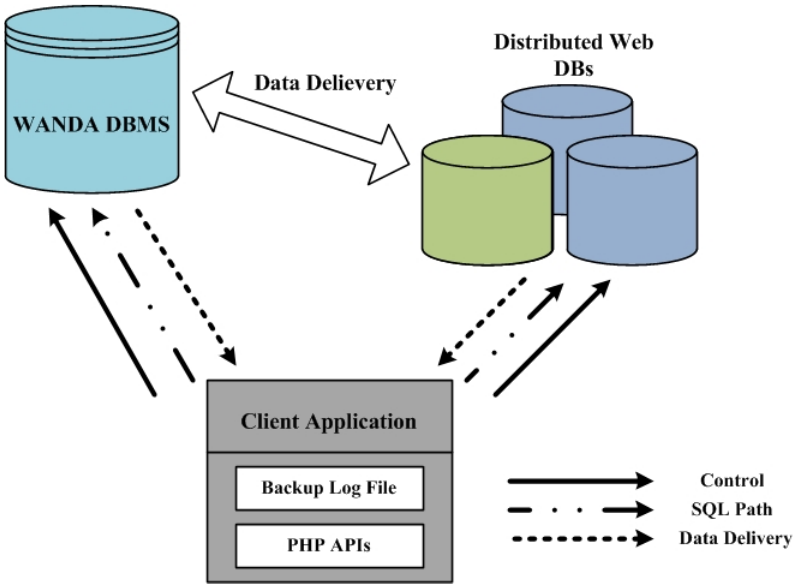


Figure 3.16: WANDA Data Backup System Configuration

central role (Figure 3.16). Using its backup log file and PHP APIs, it controls data delivery between distributed web databases and the DBMS. The developed APIs execute SQL queries, deliver data, manage database backup, and enact recovery procedures.
Just like DATALINK [BMB02], the WANDA platform utilizes a backup copy of the database for restoring the database to one of four possible consistent states.

- Offline Backup State (OBS): The state when an offline backup was taken.
- Quiescent Point State (QPS): The state when no update was being allowed to the database.
- Point-in-time State (PTS): Some arbitrary point-in-time state.
- Current Time State (CTS): The state at the time of crash.

3.3.2 Task Optimization and Early Adaptive Alarm

As part of the data analytics in the third tier, WANDA performs data transformation for quantizing sensory data readings and executes data association rule mining. Instead of using experts' knowledge, WANDA finds data clusters and their ranges in order to discretize timestamps and health-related readings and answers. Association rule analysis is a method to find interesting relations among attributes in large data sets. Rules are derived using previously collected data to help predict the current or future behaviors and health status of a patient.

First order logic from association rule mining can be used for patient task optimization and emergency event prediction, as shown in Figure 1.2. Association rule mining and its feedback are used for reducing the number of tasks required by patients while increasing information gain. One of the advantages of the task optimization step is improving patients adherence to remote health monitoring and enhancing missing data imputation results by obtaining more data or finding underlying data relationships. After task optimization for improving patient participation, missing data imputation techniques can be applied to WANDA in order to predict missing values and provide alarms to healthcare professionals when the predicted missing values are out of acceptable range [SWL11]. In addition, association rules related to emergency events can be used to generate early adaptive alarms and guidance to prevent emergent events. For example, analyzing past vital signal data can help find trends of readings, which cause adverse events such as hospitalization.

3.3.2.1 Data Discretization

Sensor readings and corresponding timestamps in remote health monitoring are continuous signals. To reduce the dimensionality and complexity of processing continuous numeric and timestamp data, it is necessary to discretize the data. Under supervised settings, patients might have regular schedules on meal, exercise and sensor data measurements such as having a meal and measure blood glucose at 6:30 am, 11:30 am and 5:00 pm [MDD69]. However, in unsupervised environments such as remote health monitoring, patients can have more flexible schedules and freedom, so it is hard to categorize time frames as morning, afternoon and evening. In addition, as patients with chronic diseases generally have readings that are abnormal or in an out-of-acceptable range, their readings can be more biased and may not follow normal distribution or standards [HFJ96]. Additionally, readings can have different distributions depending on the time of measurements. For example, results of Jarrett's study show that blood glucose levels in the afternoon or evening are generally higher than in the morning [JBK72]. As health-related readings vary depending on the time of day, discretizing sensor data equivalently for different time intervals can result in biased results.

Therefore, it is necessary to categorize and quantize data using data-driven methods instead of experts' knowledge or human intuition. In this dissertation, I assume that the distribution of collected sensor data is a mixture of Gaussian distribution and applied expectation maximization algorithm to cluster data into the pre-defined numbers of bins. The expectation maximization (EM) algorithm [Moo96] [Har58] is an iterative method to optimize the estimation of unknown parameter , given measured variables U and unmeasured variables J. The objective of the EM algorithm is maximization of the posterior probability (Equation 3.6) of the parameter given U and J.

$$\Theta^* = argmax_\Theta \sum_J P(\Theta, J|U) \tag{3.6}$$

The EM algorithm consists of two steps: The expectation step (E step) and the maximization step (M step). The E step finds a local lower-bound to the posterior distribution while the M step optimizes the bound obtained from the E step using iteration. In the E step, the algorithm calculates the expected value of the log likelihood function under the current estimate of the parameters Θ_t of the conditional distribution J given U. This step finds the best lower bound, $B(\Theta|\Theta_t)$.

$$B(\Theta|\Theta^t) = \sum_J f^t(J) \log \frac{P(U, J, \Theta^t)}{f^t(J)} \tag{3.7}$$

while $f^t(J) = \frac{P(U,J,\Theta^t)}{\sum_J P(U,J,\Theta^t)} = P(J|U, \Theta^t)$. The M step iterates and chooses Θ^{t+1} by maximizing the bound, $B(\Theta|\Theta^{t+1})$ from the E step.

$$\Theta^{t+1} = argmax_\Theta B(\Theta|\Theta^t) = argmax_\Theta [Q^t(\Theta) + \log P(\Theta)] \tag{3.8}$$

while Q^t $(\Theta)$is the expected complete log-likelihood, $\log P(U, J|\Theta)$ and $P(\Theta)$ is the prior on the parameters Θ.

The developed algorithm quantizes timestamps of sensor readings and sensor readings in each time range are discretized. Based on mean (μ) and standard deviation (σ) values of each Gaussian curve from the EM algorithm, the developed algorithm finds solution of simultaneous equations and these points are used for discretizing time and sensor data ranges.

$$\begin{cases} y = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2} \\ y = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu_2}{\sigma_2})^2} \end{cases} \tag{3.9}$$

3.3.2.2 Data Association Rule Mining

Association rule analysis is a method to find interesting and strong association among attributes in large data sets. One example is affinity analysis to find the purchase behavior of different groups of consumers and their market baskets [MH03]. The results of affinity analysis can be used for arranging items in the store, planning store promotions, etc.

In remote health monitoring for patients with chronic diseases, patients' health status changes dynamically, but health-related readings are correlated [KEL02] [VLL01]. Therefore, finding trends and associations of patients data can help to reduce the number of tasks, decide the order of tasks, and even enable to provide early adaptive alarms to prevent emergency situations.

In WANDA, rules are derived using previously collected data to help predict the status and behavior of a patient. The WANDA implementation uses data collected within a dynamic sliding window w determined by the Apriori algorithm (Algorithm 1) as a main association rule mining and the confirmation-based association rule mining algorithm (Algorithm 2) as a supplementary algorithm before the current or future measurement. Association rule mining and its feedback are used for reducing the number of tasks required by patients and predicting future outcomes while increasing information gain. In the data preprocessing step, the developed algorithm performs data cleaning and discretization for removing erroneous data and discretizing timestamp and indexing data (see Section 3.3.2.1). The system also indexes sensor readings and questionnaire response data as multiple measurements and system non-use. Additionally, information on whether a caregiver contacted the patient for each day is used.

Apriori-based Data Association Rule Mining After preprocessing data, the developed algorithm applies the Apriori algorithm [AS94] to derive rule sets (Algorithm 1). The discretized and categorized data are used in the algorithm as inputs and rules are derived by looking back a variable number of days (sliding time window). The time window is increased for each iteration. The algorithm calculates the support and confidence of each implication and chooses implications qualifying threshold limits. In each subsequent pass, the large item sets found in the previous step are used to generate the candidate sets (the largest item sets). The results of each step are large item sets of qualifying minimum support and confidence

$Result := \emptyset$;

$k := 1$;

while $C_k \neq \emptyset$ **do**

  Create a counter for each itemset in C_k;

  foreach *events in database* **do**

    Increase the counter of itemsets in C_k which appears in the events;

  end

  $L_k :=$ All candidates in Ck with support s_{min} ;

  $Resuit := Result \cup L_k$;

  $C_{k+1} := k + l$-itemsets which have all subsets of L_k ;

  $k := k + l$;

end

Algorithm 1: Apriori Algorithm.

in the given time window.

Let $I = \{i_1, i_2, , i_m\}$ be a superset of all possible task outputs. Let D be a set of events such that $D \subset I$. An association rule is an implication of $A \Rightarrow B$ where $A \subset I$, $B \subset I$ and $A \cap B = \emptyset$. Confidence c means that $c\%$ of events in D contain A and B. Support s indicates $s\%$ of events in D contain A or B. The developed algorithm requires generating association rules that have support and confidence greater than the user-defined thresholds, minimum support (s_{min}) and confidence (c_{min}).

Confirmation-based Data Association Rule Mining Another data association rule mining algorithm used in this dissertation is the Tertius algorithm that is a confirmation-based algorithm. The Tertius algorithm is an unsupervised learning method, utilizing a heuristic measure of confirmation using novelty and satisfaction to distinguish $\forall(X \rightarrow Y)$ and $\forall(Y \rightarrow X)$. The developed algorithm induces rules that is not only usual and common, but also unusual and interesting that sometimes expert knowledge does not cover as Apriori does.

The candidate rule in Algorithm 2 is used for refining rules. Possible ways to refining rules are

1. Add a new literal.
2. Unify two variables.
3. Instantiate a variable with a constant.

When n_{ij} is an observed frequency, an expected frequency μ_{ij} is defined as Equation 3.10.

$$\mu_{ij} = \frac{n_{i*} \times n_{*j}}{N} \tag{3.10}$$

Pearson statistics X^2 is used for checking how μ_{ij} and n_{ij} are close and defined as Equation 3.11.

$$X^2 = \sum_{ij} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} = \sum_{ij} \frac{{n_{ij}}^2}{\mu_{ij}} - N \tag{3.11}$$

Dividing X^2 by N is a measure of the strength of the dependency between X and Y, and it ranges between 0 (independent) and 1(dependent).

$$\Phi^2 = \frac{X^2}{N} = \sum_{ij} \frac{(n_{ij} - \mu_{ij})^2}{N \times \mu_{ij}} \tag{3.12}$$

To distinguish $\forall(X \to Y)$ and $\forall(Y \to X)$, which have the same number of confirming instances, $n_{\bar{Y}X}$, counterinstances of the rule $\forall(X \to Y)$, can be considered. The novelty of a rule $\forall(X \to Y)$ is defined as $\Delta_{\bar{Y}X} = \mu_{\bar{Y}X} - n_{\bar{Y}X}$ and the satisfaction of a rule $\forall(X \to Y)$

is defined as $\sigma_{\bar{Y}X} = \frac{\mu_{\bar{Y}X} - n_{\bar{Y}X}}{\mu_{\bar{Y}X}}$. As the satisfaction is a version of rule accuracy $p(Y|X)$, we want high novelty and satisfaction and the product of these two values, $\frac{(\mu_{\bar{Y}X} - n_{\bar{Y}X})^2}{\mu_{\bar{Y}X}}$ which is same as the sum of Φ^2 with $\bar{Y}$ and X.

$$\Phi^2 \geq {\Phi_{\bar{Y}X}}^2 = \frac{X^2}{N} = (\frac{\mu_{\bar{Y}X} - n_{\bar{Y}X}}{\sqrt{\mu_{\bar{Y}X}} - \mu_{\bar{Y}X}})^2 \tag{3.13}$$

The degree of confirmation of a rule $\forall(X \rightarrow Y)$ is defined as $\Phi_{\bar{Y}X} = \frac{\mu_{\bar{Y}X} - n_{\bar{Y}X}}{\sqrt{\mu_{\bar{Y}X}} - \mu_{\bar{Y}X}} \leq \frac{1 - p_{Y\bar{X}}}{1 + p_{Y\bar{X}}}$.

$$\begin{aligned} \Phi_{\bar{Y}X} &= \tfrac{\mu_{\bar{Y}X} - n_{\bar{Y}X}}{\sqrt{\mu_{\bar{Y}X}} - \mu_{\bar{Y}X}} \\ &\leq \tfrac{\mu_{\bar{Y}X}}{\sqrt{\mu_{\bar{Y}X}} - \mu_{\bar{Y}X}} \\ &= \tfrac{\sqrt{\mu_{\bar{Y}X}}}{1 - \sqrt{\mu_{\bar{Y}X}}} \\ &= \tfrac{p_{\bar{Y}} + p_X}{2 - p_{\bar{Y}} - p_X} \\ &= \tfrac{1 - (p_{Y\bar{X}} - p_{\bar{Y}X})}{1 + (p_{Y\bar{X}} - p_{\bar{Y}X})} \\ &\leq \tfrac{1 - p_{Y\bar{X}}}{1 + p_{Y\bar{X}}} \end{aligned} \tag{3.14}$$

Therefore, The optimistic estimate of the confirmation (OEC) value of a rule $\forall(X \rightarrow Y)$ is $\frac{1 - p_{Y\bar{X}}}{1 + p_{Y\bar{X}}}$ and this can be the low boundary of confirmation values.
In Algorithm 2, rules with a good optimistic estimate are refined first, so it is likely that one of the children will have a good confirmation. The children are considered as soon as they are generated, so they can be added to the results as soon as possible.

Main Loop in Data Association Rule Mining When Apriori and Tertius returns first order logics, $A \Rightarrow B$, the main loop in Algorithm 3 calculates their contrapositive rules, $\neg B \Rightarrow \neg A$ for the Apriori algorithm and conditional probabilities of all generated first order logics. As the Tertius algorithm in Section 3.3.2.2 calculates the optimistic estimate of the confirmation to distinguish $\forall(A \rightarrow B)$ and $\forall(A \rightarrow B)$, the main algorithm does not consider

```
CandiateRule := ∅;
while CandiateRule ≠ ∅ do
    rule := first rule;
    if rule can be refined then
        RefinedRule := refine rule;
        foreach RefinedRule do
            Calculate Optimistic Estimate of Confirmation (OEC), (1-p_{YX̄})/(1+p_{YX̄}) ;
            if OEC > OEC threshold then
                Add RefinedRule to CandidateRule;
                Calculate Confirmation;
                if Confirmation > Confirmation Threshold then
                    add RefinedRule to Result;
                end
            end
            Sort CandiateRule according to OEC ;
        end
    end
end
Return Result;
```

Algorithm 2: Confirmation-based Association Rule Learning.

contrapositive rules for Tertius. Conditional probability p indicates $p\%$ of events contains B if A happens and is denoted as below.

$$P(B|A) = \frac{P(A \cap B)}{A} \tag{3.15}$$

Algorithm 3 only chooses first order logic with a minimum conditional probability (p_{min}). If the timestamp of the consequent in either implication (original rule or contrapositive rule) is larger than the timestamp of the antecedent and the implication is not a subset of any existing rules, the generated rule is added to the rule set. The process stops when there is no new rule and the algorithm returns the final rule set.

$Rule := \emptyset$;

$w := 1$;

while $A_w{}^* \neq \emptyset$ **do**

  $A_w := Result$ from Algorithm 1 with minimum conditional probability p_{min};

  $A'_w :=$ Contrapositive of A_w;

  $A_w{}^\dagger := Result$ from Algorithm 2 with minimum conditional probability p_{min};

  $A_w{}^* :=$ Subsets of $A_w \cup A'_w \cup A_w{}^\dagger$ which antecedents' maximum timestamp is smaller than consequents' minimum timestamp and not in Rule with smaller conditional probability;

  $Rule := Rule \cup A_w{}^*$;

  $w{:=}w + 1$;

end

Algorithm 3: Main Loop of Association Rule Learning Algorithm in WANDA.

The generated rules in *Rule* are prioritized based on the conditional probability values and applied to the remote health monitoring system. Using the implication rules, the system can reduce user tasks by monitoring necessary tasks and predicting unperformed tasks and can predict future events to provide early adaptive alarms and reduce emergencies. As the Apriori algorithm has excellent scale-up properties, the developed algorithm can be applied to the system for dynamically arranging daily patient tasks and providing early adaptive alarms depending on the size of dataset [AS94]. Additionally, the Tertius algorithm is used for generating additional rules that Apriori does not find and used as a supplementary algorithm.

3.3.3 Missing Data Imputation

This section uses the missing at random (MAR) hypothesis [SG02]. MAR assumes that missing data is dependent on observed data. Hence, missing data can be predicted by resident data.

3.3.3.1 Non-binomial Data Prediction

In order to predict missing answers, we exploit the projection adjustment by contribution estimation regression algorithm (PACE) (rounding any non-integer value returned by PACE) [WW02]. This method is based on maximum likelihood estimation (MLE) and an empirical Bayes framework to minimize the Kullback-Leibler (KL) distance between the original and the estimation function. First, the PACE algorithm transforms parameters using MLEs asymptotic normality property [FK85] to convert the original parameters. The algorithm utilizes the empirical Bayes estimator in Equation 3.16.

$$\widetilde{\theta_i^{EB}} = \frac{\int \theta f(x_i|\theta) dG_k(\theta)}{\int f(x_i|\theta) dG_k(\theta)} \tag{3.16}$$

where $\tilde{\theta}(x)$ is the estimator, $f(x_i|\theta_i)$ is a probability density function (PDF) and G_k is a consistent estimator of G which is the mixing distribution of the mixture $f_G(x) = \int(f(x|\theta)dG$. Using Equation 3.17, the developed algorithm minimizes the KL distance between f and $\tilde{f}$.

$$\Delta_{KL}(f, \tilde{f}) = E_f log(\frac{f}{\tilde{f}}) = \int \log(\frac{f}{\tilde{f}})f \tag{3.17}$$

This method especially shows better results in high dimensional data spaces and is applied to complete cases that have all answered questions to evaluate the accuracy.

3.3.3.2 Binomial Data Prediction

A binomial approach is used to predict alarms normally triggered by abnormal data values (e.g., drastic weight changes, unhealthy blood pressure, etc.) given missing data. For example, the system should trigger an alarm if a patient has an extreme change in sensor reading even when the value was not collected by WANDA. We use naïve Bayes, a Bayesian network, and VFI to detect such changes in order to alert caregivers.

Naïve Bayes and Bayesian network classifiers are algorithms that approach the classification problem using the conditional probabilities of the features [DH96]. A Bayesian network is a directed acyclic graph (DAG) over a set of variables X, where the outgoing edges of a variable x_i specifies all variables that depend on x_i. The probability of an outcome is determined as Equation 3.18.

$$P(X) = \prod_{x \in X} p(x|par(x)) \tag{3.18}$$

where $X = x_1, x_2, \ldots, x_k$ is a set of variables, and $par(x)$ is the set of parents of x in a Bayesian network. The probability of the instance belonging to a single class is calculated by using the prior probabilities of classes and the feature values for an instance. Naïve Bayesian method assumes that features are independent and there are no hidden or latent attributes in the prediction process. As such, the experimental results for naïve Bayes and Bayesian network can be slightly different as $p(class) = \frac{1+N(class)}{N(class)+N(instance)}$ for naïve Bayes $p(class) = \frac{\frac{1}{2}+N(class)}{N(class)\times\frac{1}{2}+N(instances)}$ for Bayesian network where $N(x)$ is the number of sets or instances.

VFI is a categorical classification algorithm and considers each feature independently as Bayes methods [DG97]. The classification of a new instance is based on a vote among the classifications built by the value of each feature. While training, the VFI algorithm constructs intervals for each feature. For the classification, a single value and the votes of each class in that interval are calculated for each interval. For each class c, feature f gives a vote values (Equation 3.19).

$$feature\_vote[f,c] = \frac{interval\_class\_count[f,i,c]}{class\_count[c]} \tag{3.19}$$

where $interval\_class\_count[f,i,c]$ is the number of instances of class c which is a member of interval i of feature f. The class with the highest total vote is predicted to be the class of the test instance.

In the Bayes methods, each feature participates in the classification by assigning probability for each class and the final probability of a class is the product of each probability measured on each feature. In VFI, each feature distributes its vote among classes and the final vote of a class is the sum of each vote given the features.

CHAPTER 4

Result

4.1 Clinical Trials

The WANDA study was approved by the UCLA institutional review board (IRB). IRB approves, monitors, and reviews biomedical and behavioral research involving humans with the aim to protect the rights and welfare of the research subjects. There are two completed CHF clinical trials, an undergoing CHF clinical trial and an undergoing study.

4.1.1 Congestive Heart Failure

As November 2009, the WANDA system has been used for health data collection on the intervention arm including 26 different congestive heart failure patients (*CHF1*) [SCW11]. The population of this first clinical trial is approximately 68% male; 40% White, 13% Black, 32% Latino, and 15% Asian/Pacific Islander, with a mean age of approximately 68.7 $\pm$ 12.1. The gender distribution and anticipated age of participants are representative of the incidence and natural history of CHF. The combined ethnic distribution across the sites is generally representative of the population in the United States. Patients are included in the study if

1. They have a diagnosis of CHF.
2. They were recently hospitalized for CHF exacerbation during the past 30 days prior to enrollment.

3. They had an age $\geq$ 65 years.

4. They lived independently.

They were excluded if

1. They had serious complicating comorbidities, untreated malignancies, or neurological disorders that impaired cognition.

2. They were unable to understand spoken English or Spanish.

3. They had hearing or vision loss that was deemed major or uncorrected.

They were all provided with Bluetooth weight scales, blood pressure monitors, landline or Ethernet gateways, and PAM devices. Daily questions (listed in Table 1.1) are delivered to each patient and monitored by WANDA. Each captured data instance for the study comprises 37 different attributes, including: timestamps; weight; diastolic/systolic blood pressure; heart rate; metabolic equivalents (METs); calorie expenditure; and numeric responses to twelve somatic awareness questions. Each data instance is gathered from each subject once a day. The total number of instances used in this study is 1090.
The second clinical trial of WANDA started on February 2011 with 18 low literacy Latinos with heart failure (*CHF2*). This population is disproportionately affected with CHF, is more likely to be hospitalized with CHF, and is at greatest risk for re-hospitalization, and dying from CHF. The population of the participants in this study is approximately 89% male with a mean age of approximately 54. Study participants were provided with Bluetooth weight scales, blood pressure monitors, landline or Ethernet gateways, and Android activity monitoring applications. Each captured data instance for the study comprises 33 different attributes, including: timestamps; weight; diastolic/systolic blood pressure; heart rate; metabolic equivalents (METs); calorie expenditure; and numeric responses to questions in Table 4.1. Each data instance is gathered from each subject once a day. The total number

Questionnaire Items
Q1. Are your feet more swollen than usual?
Q2. Did you use an extra pillow last night?
Q3. Are you more tired than usual?
Q4. Are you coughing more than usual?
Q5. Did you wake up with shortness of breath during the night?
Q6. Did you take your heart medications?
Q7. Did you walk or exercise today?
Q8. Did you eat a low-salt diet?

Table 4.1: Questionnaire Items in *CHF2*

of instances used in this study is 1030.

Finally, the system is currently deployed in a much larger study that targets the remote monitoring of 1500 patients of 50 years or older with heart failure problems (*BEAT-HF*). This an on-going project is conducted in collaboration with the UCLA Department of Medicine, UC Davis, UCSF, UCI, UCSD, and Cedar Sinai Hospital. Chronic heart failure patients who were hospitalized at any of the six participating medical centers are being considered to be recruited in either the control or intervention arm through a randomized trial process. Data collection for this study has started in November 2011 and is scheduled to end in April 2013. Patients measure their weights and blood pressures and reply to questionnaires on a daily basis, and the collected data are transmitted to the database via a phone-line, an Internet connection at home, or through cellular networks. We have currently enrolled more than 400 patients in the study.

The data from *CHF1* and *CHF2* are used for evaluating the system effectiveness for patients with CHF in Section 4.2 and for evaluating missing data imputation algorithms in Section 4.5.

Group	Total	Male	Female	Avg. Age
Intervention	21	18	3	48.13
Control	24	16	8	65.25

Table 4.2: Patient Population and Information in *Diabetes*

4.1.2 Diabetes

The fist clinical trial for diabetes started on June 2011 with patients with Type 2 Diabetes, HbA1c > 7.5 who were recently hospitalized (*Diabetes*) [SMW12]. Patients with active malignancy or those unable to provide informed consent were excluded. In this dissertation, we used data from 14 study participants assigned to the intervention arm (Table 4.2) and the average participation duration is 59.67 days. Patients in the intervention arm are required to measure their blood sugar up to three times a day (morning, afternoon and evening) and answer four questions per day (Table 3.2). Acceptable ranges of blood glucose and questionnaire values are denoted in Table 4.5. Existence of call logs between a patient and caregivers are labeled. The total number of instances used in this study is 1117 and each data has 54 attributes per day including patients profile.

The data from *Diabetes* are used for evaluating the system effectiveness and for evaluating data discretization and association rule mining algorithms for task optimization and early adaptive alarm in Section 4.3 and Section 4.4.

4.2 Effectiveness of Remote Health Monitoring

CHF1 has enabled patients to reduce 5.6% of weight and blood pressure values that are out of the acceptable range (Figure 4.1). Test results show that the number of measurements that were out of the acceptable range was decreased as patients use the WANDA system.

Items	**Values**
Weight Changes	$< +2$ (lb./ day)
Systolic Blood Pressure	> 90 mm Hg
Diastolic Blood Pressure	> 50 mm Hg
Heart Rate	< 90 and > 40 bpm
Q1	Not at all, A little
Q2	Not at all, A little
Q3	Not at all, A little
Q4	Not at all, A little
Q5	Not at all, A little
Q6	Not at all, A little
Q7	Not at all, A little
Q8	Not at all, A little
Q9	Not at all, A little
Q10	Not at all, A little
Q11	Not at all, A little
Q12	Not at all, A little

Table 4.3: Monitored attributes and acceptable ranges for *CHF1*

Items	**Values**
Weight Changes	$< +2$ (lb./ day)
Systolic Blood Pressure	> 90 mm Hg
Diastolic Blood Pressure	> 50 mm Hg
Heart Rate	< 90 and > 40 bpm
Q1	No
Q2	No
Q3	No
Q4	No
Q5	No
Q6	Yes
Q7	Yes
Q8	Yes

Table 4.4: Monitored attributes and acceptable ranges for *CHF2*

Items	**Values**
Blood Glucose	80 - 200 mg/dl
Q1	No
Q2	No
Q3	Good, Fair
Q4	Better, About same

Table 4.5: Monitored attributes and acceptable ranges for *Diabetes*

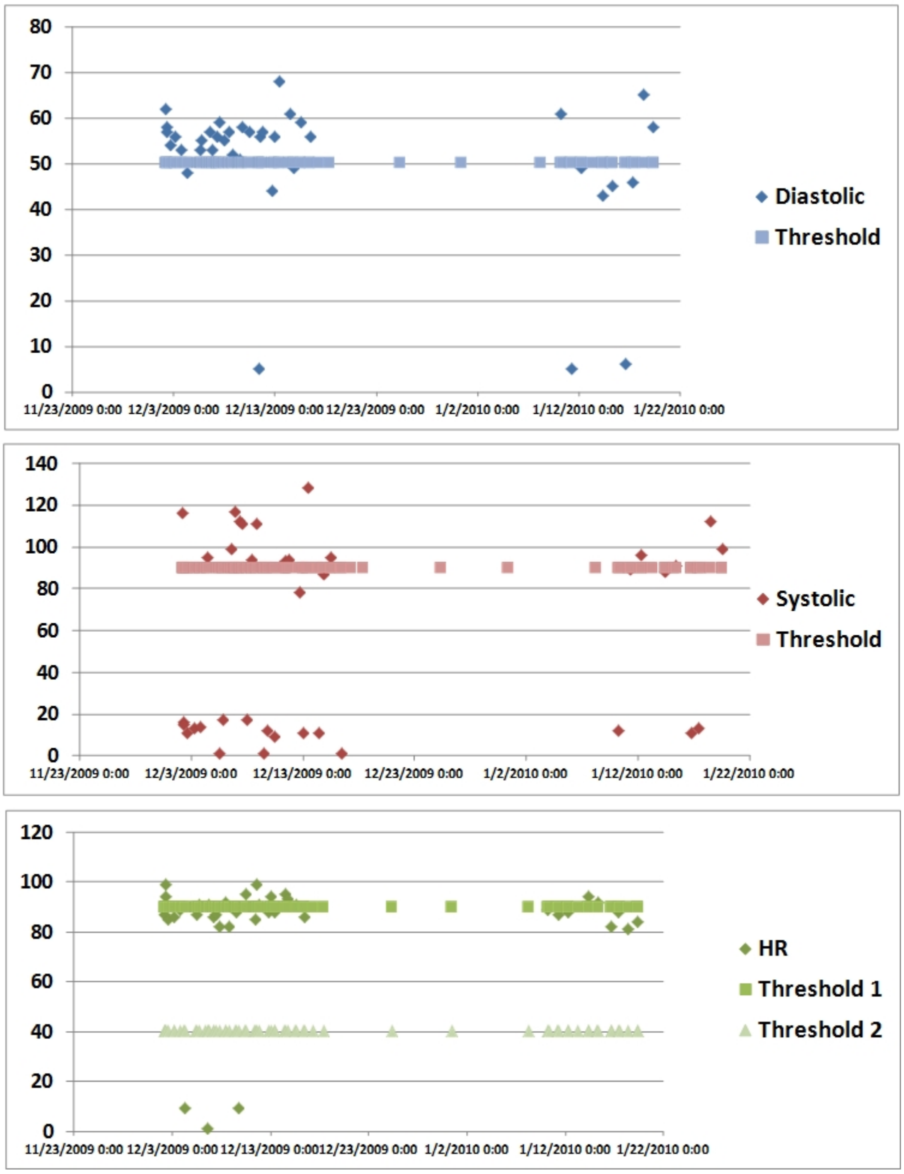


Figure 4.1: Changes in Diastolic, Systolic, HR values

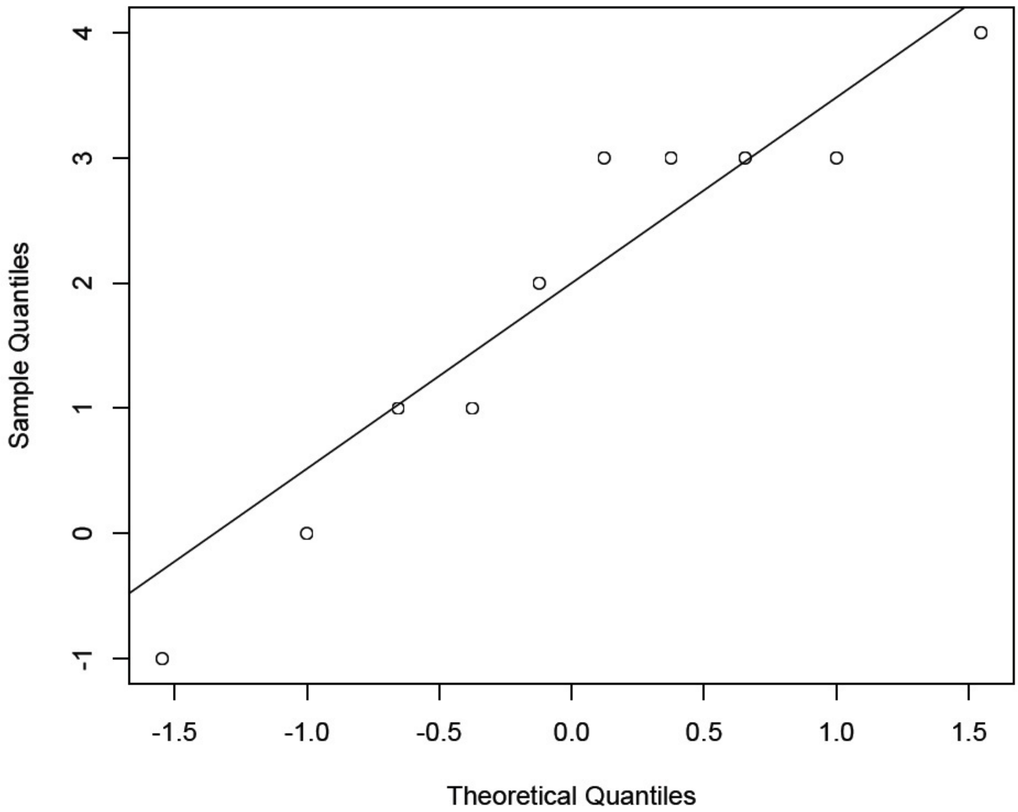


Figure 4.2: Normal Q-Q Plot for Weight Data

A paired t-test was used in order to compare paired result values where both observations are taken from the same subjects. In order to use a paired t-test, Q-Q plot is applied (Figure 4.2) for checking the normality of difference between the data. Those patients who participated in the test for more than 2 months were used in the analysis. The null hypothesis and the alternative hypothesis are given as below with 5% significant level.

Null Hypothesis (H_0) : The number of warnings during the first 15 days- The number of warnings during the last 15 days = 0.
Alternative Hypothesis (H_1): The number of warnings during the first 15days - The number of warnings during the last 15 days > 0.

Weight data received a t-value of 3.77 when applying the t-test method using the equation below. Therefore, the null hypothesis should be rejected for weight data with p-value=0.0022 which is less than 0.05.

$$t = \frac{mean\ difference}{standard\ deviation} \times \sqrt{n} \tag{4.1}$$

The t-test in Equation 4.1 assesses whether the means of two groups are statistically different from each other. This analysis is appropriate to compare the means of two groups, and especially appropriate as the analysis for our two-group randomized experimental design. For weight, the t-test results show that the WANDA study is effective for patients with CHF. Due to the high prevalence of obesity in the United States, Kenchaiah [KEL02] suggests that strategies to promote optimal body weight may reduce the population burdened with CHF. In addition, an increase in weight indicates the retention of excess fluid, which requires increasing the dosage of diuretic medication to counteract fluid accumulation. Therefore, weight loss is highly related to improving patient quality of life.

The clinical trial for low literacy CHF Latinos (*CHF2*) revealed high daily usage (>90%) of the WANDA wireless devices (weight scale, BP machine), and adherence to symptom reporting and self-care behaviors. This proves the usability and ease of use of WANDA, since the adherence rate is only about 50% for patients with chronic diseases in developed countries and lower than 50% in a real-world setting [Sab03].

The clinical trial with patients with diabetes (*Diabetes*) showed that the average phone calls to intervention group were 0.1366 a day, while 0.0107 to control group. The p-value of the study is 0.0032 for 95% confidence interval. Therefore, the preliminary result shows that the interactions from support networks were greater to the intervention group. As diabetes is a disease related to their daily habits such as monitoring symptoms, eating and exercise, sharing daily readings, providing information, technical and emotional supports

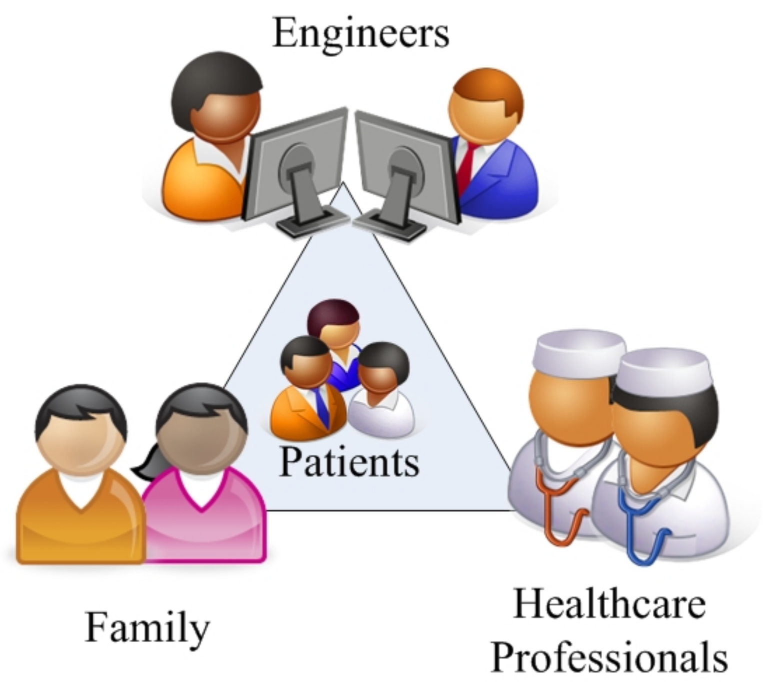


Figure 4.3: Collaboration Triangle in Remote Health Monitoring

from their family members, health care providers and even engineers in the collaboration triangle of remote health monitoring are important[WFD98] (Figure 4.3). The results of clinical trial shows the developed system and the designed clinical trial mechanisms helps promote conversation and collaboration among a patient and his/her support networks to encourage the patient and adhere their daily monitoring, since the WANDA helps caregivers know patients' health status in real time.

4.3 Data Discretization

In this section and the section below (Section 4.4), I utilized data from the *Diabetes* clinical trial.

To find the best discretization method for enhancing the data association rule mining results, I applied different discretization approaches using:

1. The experts' knowledge utilized in Hanefeld and Malherbe's studies ([HFJ96] [MDD69]), on blood glucose and timestamp accordingly

2. The EM algorithm

3. The combination of experts' knowledge and the EM algorithm

The number of bins used to discretize timestamps and blood glucose data are 3 for each attribute. The increment in sliding time window w) is 1 day per iteration.

For experts' knowledge-based discretzation ($EKTBD$), timestamps are categorized into three different time periods (T_i) and each timestamp period is T_1: 7:30:00-12:00:00, T_2: 12:00:00-16:30:00 and T_3: 16:30:00-21:00:00, and blood glucose readings are categorized into three different level (B_i) and each blood glucose level is B_1: <80 mg/dl, B_2: 80-200 mg/dl and B_3: > 200 mg/dl.

For discretizing timestamps only ($EMTD$), I applied the EM algorithm on collected timestamps of blood glucose measurements and applied experts' knowledge on blood glucose readings. Timestamps are discretized as T_1: 0:00:00 - 8:29:59, T_2: 8:29:59-15:09:45 and T_3: 15:09:45- 23:59:59 and bloodglucose readings are discretized as B_1: <80 mg/dl, B_2: 80-200 mg/dl and B_3: > 200 mg/dl (Figure 4.4).

For discretizing blood glucose only ($EMBD$), I utilized the EM algorithm on collected blood glucose readings and applied experts' knowledge on timestamps. Each timestamp period is T_1 : 7:30:00-12:00:00, T_2: 12:00:00-16:30:00 and T_3: 16:30:00-21:00:00, and blood glucose readings are categorized as B_1: < 170.8 mg/dl, B_2: 170.8-274.0 mg/dl and B_3: > 274.0 mg/dl (Figure 4.5).

For quantizing both timestamps and blood glucose readings ($EMTBD$), I utilized EM algorithm on collected timestamps of blood glucose measurements to discretize data into three bins and readings collected in each time period is also discretized into three levels. In other

		T_1	T_2	T_3
Level 1	**Mean**	124.9	133.6	139.9
	STD	28.90	27.52	38.69
Level 2	**Mean**	213.2	187.1	215.5
	STD	53.35	60.65	68.41
Level 3	**Mean**	426.4	273.8	438.3
	STD	16.24	138.6	138.6

Table 4.6: Timestamp and Blood Glucose Ranges in *EMTBD*. Note: T_1 is 00:00:00-08:29:59, T_2 is 08:30:00-15:09:45 and T_3 is 15:09:46-23:59:59.

words, each time interval has different standards of categorizing blood glucose data in three different levels. Each discretized timestamp period and its three different blood glucose levels are in Table 4.6. I assume that the readings are mixture of Gaussian distribution and find intersection of Gaussian curves. The obtained B_1, B_2, B_3 ranges of T_1 (B_{11}, B_{12}, B_{13}) are < 166 , 166-372, and > 372 and B_{21}, B_{22}, B_{23}, B_{31}, B_{32}, B_{33} are < 169 , 169-265, > 265, $<$ 185, 185-318, > 318 mg/dl accordingly.

The experimental results show that the maximum sliding window size to make the conditional probabilities of ten best Apriori first order logic rules 1.00 in Algorithm 1 is 4 days in *EMTBD*, while other methods require 5 days (Figure 4.7). Therefore, the developed algorithm can reduce the required time windows to 4 days to reach the maximum conditional probability, while utilizing experts' knowledge requires 5 days. Furthermore, discretizing timestamp and quantizing blood glucose data of each time frame using the EM algorithm yields less computational power and maximizes information gain in a shorter period of time compared to using experts' knowledge-based methods.

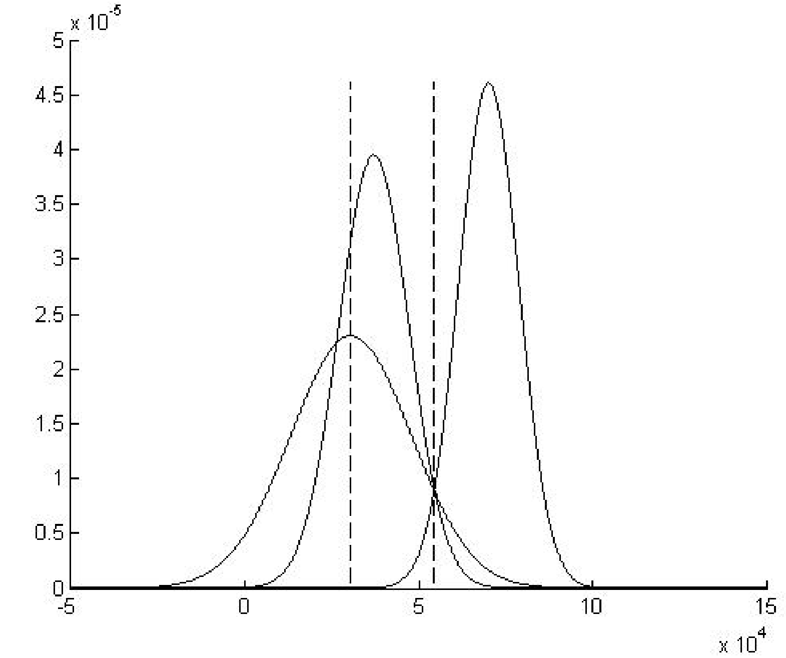


Figure 4.4: Gaussian Mixture of Timestamp (seconds) in *EMTD*

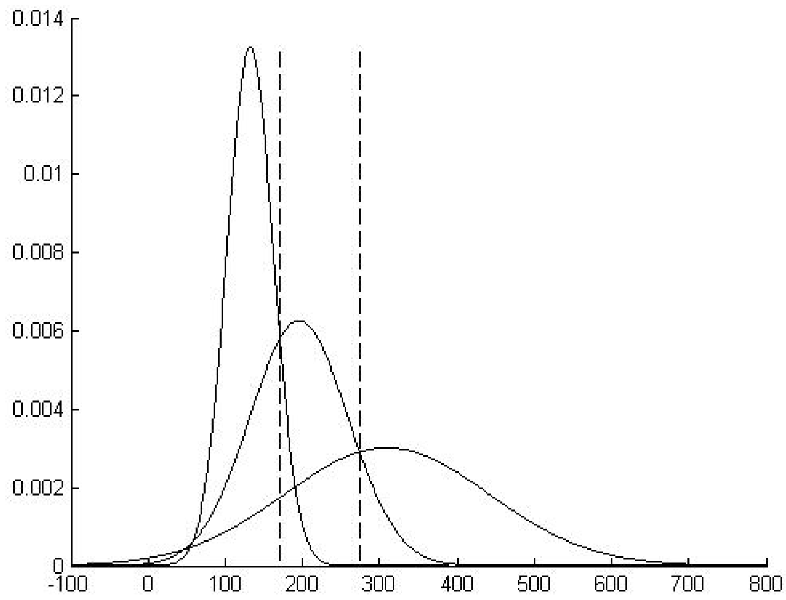
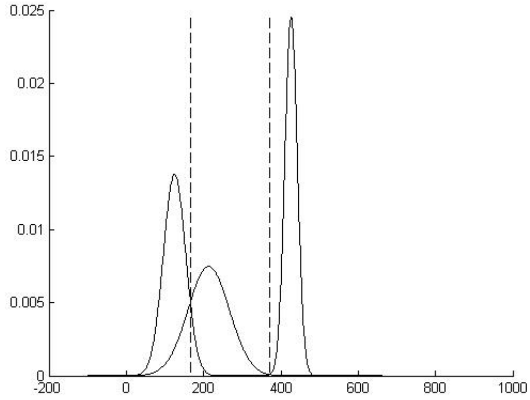
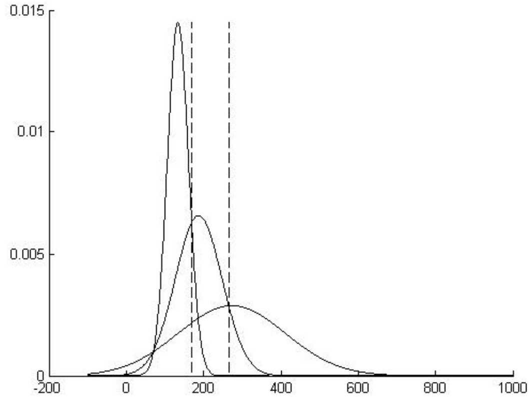


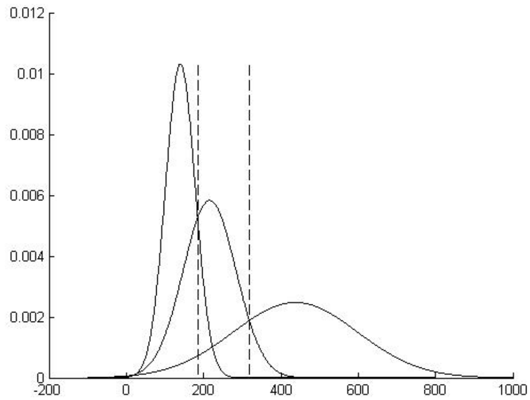
Figure 4.5: Gaussian Mixture of Blood Glucose (mg/dl) in *EMBD*



(a) Gaussian Mixture of Blood Glucose (mg/dl) in T_1



(b) Gaussian Mixture of Blood Glucose (mg/dl) in T_2



(c) Gaussian Mixture of Blood Glucose (mg/dl) in T_3

Figure 4.6: Gaussian Mixture of Blood Glucose (mg/dl) in *EMTBD*

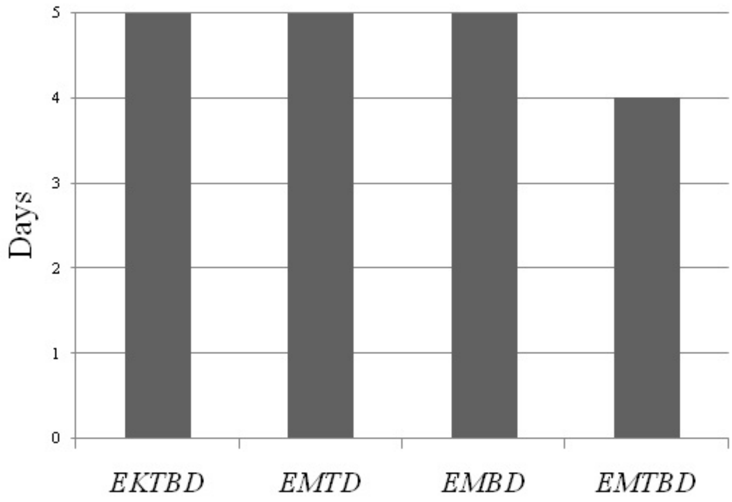


Figure 4.7: Required Window Size (Days) to Reach Maximum Conditional Probability, 1

4.4 Data Association Rule Mining

For evaluating developed methods, the data from the *Diabetes* clinical trial in Section 4.1.2 are utilized in this section. As mentioned earlier, the Apriori algorithm is used as the main association rule mining and the confirmation-based association rule mining (Tertius) algorithm is used as a supplementary algorithm.

4.4.1 Apriori-based Data Association Rule Mining

The pre-defined threshold for s_{min} and c_{min} in Algorithm 1 are both 0.95 and p_{min} is 0.85 in Algorithm 3. The proposed algorithm had optimal results with a look- back window of 5 days. The minimum confidence (c_{min}) in Algorithm 1 peaks at 1.00 at 2 to 5 days. Compared to my other study which only utilizes experts' knowledge [SMW12], the combination of EM algorithm-based discretization and Apriori algorithm shows improvement in 1.292%

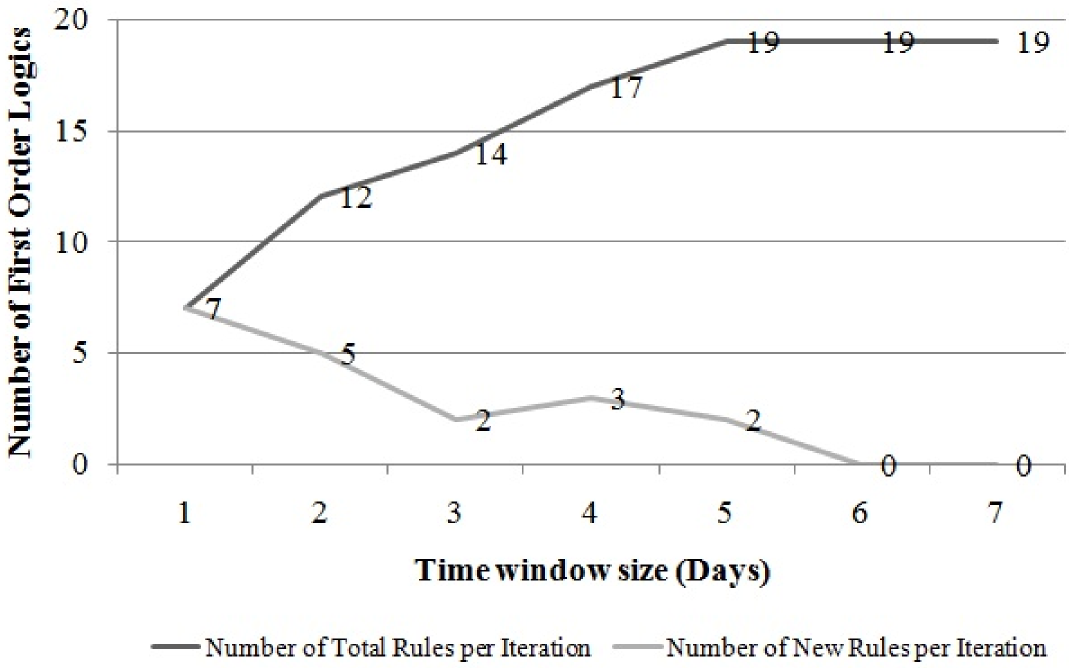


Figure 4.8: Number of First Order Logics Added or Updated per Iteration from Apriori-based Association Rule Mining

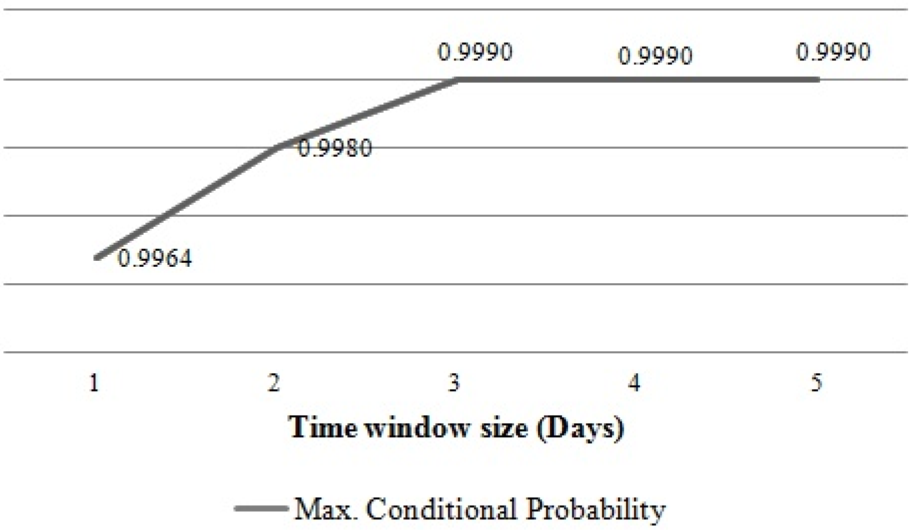


Figure 4.9: Maximum Conditional Probability per Iteration from Apriori-based Association Rule Mining

of minimum confidence.

Figure 4.8 shows the number of new rules added or updated with increasing window size. A total of 7 rules were added with a window size of one day and a total of 19 rules in Table 4.7 were updated with a window size of 5 days. No rules were added or updated with a look-back window of more than 5 days. Compared with results in [SMW12], a larger amount of data (546 data in [SMW12] and 1117 data in this study after data cleaning) and the EM algorithm-based discretization yields more rules with larger size of sliding window from the Apriori algorithm. As shown in Figure 4.9, the maximum conditional probability of generated first order logic increased as the size of sliding time window increases.

The total number of patient tasks was reduced by up to 76.19% with negligible information loss. The reduction in patient tasks allows the system to generate additional tasks for patients to increase information gain. For example, as shown in Table 4.7, it was found that responses to Q3 and Q4 in Table 3.2 can be inferred from each other. This allows the system to generate a new unrelated question to replace Q3 to learn additional information about this patient (with no added work by the patient).

Compared with the algorithm in Section 3.3.2.2 [FL01], the Apriori algorithm shows higher efficiency. However, as Flach's confirmation-based algorithm finds new rules that the developed algorithm doesn't generate, it can be used for generating more first order logics.

IF	THEN	Cond. Prob
Answer of Q3 on Day 1 is Good/Fair AND Answer of Q4 on Day 2 is Better/About Same	Answer of Q3 on Day 2 is Good/Fair	0.9970
Answer of Q3 on Day 1 is Good/Fair AND Answer of Q4 on Day 2 is Better/About Same AND Answer of Q4 on Day 3 is Good/Fair	Answer of Q3 on Day 3 is Good/Fair	0.9990
Answer of Q3 on Day 1 is Poor	Answer of Q3 on Day 2 is Poor Answer of Q3 on Day 3 is Poor Answer of Q4 on Day 1 is Worse Answer of Q4 on Day 2 is Worse Blood glucose is above 372 before 8:30 am on Day 1 Blood glucose is above 372 before 8:30 am on Day 2 Blood glucose is above 372 before 8:30 am on Day 3	0.8889

Answer of Q4 on Day 1 is Better/About Same	Answer of Q3 on Day 1 is Good/Fair	0.9961
Answer of Q4 on Day 1 is Better/About same AND Answer of Q4 on Day 2 is Better/About same	Answer of Q3 on Day 2 is Good/Fair	0.9980
Blood glucose is less than 372 before 8:30 am on Day 1 AND Answer of Q4 on Day 2 is Better/About Same AND Answer of Q4 on Day 3 is Better/About Same	Answer of Q3 on Day 3 is Good/Fair	0.9990
Blood glucose is not below 166 AND not above 372 before 8:30 am on Day 1	No Multiple measurement before 8:30 am on Day 1	0.9964
Multiple measurement before 8:30 am on Day 1	Blood glucose is above 372 before 8:30 am on Day 1 Blood glucose is below 166 before 8:30 am on Day 1 Blood glucose is above 265 between 8:29:59am and 15:09:45 pm on Day 1 Multiple measurement between 8:29:59 am and 15:09:45 pm on Day 1	0.9333

	Blood glucose is above 318 after 15:09:45 on Day 1 Multiple measurement after 15:09:45 pm on Day 1	

Table 4.7: Results of Apriori-based Data Association Rule Mining

4.4.2 Confirmation-based Data Association Rule Mining

The pre-defined threshold for $Confirmation\ Threshold$ in Algorithm 2 is 0.95 and p_{min} is 0.85 in Algorithm 3. The proposed algorithm had optimal results with a look- back window of 2 days. As the size of sliding windows increases, the frequency of counterinstances decreases (Figure 4.11).

Most of first order logics returned from Algorithm 2 satisfying with $Confirmation\ Threshold$ does not satisfy with p_{min} and only total of 3 rules in Table 4.8 are generated (Figure 4.10). No rules were added or updated with a look-back window of more than 2 days, using the confirmation-based association rule mining.

As the number of considered rules increases exponentially as the size of sliding time window increases, the algorithm is hard to be applied in real-time and it is better to use the confirmation-based rule mining as a supplementary of Apriori.

After combining results from Section 4.4.1 and 4.4.2, all generated rule satisfying the pre-defined conditions are classified as task optimization (Table 4.9) and early adaptive alarms(Table 4.10).

The experimental results of combination of Apriori and confirmation-based association rule mining show that the developed algorithm can reduce the number of tasks by up to 76.19 % with the maximum time window size of 5 days and predict 10 future events (27.08 %) for providing early adaptive alarms.

IF	**THEN**

Male AND Blood glucose is NOT below 185 and NOT between 185 and 315 after 3:09:46 pm on Day1	No measurement after 3:09:46 pm on Day 1
Blood glucose is above 372 on Day 1 and Measurement existing after 3:09:46 pm on Day 2	Blood glucose is below 185 after 3:09:46 pm on Day 2 Blood glucose is between 185 and 318 after 3:09:46 pm on Day 2
Answer of Q3 on Day 1 is Poor	Blood glucose is above 372 before 8:30 am on Day 1 Blood glucose is above 372 before 8:30 am on Day 2 Blood glucose is above 372 before 8:30 am on Day 3
Multiple measurement before 8:30 am on Day 1	Blood glucose is above 372 before 8:30 am on Day 1 Blood glucose is below 166 before 8:30 am on Day 1 Blood glucose is above 265 between 8:29:59 am and 15:09:45 pm on Day 1 Blood glucose is above 318 after 15:09:45 on Day 1
Answer of Q3 on Day 1 is Good/Fair AND Answer of Q4 on Day 2 is Better/About Same	Answer of Q3 on Day 2 is Good/Fair

Answer of Q3 on Day 1 is Good/Fair AND Answer of Q4 on Day 2 is Better/About Same AND Answer of Q4 on Day 3 is Good/Fair	Answer of Q3 on Day 3 is Good/Fair
Answer of Q3 on Day 1 is Poor	Answer of Q3 on Day 2 is Poor Answer of Q3 on Day 3 is Poor Answer of Q4 on Day 1 is Worse Answer of Q4 on Day 2 is Worse
Answer of Q4 on Day 1 is Better/About Same	Answer of Q3 on Day 1 is Good/Fair
Answer of Q4 on Day 1 is Better/About same AND Answer of Q4 on Day 2 is Better/About same	Answer of Q3 on Day 2 is Good/Fair
Blood glucose is less than 372 before 8:30 am on Day 1 AND Answer of Q4 on Day 2 is Better/About Same AND Answer of Q4 on Day 3 is Better/About Same	Answer of Q3 on Day 3 is Good/Fair

Table 4.10: First Order Logic for Early Adaptive Alarm

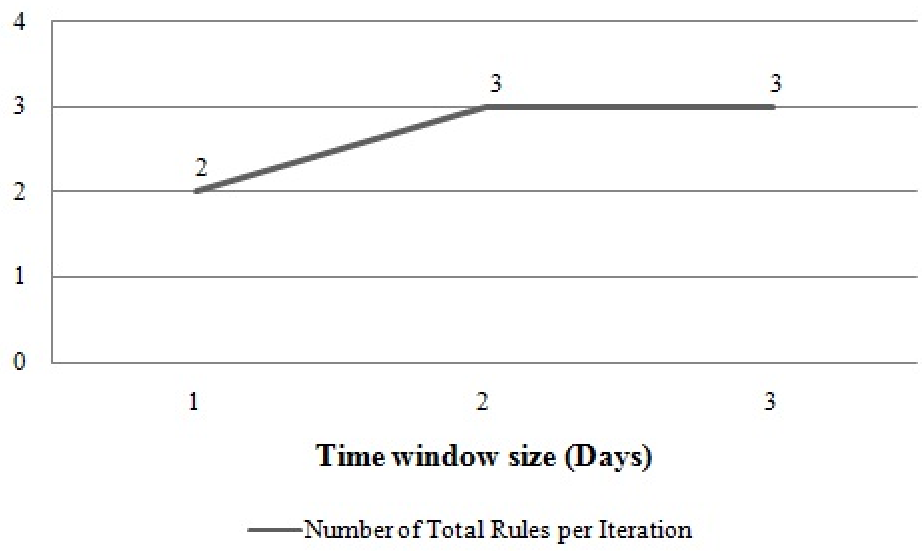


Figure 4.10: Number of First Order Logics Added or Updated per Iteration from Confirmation-based Association Rule Mining

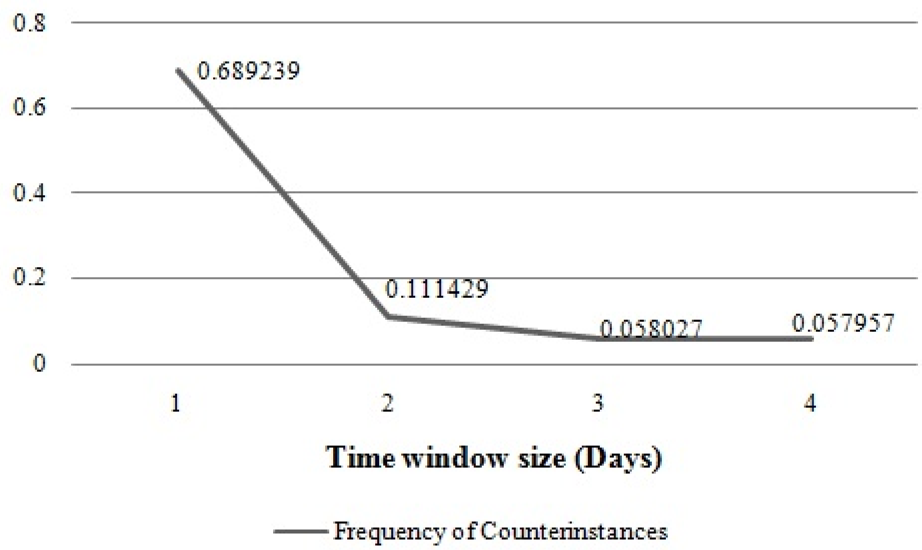


Figure 4.11: Number of Counterinstances per Iteration from Confirmation-based Association Rule Mining

IF	**THEN**	**Confirmation**	**Conditional Probability**
Male AND Blood Glucose is NOT below 185 and NOT between 185 and 315 after 3:09:46 pm on Day1	No measurement after 3:09:46 pm on Day 1	0.9981	0.9711
Blood glucose is above 372 on Day 1 and Measurement existing after 3:09:46 pm on Day 2	Blood glucose is below 185 after 3:09:46 pm on Day 2 Blood glucose is between 185 and 318 after 3:09:46 pm on Day2	0.9961	0.9694

Table 4.8: Results of Confirmation-based Data Association Rule Mining

IF	THEN
Answer of Q3 on Day 1 is Good/Fair AND Answer of Q4 on Day 2 is Better/About Same	Answer of Q3 on Day 2 is Good/Fair
Answer of Q3 on Day 1 is Good/Fair AND Answer of Q4 on Day 2 is Better/About Same AND Answer of Q4 on Day 3 is Good/Fair	Answer of Q3 on Day 3 is Good/Fair
Answer of Q3 on Day 1 is Poor	Answer of Q3 on Day 2 is Poor Answer of Q3 on Day 3 is Poor Answer of Q4 on Day 1 is Worse Answer of Q4 on Day 2 is Worse
Answer of Q4 on Day 1 is Better/About Same	Answer of Q3 on Day 1 is Good/Fair
Answer of Q4 on Day 1 is Better/About same AND Answer of Q4 on Day 2 is Better/About same	Answer of Q3 on Day 2 is Good/Fair
Blood glucose is less than 372 before 8:30 am on Day 1 AND Answer of Q4 on Day 2 is Better/About Same AND Answer of Q4 on Day 3 is Better/About Same	Answer of Q3 on Day 3 is Good/Fair

Table 4.9: First Order Logic for Task Optimization

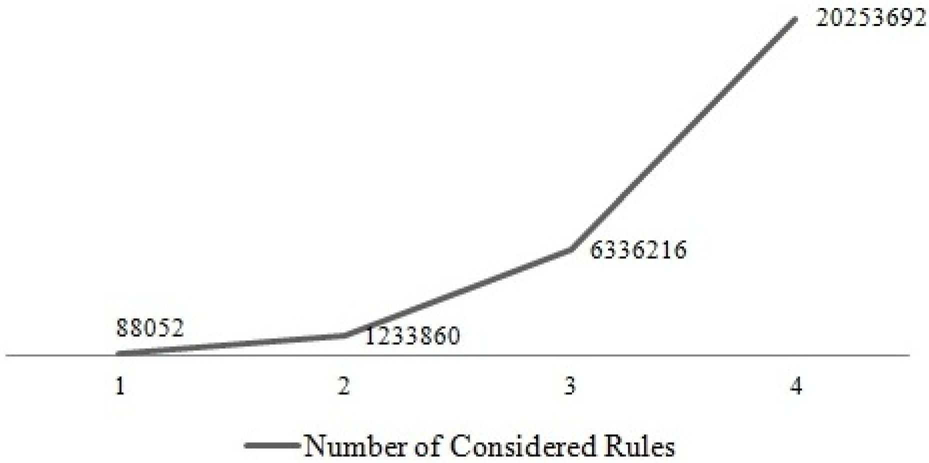


Figure 4.12: Number of Considered Hypothesis per Iteration from Confirmation-based Association Rule Mining

4.5 Missing Data Imputation

4.5.1 Non-binomial Missing Data Imputation

CHF1 employs the Heart Failure Somatic Awareness Scale (HFSAS in Table 1.1, [JFR06]), which is a 12-item Likert-scale to measure awareness of signs and symptoms specific to CHF. A 4-point Likert-type scale is used to ascertain how much a patient is bothered by a symptom (0: not at all, 1: a little, 2: a great deal, 3: extremely). Patients in *CHF2* were asked subset of 8 questions in Table 4.1 (0: no, 1: yes).

For non-binomial data, PACE, linear [DS98], simple linear [HN09] and isotonic regression [BC90] methods were applied. Table 4.11 shows the correlation coefficient values of each method to predict each answer a_i corresponding to Q_i of *CHF1*. Correlation coefficient is a measure of the least square fitting values between the predicted and original data. For a given N data points (X, Y), the correlation coefficient $\rho_{X,Y}$ is given as Equation 4.2 where $COV(X, Y)$ is a covariance between X and Y and σ_X,σ_Y are standard deviation values of

X and Y. The experimental results show that PACE regression method works better on average than other given regression methods.

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \times \sigma_Y} \tag{4.2}$$

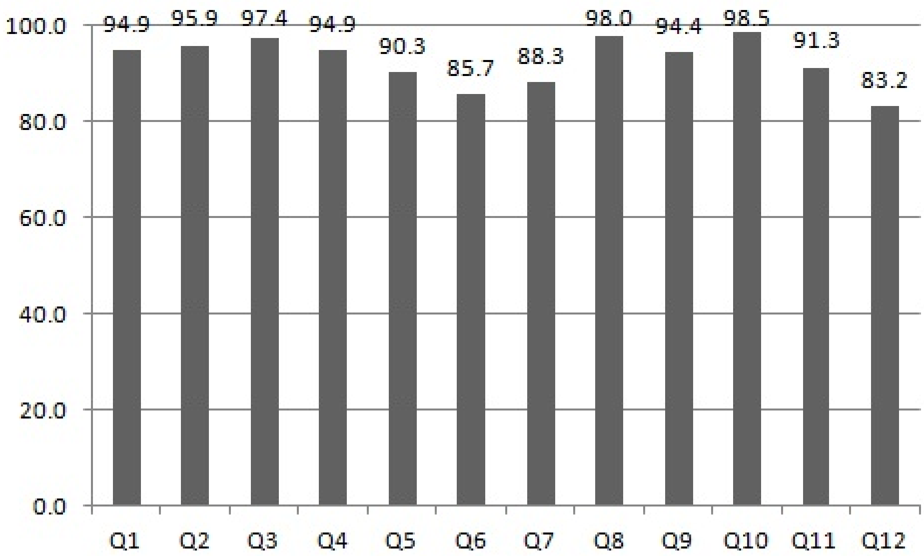
After calculating the coefficient and constant variables, the developed algorithm determines missing values using PACE regression (rounding any non-integer value returned by PACE). The accuracies of the obtained values range between 85.7% and 98.5% for *CHF1* and between 91.4% and 99.0% for *CHF2* (Table 4.13).

4.5.2 Binomial Missing Data Imputation

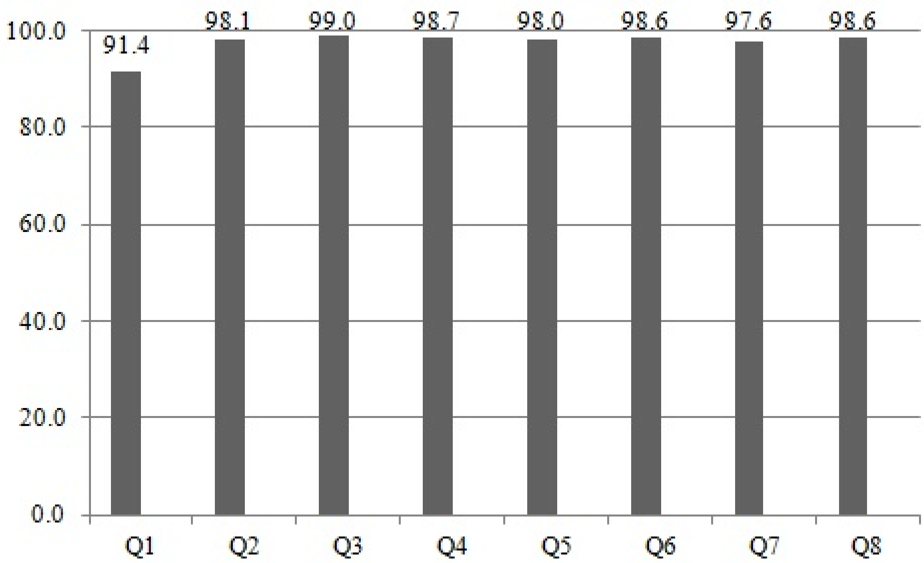
The binomial case predicts a potential abnormal vital sign when no data value exists within WANDAs database using other existing attributes. C4.5 [Qui93], Random Tree [Ald91], Naïve Bayes [JL95], Voting Feature Interval (VFI) [DG97], Nearest Neighbor [Mar95], PART [FW98], Decision Table and Naïve Bayes (DTNB) [HF08] and Decision Table [Koh95] algorithms were applied and their recall values were compared (Table 4.13, Table 4.14). For each method, ten fold cross validation was applied. In ten fold validation, the original sample is randomly partitioned into ten subsets and a single subset is held as a testing model, with the remaining nine subsets are used as training data. This cross-validation process is then repeated ten times, using a new subset as a testing model for each repetition. Recall values are given as Equation 4.3.

$$recall = \frac{T_p}{T_p + F_n} \tag{4.3}$$

where T_p is true positive and F_n is false negative. The experimental result (Table 4.14) shows that naïve Bayes and VFI have recall values of up to to 0.900 for weight, 0.946 for systolic blood pressure, 0.833 for diastolic blood pressure and 0.922 for heart rate values.



(a) Questions in Table 1.1 for *CHF1*



(b) Questions in Table 4.1 for *CHF2*

Figure 4.13: Accuracies of the Predicted Data for Non-binomial Cases (%)

	LR	SLR	**PR**	IR
a_1	0.61	0.50	**0.62**	0.53
a_2	0.79	0.75	**0.79**	0.76
a_3	0.41	0.34	**0.42**	0.23
a_4	0.28	0.00	**0.30**	0.12
a_5	0.30	0.20	**0.32**	0.18
a_6	0.78	0.85	**0.82**	0.86
a_7	0.44	0.32	**0.52**	0.38
a_8	0.88	0.90	**0.88**	0.89
a_9	0.29	0.24	**0.29**	0.10
a_{10}	0.90	0.92	**0.90**	0.92
a_{11}	0.42	0.40	**0.42**	0.29
a_{12}	0.84	0.85	**0.85**	0.85

Table 4.11: Correlation Coefficient Values of Each Technique for Non-binomial Data in *CHF1*. Note: LR(Linear Regression), SLR(Simple Linear Regression), PR(Pace Regression), IR(Isotonic Regression)

Algorithm	Description
Linear Regression	The linear regression constructs a linear model of one or more independent variables, X and a dependent variable, Y, using Akaike information criterion. [Kem03][Aka74] $Y = X\beta + \epsilon$, where $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $X = \begin{pmatrix} {x_1}^T \\ {x_2}^T \\ \vdots \\ {x_n}^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$, $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$.
Simple Linear Regression	Least squares estimator of a linear regression model with a single explanatory variable, X. [KK62] For $Y = X\beta + \epsilon$, find $min_{\epsilon,\beta}\ Q(\epsilon,\beta)$ where $Q(\epsilon,\beta) = \sum_{i=1}^{n}(y_i - \epsilon - \beta x_i)^2$.
Isotonic Regression	Weighted least square fit of $x \in \mathbb{R}^n$ to a vector $a \in \mathbb{R}^n$ with weighted vector $w \in \mathbb{R}^n$. [BC90] Find $min \sum_{i=1}^{n} w_i \lvert x_i - a_i \rvert^p$.

Table 4.12: Compared Algorithms in Non-binomial Missing Data imputation

Algorithm	Description
C4.5	C4.5 constructs a decision tree based on information entropy. Each feature in the training set is evaluated for its information gain. $IG(T,a) = H(T) - H(T\|a)$ $H(T) = -\sum_{i=1}^{n} p(x_i) \ln p(x_i)$ Feature with the largest IG and have yet been used is chosen to split the decision tree at each node until the confidence falls below certain threshold.
Nearest Neighbor	An unlabeled instance is classified based on its nearest k neighbors from the training set based on majority vote. Euclidean distance of the features is often used to determine the closeness of two instances. $d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ where x_i and y_i are corresponding features from the two instances.
Random Tree	The algorithm constructs a tree which considers K-randomly-chosen attributes at each node without pruning.
PART	PART is a hybrid method of C4.5 tree and RIPPER rule mining [Wil95]. The algorithm refine an initial rule set by discarding (C4.5) or adjusting (RIPPER) each rule.
DTNB	The algorithm uses forward selection for choosing attributes modeled by Decision Table (DT) or Naïve Bayes (NB). $Q(y\|X) = \alpha \times Q_{DT}(y\|X^{DT}) \times Q_{NB}(y\|X^{NB})/Q(y)$

	where X^{DT} is a set of attributes in the DT, X^{NB} is a set of attributes in the NB, $Q_{DT}(y\|X^{DT})$ and $Q_{NB}(y\|X^{NB})$ are class probability estimates obtained from DT and NB, α is a normalization constant and $Q(y)$ is the prior probability of the class. Leave-one-out cross-validated area-under-curve (AUC) is used to split based on the probability estimates generated by the combined model.
Decision Table	It builds a simple majority classifier using a schema and a body, while a schema is a set of features in the table and a body consists of labelled features defined by the schema. Given an unlabelled instance, a decision table searches for exact matches in the decision table using the features in the schema. Let $A = X_1, ..., X_n$ be a set of features and S be a sample over features in A. DTM is a decision table majority function and f is a given target function. The objective of the algorithm is finding $A^* = argmin$ $err(DTM(A', S), f)$

Table 4.13: Compared Algorithms in Binomial Missing Data imputation

Classifiers were trained in two ways. First, unique classifiers were created for each individual where only data collected from an individual was used to predict values from the same individual. Second, a grouped classifier was created using data from the entire population. Both the individual and grouped classifiers were compared using ten-fold validation to test data from 16 patients in *CHF1*. The recall values of weight, blood pressure, and heart rate are improved when training on the entire groups data as compared with training each individuals data separately (Table 4.15). For questionnaire data, the accuracies of results were

	Weight	**Systolic**	**Diastolic**	**HR**
C 4.5	0.308	0.857	0.803	0.592
Random Tree	0.254	0.875	0.788	0.592
Naïve Bayes	**0.854**	**0.857**	**0.773**	**0.922**
VFI	**0.900**	**0.946**	**0.833**	**0.786**
Nearest Neighbor	0.238	0.875	0.773	0.592
PART	0.354	0.875	0.758	0.689
DTNB	0.223	0.911	0.803	0.612
Decision Table	0.200	0.857	0.788	0.466

Table 4.14: Recall Values of Weight, Systolic, Diastolic Blood Pressure and Heart Rate Values

also better when training on all patients data. When training individually, 75% of patients data showed 0% accuracy. This is because the entire group has bigger number of data and many individual share similarities in monitored attributes, such as age, symptoms of CHF, etc.

	Weight			Systolic			Diastolic			HR		
	NB	BN	VFI	NB	BN	VFI	NB	BN	VFI	NB	BN	VFI
P1	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
P2	.00	.00	.00	.33	.00	.33	.85	.39	.62	.54	.31	.69
P3	.00	.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
P4	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
P5	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
P6	.33	.33	.33	.00	.00	.00	.00	.00	.00	.00	.00	.00
P7	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
P8	.00	.00	.33	.00	.00	.00	.00	.00	.00	.00	.00	.00
P9	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
P10	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
P11	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
P12	1.00	.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
P13	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
P14	.65	.88	.82	.00	.00	.00	.00	.00	.00	.00	.00	.00
P15	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
P16	.00	.00	.00	.00	.00	.00	.00	.00	.00	.92	.75	.67
Group	**.70**	**.65**	**.65**	**.71**	**.71**	**.71**	**.89**	**.78**	**.67**	**.91**	**.78**	**.78**

Table 4.15: Recall Values of Data for Individual and Group, NB :Naïve Bayes, BN: Bayes Net, VFI: Voting Feature Interval

CHAPTER 5

Conclusion

5.1 Summary

This dissertation presented WANDA, a remote health monitoring system and its data analytics functions including task optimization, missing data imputation and early adaptive alarm. This dissertation made the following contributions:

1. The development of two versions of remote health monitoring system (Section 3.1, Section 3.2);

2. The development of the task optimization and early adaptive alarm function (Section 3.3.2);

3. The development of the missing data imputation function (Section 3.3.3); and

4. The deployment of the system for clinical trials (Section 4.1).

The WANDA system was developed in conjunction with the University of California Los Angeles, Computer Science, the Ronald Regan Medical Center and the School of Nursing. WANDA is developed to monitor patients with chronic diseases and provide real-time feedbacks and built on a three-tier architecture. The first tier consists of sensors that measure patients' health related measurements and transmits data to the second tier. WANDA utilizes a Bluetooth-enabled off-the-shelf devices including weight scale, blood pressure monitor, blood glucose monitor, etc. to collect health related measurements and transmit data. The second tier consists of web servers that receive data from the first tier and maintains data integrity and provide monitoring applications. In addition, when the obtained values are out of the acceptable range, the second tier sends alert messages to healthcare providers via text message or e-mail. The third tier is a back-end database server that performs backup and recovery jobs by applying an offline backup and performs data analytics. The data anlytic functions on the third tier includes task optimization, missing data imputation, and early early adaptive alarms.

The clinical trials including *CHF1*,*CHF2* and *Diabetes* in Section 4.1 have been conducted. The outcomes from the clinical trials for CHF and diabetes showed reduced number of abnormal health-related readings, more supports from healthcare professionals and higher system adherence rates.

The developed system applies EM-based data discretization for quantizing continuous values such as timestamp and sensory data. I assumed that sensor readings from patients are Gaussian mixture and quantize continuous features and applied Apriori and confirmation-based association rule mining algorithm which efficiently finds related data with support and confirmation thresholds. The designed algorithm minimizes the number of action items and reorganizes series of tasks for maximizing information gain. Also, the developed methods predicts future events such as readings out of an acceptable range and system non-use to provide early alarms for preventing emergencies.

In this dissertation, I applied the data discretization and association rule mining algorithms in Section 3.3.2 to 1117 data sets from 21 patients with diabetes enrolled in the intervention arm of the *Diabetes* clinical trial. Patients are required to measure their blood sugar up to three times a day and answer four questionnaires daily. The experimental results show that the developed algorithm can reduce the number of tasks by up to 76.19% with minimum support 0.95, minimum confidence 0.95 and maximum time window size of 5 days. The developed algorithm can predict up to 27.08% of future events for generating early adaptive alarms. Compared to our earlier study [SMW12], the EM-based discretization helps improve confidence levels of first order logics and predict further events. As the Apriori algorithm, which is the main association rule mining in this dissertation has excellent scale-up properties [AS94], the developed algorithm can be applied to the remote patient with low complexity. Confirmation-based association rule mining is only used as a supplementary algorithm because of its low efficiency compared to Apriori.

The developed missing data imputation algorithm enhanced the accuracy of the missing data using the PACE regression method for predicting and imputing non-binomial data; and Naïve Bayes method and voting feature interval for binomial data in data from *CHF1* and *CHF2*. The experimental results show that PACE regression works better than linear regression, simple linear regression, and isotonic regression methods with accuracy values of more than 85.7%. The experiment comparing Nave Bayes and VFI methods with other algorithms proves that Bayes and VFI algorithms work better with recall values of up to 0.90 for weight, 0.95 for systolic blood pressure, 0.83 for diastolic blood pressure and 0.92 for heart rate values.

5.2 Future Work in Association Rule Mining

For the clinical trials, paper-based surveys have been conducted. However, the information for designing and predicting tasks and future events have not been combined with patients' profile, sensory and questionnaire data. As Klonoff's study appplied [Klo09], it would be helpful to combine information about study participants' perception and experience of technologies for enhancing the conditional probabilities of task optimization and early adaptive alarms. Table 5.1 and Table 5.2 are parts of the survey from the *CHF1* and *Diabetes* clinical trials respectively. In addition, automating the survey process using a tablet (Figure 5.1) or smartphone should reduce the time and efficiency of processing the paper-based information [SLA11]. As computer-based questionnaires are more standardized, answers can be highly completed and the process improves efficiency and reduces costs by applying computerized surveys [MWL94].

In this dissertation, we collected healthcare providers' comments and call logs. As most of the comments included medical terms and their abbreviations, standardizing and processing terms by using UMLS (Unified Medical Language System) would be useful for reducing gaps between lay users and experts [Aro01][Med12].

Future studies will investigate and validate the significance of the obtained first order logic rules in this dissertation. Especially, I will verify the effectiveness and significance of the EM-discretization in remote health monitoring. To make the first-order logic richer to reduce required patient tasks dynamically, more data association rule mining techniques will be exploited and maximize the conditional probability and confirmation. Future works will also apply Bayes algorithms to enhance the developed task optimization and early adaptive alarm methods and evaluate precision and recall.

5.3 Future Work in Missing Data Imputation

As the developed missing data algorithm did not apply the sliding window and discretization methods, I will apply the methods to enhance the effectiveness and predict missing values using trends of data and their relationships. Evaluating missing data rate after applying task optimization and the number of adverse events with early adaptive alarms would be another task for future works.

Questions
I am satisfied with the remote monitoring system.
It is simple to use.
It is wonderful.
It is useful.
I don't notice any inconsistencies as I use it.
It is fun to use.
It does everything I would expect it to do.
It requires the fewest steps possible to accomplish what I want to do with it.
It is user friendly.
It would give me more control over activities in my life.
I feel I need to have it.
I can use it successfully every time.
It is flexible.
I learned to use it quickly.
It would make the things I want to accomplish easier to get done.
I can recover from mistakes quickly and easily.
It would help me be more effective.
Using it is effortless.
I can use it without written instructions.
It meets my needs.
It works the way I want it to work.
I easily remember how to use it.
It is pleasant to use.
I would recommend it to a friend.
It is easy to learn to use it.

Table 5.1: Perceptions of Usability Questionnaire in *CHF1*

Questions	**Answer**
I keep other health information (i.e. blood pressure, weight) on	Paper, Computer, I do not keep information
I electronically send blood sugar values to	A professional, Another person, No one
I electronically send other health information to	A professional, Another person, No one
Logging my diabetic information makes me cope better	A professional, Another person,No one
I electronically ask questions to	Agree, Disagree, Not sure
Logging blood sugar values makes me feel healthier	Agree, Disagree, Not sure
Logging other health information makes me feel healthier	Agree, Disagree, Not sure
I will use telemonitoring from the home to send blood sugar values to my diabetes professional	Agree, Disagree, Not sure
I will use telemonitoring from the home to send other health information to my diabetes professional	Agree, Disagree, Not sure
I feel using telemonitoring could help improve my diabetes	Agree, Disagree, Not sure
Currently, my level of comfort with computer technology is	Excellent,Good, Fair, Poor
Highest educational degree	No Degree, High School, College, Beyond College

Table 5.2: Survey Questions in *Diabetes* at the Baseline

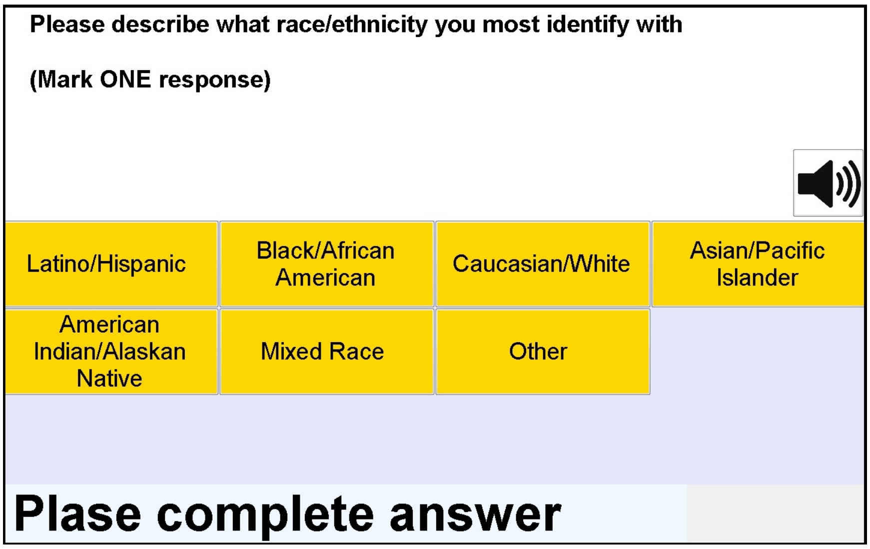


Figure 5.1: WHI Questionnaire System

References

[A12] Inc. A&D Engineering· "A&D. http://www.andonline.com/." 2012.

[Aka74] H. Akaike. "A new look at the statistical model identification." *Automatic Control, IEEE Transactions on*, **19**(6):716–723, 1974.

[Ald91] D. Aldous. "The continuum random tree II: an overview." *Stochastic Analysis*, **167**:23–70, 1991.

[Ale12] Alere. "Alere. http://www.alere.com/." 2012.

[AM85] E.K. Antonsson and R.W. Mann. "The frequency content of gait." *Journal of biomechanics*, **18**(1):39–47, 1985.

[Ame11] American Diabetes Association. "Diabetes Statistics. http://www.diabetes.org/diabetes-basics/diabetes-statistics/." 2011.

[Aro01] A.R. Aronson. "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." In *Proceedings of the AMIA Symposium*, p. 17. American Medical Informatics Association, 2001.

[AS94] R. Agrawal, R. Srikant, et al. "Fast algorithms for mining association rules." In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, volume 1215, pp. 487–499, 1994.

[Ass08] American Heart Association et al. "Heart Disease and Stroke Statistics, 2008 Update At-a-Glance." *Dallas, Texas: American Heart Association*, 2008.

[BC90] M.J. Best and N. Chakravarti. "Active set algorithms for isotonic regression; a unifying framework." *Mathematical Programming*, **47**(1):425–439, 1990.

[Bla08] K.D. Blanchet. "Telehealth and diabetes monitoring." *Telemedicine and e-Health*, **14**(8):744–746, 2008.

[BMB02] S. Bhattacharya, C. Mohan, K.W. Brannon, I. Narang, H.I. Hsiao, and M. Subramanian. "Coordinating backup/recovery and data consistency between database and file systems." In *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, pp. 500–511. ACM, 2002.

[BS01] G.B. Bell and A. Sethi. "Matching records in a national medical patient index." *Communications of the ACM*, **44**(9):83–88, 2001.

[Bur11] U.S. Census Bureau. "Census Bureau. http://www.census.gov/." 2011.

[CBM07] S.I. Chaudhry, B. Barton, J. Mattera, J. Spertus, and H.M. Krumholz. "Randomized trial of telemonitoring to improve heart failure outcomes (Tele-HF): study design." *Journal of cardiac failure*, **13**(9):709–714, 2007.

[CG94] Diabetes Control, Complications Trial Research Group, et al. "Effect of intensive diabetes treatment on the development and progression of long-term complications in adolescents with insulin-dependent diabetes mellitus: Diabetes Control and Complications Trial." *J Pediatr*, **125**(2):177–188, 1994.

[CHF96] San Francisco. University of California, Institute for Health & Aging, and Robert Wood Johnson Foundation. *Chronic care in America: a 21st century challenge.* The Robert Wood Johnson Foundation, 1996.

[CKR03] P.S. Ciechanowski, W.J. Katon, J.E. Russo, and I.B. Hirsch. "The relationship of depressive symptoms to symptom reporting, self-care and glucose control in diabetes." *General Hospital Psychiatry*, **25**(4):246–252, 2003.

[CKV06] J. Carpenter, M. Kenward, and S. Vansteelandt. "A comparison of multiple imputation and inverse probability weighting for analyses with missing data." *Journal of the Royal Statistical Society, Series A*, **169**(3):571–84, 2006.

[CMC10] S.I. Chaudhry, J.A. Mattera, J.P. Curtis, J.A. Spertus, J. Herrin, Z. Lin, C.O. Phillips, B.V. Hodshon, L.S. Cooper, and H.M. Krumholz. "Telemonitoring in patients with heart failure." *New England Journal of Medicine*, **363**(24):2301–2309, 2010.

[DG97] G. Demiröz and H. Güvenir. "Classification by voting feature intervals." *Machine Learning: ECML-97*, pp. 85–92, 1997.

[DH96] R.O. Duda and P.E. Hart. *Pattern classification and scene analysis.* Wiley, 1996.

[DMM09] A.D.V. Dabbs, B.A. Myers, K.R. Mc Curry, J. Dunbar-Jacob, R.P. Hawkins, A. Begey, and M.A. Dew. "User-centered design and interactive health technologies for patients." *Computers, informatics, nursing: CIN*, **27**(3):175, 2009.

[DMZ08] T. Dall, SE Mann, Y. Zhang, J. Martin, Y. Chen, P. Hogan, et al. "Economic costs of diabetes in the US in 2007." *Diabetes Care*, **31**(3):596–615, 2008.

[DP12] CDC (Centers for Disease Control and Prevention). "Centers for Disease Control and Prevention. http://www.cdc.gov/." 2012.

[DS98] N.R. Draper and H. Smith. "Applied regression analysis (wiley series in probability and statistics)." 1998.

[DS10] A.S. Desai and L.W. Stevenson. "Connecting the circle from home to heart-failure disease management." *New England Journal of Medicine*, **363**(24):2364–2367, 2010.

[FK85] L. Fahrmeir and H. Kaufmann. "Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models." *The Annals of Statistics*, **13**(1):342–368, 1985.

[FL01] P.A. Flach and N. Lachiche. "Confirmation-guided discovery of first-order rules with Tertius." *Machine Learning*, **42**(1):61–95, 2001.

[FW98] E. Frank and I.H. Witten. "Generating accurate rule sets without global optimization." 1998.

[Geo70] E.P. George. *Time series analysis: Forecasting and control.* Holden-D., 1970.

[GHS11] R. Ghosh, J. Heit, and S. Srinivasan. "Telehealth at scale: the need for interoperability and analytics." In *Proceedings of the first international workshop on Managing interoperability and complexity in health systems*, pp. 63–66. ACM, 2011.

[Goo11] Google. "Google Maps API Family. http://code.google.com/apis/maps/." 2011.

[GS11] R. Ghosh and H. Schellhorn. "Development of a telemedical platform: Challenges, requirements, and solutions." In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011 5th International Conference on*, pp. 270–273. IEEE, 2011.

[Har58] HO Hartley. "Maximum likelihood estimation from incomplete data." *Biometrics*, **14**(2):174–194, 1958.

[HF08] M. Hall and E. Frank. "Combining naive Bayes and decision tables." In *Proc 21st Florida Artificial Intelligence Research Society Conference, Miami, Florida. AAAI Press*, 2008.

[HFJ96] M. Hanefeld, S. Fischer, U. Julius, J. Schulze, U. Schwanebeck, H. Schmechel, HJ Ziegelasch, and J. Lindner. "Risk factors for myocardial infarction and death in newly detected NIDDM: the Diabetes Intervention Study, 11-year follow-up." *Diabetologia*, **39**(12):1577–1583, 1996.

[HFK95] J.D. Harnett, R.N. Foley, G.M. Kent, P.E. Barre, D. Murray, P.S. Parfrey, et al. "Congestive heart failure in dialysis patients: prevalence, incidence, prognosis and risk factors." *Kidney international*, **47**:884–884, 1995.

[HFN00] P.C.W. van den Hoogen, E.J.M. Feskens, N.J.D. Nagelkerke, A. Menotti, A. Nissinen, and D. Kromhout. "The relation between blood pressure and mortality due to coronary heart disease among men in different parts of the world." *New England Journal of Medicine*, **342**(1):1–8, 2000.

[HFW98] R. Hambrecht, E. Fiehn, C. Weigl, S. Gielen, C. Hamann, R. Kaiser, J. Yu, V. Adams, J. Niebauer, and G. Schuler. "Regular physical exercise corrects endothelial dysfunction and improves exercise capacity in patients with chronic heart failure." *Circulation*, **98**(24):2709–2715, 1998.

[HHM03] D. Helitzer, D. Heath, K. Maltrud, E. Sullivan, and D. Alverson. "Assessing or predicting adoption of telehealth using the diffusion of innovations theory: a practical example from a rural program in New Mexico." *Telemedicine Journal and E-health*, **9**(2):179–187, 2003.

[HLF03] A.W. Haider, M.G. Larson, S.S. Franklin, and D. Levy. "Systolic blood pressure, diastolic blood pressure, and pulse pressure as predictors of risk for congestive heart failure in the Framingham Heart Study." *Annals of Internal Medicine*, **138**(1):10–16, 2003.

[HMY06] T. Hara, Y. Matsumura, M. Yamamoto, T. Kitado, H. Nakao, H. Nakao, T. Suzuki, T. Yoshikawa, and S. Fujimoto. "The relationship between body weight reduction and intensity of daily physical activities assessed with 3-dimension accelerometer." *Jpn J Phys Fit Sports Med*, **55**:385–91, 2006.

[HN09] R.M. Heiberger and E. Neuwirth. "Simple Linear Regression." *R Through Excel*, pp. 193–212, 2009.

[HOB01] J. He, L.G. Ogden, L.A. Bazzano, S. Vupputuri, C. Loria, and P.K. Whelton. "Risk factors for congestive heart failure in US men and women: NHANES I epidemiologic follow-up study." *Archives of internal medicine*, **161**(7):996, 2001.

[Hua10] J.C. Huang. "Remote health monitoring adoption model based on artificial neural networks." *Expert Systems with Applications*, **37**(1):307–314, 2010.

[HWE06] R.F. Hamman, R.R. Wing, S.L. Edelstein, J.M. Lachin, G.A. Bray, L. Delahanty, M. Hoskin, A.M. Kriska, E.J. Mayer-Davis, X. Pi-Sunyer, et al. "Effect of weight loss with lifestyle intervention on risk of diabetes." *Diabetes care*, **29**(9):2102–2107, 2006.

[Ide11] Ideal Life. "Ideal Life. http://www.ideallifeonline.com/." 2011.

[Inc12] Roving Networks Inc. "Roving Networks. http://www.rovingnetworks.com/." 2012.

[JBK72] RJ Jarrett, IA Baker, H. Keen, and NW Oakley. "Diurnal variation in oral glucose tolerance: blood sugar and plasma insulin levels morning, afternoon, and evening." *British Medical Journal*, **1**(5794):199–201, 1972.

[JC88] N.L. Jones and E.J.M. Campbell. "Clinical exercise testing." 1988.

[JCJ07] V.R. Jakkula, D.J. Cook, and G. Jain. "Prediction models for a smart home based health care system." In *Advanced Information Networking and Applications Workshops, 2007, AINAW'07. 21st International Conference on*, volume 2, pp. 761–765. IEEE, 2007.

[JFR06] C.Y. Jurgens, J.A. Fain, and B. Riegel. "Psychometric testing of the heart failure somatic awareness scale." *Journal of cardiovascular nursing*, **21**(2):95, 2006.

[JL95] G.H. John and P. Langley. "Estimating continuous distributions in Bayesian classifiers." In *Proceedings of the eleventh conference on uncertainty in artificial intelligence*, pp. 338–345. Morgan Kaufmann Publishers Inc., 1995.

[JOH09] Z. Jin, J. Oresko, S. Huang, and A.C. Cheng. "HeartToGo: A personalized medicine technology for cardiovascular disease prevention and detection." In *Life Science Systems and Applications Workshop, 2009. LiSSA 2009. IEEE/NIH*, pp. 80–83. IEEE, 2009.

[JSC09] Z. Jin, Y. Sun, and A.C. Cheng. "Predicting cardiovascular disease from real-time electrocardiographic monitoring: An adaptive machine learning approach on a cell phone." In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pp. 6889–6892. IEEE, 2009.

[KEL02] S. Kenchaiah, J.C. Evans, D. Levy, P.W.F. Wilson, E.J. Benjamin, M.G. Larson, W.B. Kannel, and R.S. Vasan. "Obesity and the risk of heart failure." *New England Journal of Medicine*, **347**(5):305–313, 2002.

[Kem03] F. Kemp. "Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences." *Journal of the Royal Statistical Society: Series D (The Statistician)*, **52**(4):691–691, 2003.

[KHX08] H.C. Kung, D.L. Hoyert, J. Xu, S.L. Murphy, et al. "Deaths: final data for 2005." *Natl Vital Stat Rep*, **56**(10):1–120, 2008.

[KK62] JF Kenney and ES Keeping. "Linear regression and correlation." *Mathematics of statistics*, **1**:252–285, 1962.

[Klo09] D.C. Klonoff. "Using telemedicine to improve outcomes in diabetesan emerging technology." *Journal of diabetes science and technology (Online)*, **3**(4):624, 2009.

[KNM06] D.M. Karantonis, M.R. Narayanan, M. Mathie, N.H. Lovell, and B.G. Celler. "Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring." *Information Technology in Biomedicine, IEEE Transactions on*, **10**(1):156–167, 2006.

[Koh95] R. Kohavi. "The power of decision tables." *Machine Learning: ECML-95*, pp. 174–189, 1995.

[LAC09] D. Lloyd-Jones, R. Adams, M. Carnethon, G. De Simone, T.B. Ferguson, K. Flegal, E. Ford, K. Furie, A. Go, K. Greenlund, et al. "Heart Disease and Stroke Statistics2009 Update A Report From the American Heart Association Statistics Committee and Stroke Statistics Subcommittee." *Circulation*, **119**(3):480–486, 2009.

[LG94] H.P. Lu and D.H. Gustafson. "An empirical study of perceived usefulness and perceived ease of use on computerized support system use over time." *International Journal of Information Management*, **14**(5):317–329, 1994.

[Lis87] B. Liskov. "Keynote address-data abstraction and hierarchy." In *ACM Sigplan Notices*, volume 23, pp. 17–34. ACM, 1987.

[LSA12] M. Lan, L. Samy, N. Alshurafa, M. Suh, H. Ghasemzadeh, A. Macabasco-O'Connell, and M. Sarrafzadeh. "WANDA: An End-to-End Remote Health Monitoring and Analytics System for Heart Failure Patients." 2012.

[Mag01] M. Maguire. "Methods to support human-centred design." *International journal of human-computer studies*, **55**(4):587–634, 2001.

[Mar95] B. Martin. *Instance-based learning: nearest neighbour with generalisation.* PhD thesis, University of Waikato, 1995.

[MDD69] C. Malherbe, M. De Gasparo, R. De Hertogh, and JJ Hoem. "Circadian variations of blood sugar and plasma insulin levels in man." *Diabetologia*, **5**(6):397–404, 1969.

[Med12] National Institutes of Health U.S. National Library of Medicine. "Unified Medical Language System. http://www.nlm.nih.gov/research/umls/." 2012.

[MGW94] J.D. McKeen, T. Guimaraes, and J.C. Wetherbe. "The relationship between user participation and user satisfaction: an investigation of four contingency factors." *Mis Quarterly*, pp. 427–451, 1994.

[MH03] T. Menzies and Y. Hu. "Data mining for very busy people." *Computer*, **36**(11):22–29, 2003.

[MKI09] S.G. Mougiakakou, I. Kouris, D. Iliopoulou, A. Vazeou, and D. Koutsouris. "Mobile technology to empower people with diabetes mellitus: design and development of a mobile application." In *Information Technology and Applications in Biomedicine, 2009. ITAB 2009. 9th International Conference on*, pp. 1–4. Ieee, 2009.

[MN93] C. Mohan and I. Narang. *An efficient and flexible method for archiving a data base*, volume 22. ACM, 1993.

[MOJ06] A. Milenković, C. Otto, and E. Jovanov. "Wireless sensor networks for personal health monitoring: Issues and an implementation." *Computer communications*, **29**(13):2521–2533, 2006.

[Moo96] T.K. Moon. "The expectation-maximization algorithm." *Signal Processing Magazine, IEEE*, **13**(6):47–60, 1996.

[MWL94] C.A. McHorney, J.E. War Jr, J.F.R. Lu, and C.D. Sherbourne. "The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups." *Medical care*, pp. 40–66, 1994.

[MYG00] K. Malmberg, S. Yusuf, H.C. Gerstein, J. Brown, F. Zhao, D. Hunt, L. Piegas, J. Calvin, M. Keltai, A. Budaj, et al. "Impact of diabetes on long-term prognosis in patients with unstable angina and non–Q-wave myocardial infarction: results of the OASIS (Organization to Assess Strategies for Ischemic Syndromes) Registry." *Circulation*, **102**(9):1014–1019, 2000.

[Noy04] N.F. Noy. "Semantic integration: a survey of ontology-based approaches." *ACM Sigmod Record*, **33**(4):65–70, 2004.

[OKS11] C.K.L. Or, B.T. Karsh, D.J. Severtson, L.J. Burke, R.L. Brown, and P.F. Brennan. "Factors affecting home care patients' acceptance of a web-based interactive self-management technology." *Journal of the American Medical Informatics Association*, **18**(1):51–59, 2011.

[PFS10] C. Pérez-Gandía, A. Facchinetti, G. Sparacino, C. Cobelli, EJ Gómez, M. Rigla, A. de Leiva, and ME Hernando. "Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring." *Diabetes Technology & Therapeutics*, **12**(1):81–88, 2010.

[Pha12] Pharos Innovations. "Pharos Innovations. http://www.pharosinnovations.com/." 2012.

[PMB10] L. Pecchia, P. Melillo, and M. Bracale. "Remote health monitoring of heart failure with data mining via CART method on HRV features." *Biomedical Engineering, IEEE Transactions on*, (99):1–1, 2010.

[PPA77] R. Peto, MC Pike, P. Armitage, N.E. Breslow, DR Cox, SV Howard, N. Mantel, K. McPherson, J. Peto, and PG Smith. "Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples." *British journal of cancer*, **35**(1):1, 1977.

[PSS07] H. Patel, M. Shafazand, M. Schaufelberger, and I. Ekman. "Reasons for seeking acute care in chronic heart failure." *European journal of heart failure*, **9**(6-7):702–708, 2007.

[PTM09] S. Pettifer, D. Thorne, P. McDermott, J. Marsh, A. Villéger, D.B. Kell, and T.K. Attwood. "Visualising biological data: a semantic approach to tool and database integration." *BMC bioinformatics*, **10**(Suppl 6):S19, 2009.

[Qui93] J.R. Quinlan. *C4. 5: programs for machine learning.* Morgan kaufmann, 1993.

[RJB03] M.M. Redfield, S.J. Jacobsen, J.C. Burnett Jr, D.W. Mahoney, K.R. Bailey, and R.J. Rodeheffer. "Burden of systolic and diastolic ventricular dysfunction in the community." *JAMA: the journal of the American Medical Association*, **289**(2):194–202, 2003.

[Sab03] E. Sabaté. *Adherence to long-term therapies: evidence for action.* World Health Organization, 2003.

[SCW11] M. Suh, C.A. Chen, J. Woodbridge, M.K. Tu, J.I. Kim, A. Nahapetian, L.S. Evangelista, and M. Sarrafzadeh. "A Remote Patient Monitoring System for Congestive Heart Failure." *Journal of medical systems*, pp. 1–15, 2011.

[SEC10] M. Suh, L.S. Evangelista, C.A. Chen, K. Han, J. Kang, M.K. Tu, V. Chen, A. Nahapetian, and M. Sarrafzadeh. "An automated vital sign monitoring system for congestive heart failure patients." In *Proceedings of the 1st ACM International Health Informatics Symposium*, pp. 108–117. ACM, 2010.

[Set08] E. Seto. "Cost comparison between telemonitoring and usual care of heart failure: a systematic review." *Telemedicine and e-Health*, **14**(7):679–686, 2008.

[SF97] C. Speier and M. Frese. "Generalized self efficacy as a mediator and moderator between control and complexity at work and personal initiative: A longitudinal field study in East Germany." *Human Performance*, **10**(2):171–192, 1997.

[SFP10] O.Z. Soran, A.M. Feldman, I.L. Piña, G.A. Lamas, S.F. Kelsey, F. Selzer, J. Pilotte, and J.R. Lave. "Cost of medical services in older patients with heart failure: those receiving enhanced monitoring using a computer-based telephonic monitoring system compared with those in usual care: the heart failure home care trial." *Journal of cardiac failure*, **16**(11):859–866, 2010.

[SG02] J.L. Schafer and J.W. Graham. "Missing data: our view of the state of the art." *Psychological methods*, **7**(2):147, 2002.

[SHS02] J. Starren, G. Hripcsak, S. Sengupta, CR Abbruscato, P.E. Knudson, R.S. Weinstock, and S. Shea. "Columbia University's Informatics for Diabetes Education and Telemedicine (IDEATel) Project." *Journal of the American Medical Informatics Association*, **9**(1):25, 2002.

[SLA11] K.W. Singleton, M. Lan, C. Arnold, M. Vahidi, L. Arangua, L. Gelberg, and A.A.T. Bui. "Wireless Data Collection of Self-administered Surveys using Tablet Computers." In *AMIA Annual Symposium Proceedings*, volume 2011, p. 1261. American Medical Informatics Association, 2011.

[SMW12] M. Suh, T. Moin, J. Woodbridge, M. Lan, H. Ghasemzadeh, S. Ahmadi, A. Bui, and M. Sarrafzadeh. "Dynamic Self-adaptive Remote Health Monitoring System for Diabetics." 2012.

[SPL08] O.Z. Soran, I.L. Piña, G.A. Lamas, S.F. Kelsey, F. Selzer, J. Pilotte, J.R. Lave, and A.M. Feldman. "A randomized clinical trial of the clinical effects of enhanced heart failure monitoring using a computer-based telephonic monitoring system in older minorities and women." *Journal of cardiac failure*, **14**(9):711–717, 2008.

[SRR99] D.O. Scharfstein, A. Rotnitzky, and J.M. Robins. "Adjusting for nonignorable drop-out using semiparametric nonresponse models." *Journal of the American Statistical Association*, pp. 1096–1120, 1999.

[SWC09] J.A.C. Sterne, I.R. White, J.B. Carlin, M. Spratt, P. Royston, M.G. Kenward, A.M. Wood, and J.R. Carpenter. "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls." *BMJ: British Medical Journal*, **338**, 2009.

[SWL11] M. Suh, J. Woodbridge, M. Lan, A. Bui, L.S. Evangelista, and M. Sarrafzadeh. "Missing data imputation for remote CHF patient monitoring systems." In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pp. 3184–3187. IEEE, 2011.

[SWT09] S. Shea, R.S. Weinstock, J.A. Teresi, W. Palmas, J. Starren, J.J. Cimino, A.M. Lai, L. Field, P.C. Morin, R. Goland, et al. "A randomized trial comparing telemedicine case management with usual care in older, ethnically diverse, medically underserved patients with diabetes mellitus: 5 year results of the IDEATel study." *Journal of the American Medical Informatics Association*, **16**(4):446–456, 2009.

[The02] The Diabetes Prevention Program Research Group. "The Diabetes Prevention Program (DPP) : Description of lifestyle intervention.", 2002.

[TJT06] Z. Tang, T.R. Johnson, R.D. Tindall, and J. Zhang. "Applying heuristic evaluation to improve the usability of a telemedicine system." *Telemedicine Journal & E-Health*, **12**(1):24–34, 2006.

[TNG97] F. Talbot, A. Nouwen, J. Gingras, M. Gosselin, and J. Audet. "The assessment of diabetes-related cognitive and social factors: the Multidimensional Diabetes Questionnaire." *Journal of Behavioral Medicine*, **20**(3):291–312, 1997.

[Var07] U. Varshney. "Pervasive healthcare and wireless health monitoring." *Mobile Networks and Applications*, **12**(2-3):113–127, 2007.

[VB91] W. Vach and M. Blettner. "Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables." *American Journal of Epidemiology*, **134**(8):895–907, 1991.

[VH03] S. Vijan and R.A. Hayward. "Treatment of hypertension in type 2 diabetes mellitus: blood pressure goals, choice of agents, and setting priorities in diabetes care." *Annals of internal medicine*, **138**(7):593–602, 2003.

[VLL01] R.S. Vasan, M.G. Larson, E.P. Leip, J.C. Evans, C.J. O'Donnell, W.B. Kannel, and D. Levy. "Impact of high-normal blood pressure on the risk of cardiovascular disease." *New England journal of medicine*, **345**(18):1291–1297, 2001.

[WFD98] G.C. Williams, Z.R. Freedman, and E.L. Deci. "Supporting autonomy to motivate patients with diabetes for glucose control." *Diabetes care*, **21**(10):1644–1651, 1998.

[WG00] S. Wu and A. Green. "Projection of Chronic Illness Prevalence and Cost Inflation. Santa Monica, CA: RAND Health.", 2000.

[WHI12] UCLA Wireless Health Institute (UCLA WHI). "UCLA Wireless Health Institute. http://www.wirelesshealth.ucla.edu/." 2012.

[Wil95] W.C. William et al. "Fast effective rule induction." 1995.

[WW02] Y. Wang and I.H. Witten. "Modeling for optimal probability prediction." 2002.

[WWL07] J.H. Wu, S.C. Wang, and L.M. Lin. "Mobile computing acceptance factors in the healthcare industry: A structural equation model." *International journal of medical informatics*, **76**(1):66–77, 2007.

[WWT04] A.M. Wood, I.R. White, and S.G. Thompson. "Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals." *Clinical Trials*, **1**(4):368–376, 2004.

[YYN09] Y. Yamada, K. Yokoyama, R. Noriyasu, T. Osaki, T. Adachi, A. Itoi, Y. Naito, T. Morimoto, M. Kimura, and S. Oda. "Light-intensity activities are important for estimating physical activity energy expenditure using uniaxial and triaxial accelerometers." *European journal of applied physiology*, **105**(1):141–152, 2009.

[ZBS08] M.R. Zile, T.D. Bennett, M.S.J. Sutton, Y.K. Cho, P.B. Adamson, M.F. Aaron, J.M. Aranda Jr, W.T. Abraham, F.W. Smart, L.W. Stevenson, et al. "Transition from chronic compensated to acute decompensated heart failure." *Circulation*, **118**(14):1433–1441, 2008.

[ZYA10] F. Zhou, H.I. Yang, J. Álamo, J. Wong, and C. Chang. "Mobile personal health care system for patients with diabetes." *Aging Friendly Technology for Health and Independence*, pp. 94–101, 2010.