TEMPORAL MINING FRAMEWORK FOR RISK REDUCTION AND EARLY DETECTION OF CHRONIC DISEASES

A Thesis

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

---

by

SOWJANYA PALADUGU

Dr. Chi-Ren Shyu, Thesis Supervisor

MAY 2010

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled

TEMPORAL MINING FRAMEWORK FOR RISK REDUCTION AND EARLY DETECTION OF CHRONIC DISEASES

presented by Sowjanya Paladugu,

a candidate for the degree of Master of Science,

and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Chi-Ren Shyu

Dr. Dmitry Korkin

Dr. Jane M. Armer

ACKNOWLEDGEMENTS

There are several people who played an integral part in the research that led to this thesis. First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Chi-Ren Shyu, for his continuous support throughout my graduate career. He has consistently provided valuable guidance and encouragement which motivated me to work harder and helped me stay focused. I would also like to thank Dr. Jane M. Armer and Dr. Dmitry Korkin for kindly accepting to serve on my committee. Special thanks to Dr. Armer for sharing her expertise on Lymphedema, for providing the Lymphedema dataset, and for her constant encouragement. The research conducted in this thesis was supported by grants from National Institute for Nursing Research and National Institutes of Health.

I would like to thank my peers at the MedBio lab for all their help and support. A special appreciation goes to Jason M. Green for having numerous discussions during the course of this research, and for his input and valuable comments. I am also grateful to Robin P. Shook and Dr. Bob R. Stewart for helping me understand the Lymphedema dataset.

I am forever indebted to my family for their unconditional love and support. Unending support and encouragement came from my husband, Praveen Edara. He was always there to offer suggestions and to cheer me up.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

TEMPORAL MINING FRAMEWORK FOR RISK REDUCTION AND EARLY DETECTION OF CHRONIC DISEASES

Sowjanya Paladugu

Dr. Chi-Ren Shyu, Thesis Supervisor

ABSTRACT

Chronic diseases significantly affect the quality of life of over 25 million Americans and are among the most common and costly health problems. Due to the complexity of these diseases, it is difficult for clinicians to analyze trends in patient data and correlate these trends with other patient information such as demographic data. Therefore, there is a need for informatics tools to efficiently monitor disease progression and to analyze trends in patient data to improve disease management. Moreover, because chronic diseases have been identified as among the most preventable diseases, these tools can also be used to identify patients at risk and provide information for early intervention. To this end, a temporal mining framework was developed to identify frequently occurring temporal patterns in patient measurements that may lead to development of diseases.

The developed framework uses patient data collected over a series of regularly-scheduled clinical visits. Temporal sequences were preprocessed and discretized based on user preferences. Temporal mining was then conducted to identify frequent episodes in measurement sequences before the onset of a disease. The relevance and importance of these episodes were determined by examining the episode frequency and confidence.

Contrast mining was also performed to determine episodes significant to specific patient groups and to conduct side-by-side comparisons of episodes shared among patient groups. The efficacy of the temporal mining framework was evaluated via a case study of lymphedema. The framework was applied to a dataset to study the incidence and severity of lymphedema in post breast cancer patients. Temporal changes in limb volume (LV) measurement data were analyzed via the framework, with patients grouped based on body mass index, occurrence of post-operative swelling, and age. The analysis indicated that similar LV change episodes have varying probabilities of leading to lymphedema in various populations. This framework facilitates the identification of patients at risk of developing a chronic disease and provides useful evidence-based guidelines for making personalized and targeted treatment decisions.

CHAPTER 1

INTRODUCTION

1.1 Motivation

Chronic diseases such as cancer, heart disease, and diabetes are the most common causes of death in the United States and are responsible for as many as 7 out of 10 American deaths each year [1]. They account for 1.7 million deaths in the US each year and significantly affect the quality of life of an additional 25 million Americans [1]. Since detection and management of chronic diseases requires monitoring of a variety of patient measurements over prolonged periods of time, it becomes difficult for physicians and clinicians to keep track of trends in patient data and even more difficult to identify correlations between the trends and other demographic patient information such as comorbidities. There is thus a need for informatics tools to monitor disease progression, analyze trends in patient data, and match these trends to known patient profiles in order to improve disease management.

Although chronic diseases are among the most prevalent, they have also been identified as among the most preventable [1]. Early detection and healthy behaviors play an important role in preventing and controlling the effects of these diseases [1-2]. The relative success of chronic disease treatments was also found to be dependent on the earliness of detection. Thus, it would be very beneficial to develop systematic informatics tools to identify patients at risk of developing a disease and provide information for early intervention. Such early interventions are particularly important in the management of chronic diseases given the irreversible nature of many chronic conditions and the long and costly treatments required to manage them.

To analyze frequently occurring trends in patient measurements that lead to chronic diseases in different patient groups and to identify patients at risk of developing such diseases, a temporal mining framework was developed and is presented in this thesis.

1.2 Needs and Goals for the Temporal Mining Framework

The proposed temporal mining based framework will be useful for analyzing chronic disease datasets with temporal components, with specific focus on extracting significant trends in patient measurements that lead to the development of a specific disease. The significant trends enable the identification of patients at risk of a disease and may predict the future onset of the disease. The recognition of at-risk patients provides opportunities for early interventions for better disease management in such patients.

Since clinical datasets typically contain large amount of patient data with a variety of measurement trends that could possibly lead to diseases, it is important that the temporal mining results be well structured and organized to deduce meaningful and useful information. This is achieved by grouping patients and performing temporal mining in patient groups based on domain-concept mining techniques [3]. By analyzing the significant trends leading to diseases in individual patient groups, evidence-based guidelines can be established in each patient group which would help in the formulation of personalized treatment decisions. In addition to clinical datasets being voluminous in nature, they also differ widely in their data structures as different clinical settings use different sources and procedures to collect patient data. Thus, the temporal mining framework needs to be generic to accommodate these differences.

The growing volumes of electronic patient records and the prevalence of chronic diseases in the United States provide abundant patient data to develop nationally representative patient samples to obtain evidence-based results. Such evidence-based findings can be used to propose recommendations for risk reduction that can be included in the Best Practices documents for the management of diseases.

1.3 Thesis Outline

This thesis document is organized as follows: In CHAPTER 2, the data mining and temporal mining concepts are introduced and a survey of previous studies in clinical domain using these mining techniques is presented. CHAPTER 2 also introduces lymphedema, a chronic condition observed in breast cancer survivors, which is

considered as a case study of the temporal mining framework. In CHAPTER 3, the process flowchart for the proposed framework is described and the problem definition is presented using a formal representation of the framework. In CHAPTER 4, the division of patients into groups and temporal modeling of patient measurements in a dataset are discussed and illustrated with a sample clinical study. CHAPTER 4 also discusses the extraction of frequent episodes leading to a disease using episode mining and the comparison of the extracted episodes across different patient groups. CHAPTER 5 presents a case study of the temporal framework for lymphedema dataset. The patient dataset and user interfaces for data collection are reviewed and the steps described in CHAPTER 4 for developing the temporal mining framework are applied to the lymphedema dataset. CHAPTER 6 includes a discussion and evaluation of the proposed framework using results from the lymphedema case study. It also discusses the limitations of the framework. The thesis is concluded in CHAPTER 7 with a discussion of possible future work.

CHAPTER 2

LITERATURE REVIEW

This chapter focuses on three topics. First, a review of previously published data mining related articles in medical and clinical domains is conducted. The second section focuses on literature on temporal data mining to extract relationships from medical datasets with temporal components. The third section introduces lymphedema, a side effect of breast cancer surgery, which is used as a case study of the temporal mining framework in this thesis.

2.1 Data Mining in Medical Informatics

Data mining is commonly defined as the extraction of previously unknown and potentially useful information from a database [4]. With the growing volumes of electronic patient records, data mining has become popular to extract hidden patterns in patient data for better understanding of relationships within the data. Data mining in medical domain is unique from that in other domains due to the special characteristics of

medical datasets. Medical datasets are often privacy-sensitive, voluminous and heterogeneous with data collected from different sources [5]. The collected data may also need to be characterized mathematically. The rest of the section discusses a few data mining studies that have been conducted in medical and clinical areas.

In clinical domain, data mining techniques have been applied to large clinical repositories containing clinical and administrative data collected from electronic sources to identify new disease associations [6]. The techniques applied include pattern discovery to identify commonly occurring associations in the dataset, predictive analysis to predict future outcome for a patient based on the existing patient records, and association mining to extract interesting rules from the identified associations.

There have been many recent studies to predict the survival of patients with fatal diseases and to predict treatment outcomes. Studies were conducted by Oztekin et. al. [7] to predict the survivability of heart-lung transplantation patients and by Delen et. al. [8] to predict the survivability of breast cancer patients using prediction models such as neural networks [9], decision trees [10], and regression [11]. Decision tree algorithms were also used to effectively predict the survival period of kidney dialysis patients [12] and bladder cancer treatment outcomes [13]. Decision trees based on rules were created and decision making algorithms were used to predict outcomes.

A study by Richards et. al. [14] involved the generation of association rules to find indicators for early mortality. The association between the initial patient visits'

observations and early mortality was investigated and rules were generated based on clinical records of 21000 diabetes patients. Association/dependency rule concepts were also applied to identify head trauma patient needing computed tomography (CT) scans [15]. Rules showing dependencies between patient criteria such as age, sex, intoxication, headache, etc., and CT scan necessity were identified. The rules were evaluated on test data and selected based on the support and confidence.

Data mining has also been used for other functions, for instance to fill knowledge gaps in clinical guidelines used in clinical decision support systems [16]. Clinical guidelines hold medical evidence and provide recommendations for clinical conditions that may be encountered in practice. Data mining techniques such as decision-tree algorithms were used to extract rules pertaining to the choice of treatment, choice of drugs, etc. from patient records. The rules were extracted for patient subgroups with conditions that were either not addressed in the guidelines or had incomplete rules with missing or imprecise recommended action in the guidelines.

Data mining techniques were also used to detect possible adverse drug events (ADEs) [17]. Data analysis in the ADE area is particularly important for the generation of reports of suspected adverse reactions when new medical products are introduced into the market and for the better detection of ADEs to reduce mortality in hospitalized patients.

Automated methods for extracting only the interesting patterns from a large set of mined patterns have been proposed by Siadaty et. al. [18]. They used a dual-mining approach

that is based on a comparison of the strength of a mined pattern from a database to the strength of the equivalent pattern extracted from a knowledgebase. If the strengths of the two patterns are different, then the pattern is considered as potentially interesting.

Although the above mentioned studies make important contributions to the medical field by analyzing patient records using data mining techniques, there are many patient datasets containing time-stamped data which need specialized temporal techniques for analysis. Many clinical tasks such as tracking of chronic diseases and collection of patient measurements over a number of visits are time-oriented. Thus, clinical datasets often contain medical data with a temporal dimension, and therefore temporal reasoning and temporal mining are becoming popular in medical informatics to extract sequential relationships containing significant clinical meaning from such datasets [19].

2.2 Temporal Analysis in Medical Informatics

Temporal/episode mining emerged as a specialized stream of data mining to extract relationships from datasets with temporal dependencies [20-21]. Two focus areas of temporal mining are temporal causal relationship mining and association mining [22]. Temporal data mining has a few distinguishing characteristics when compared to conventional data mining [22]. In time series data, the value of the series at any given instance may depend on values at previous time instances. Also, the association rules generated using temporal mining need to be connected to time periods and are valid only for those time periods [22].

In temporal mining, the time series data is considered as sequence of events with associated times of occurrence. Mining algorithms have been proposed to identify frequent ordered collections of events, called 'episodes' in event sequences [20-21]. Candidate episodes are recognized by sliding a window across event sequences and the identified episodes are used to obtain rules for predicting the behavior of episodes and for evaluating the episode occurrence frequency.

The importance of temporal information in electronic medical records and the effective use of the information for temporal inference such as temporal abstraction, time-oriented decision support, and forecasting tasks have been recognized over the past few years [23]. Temporal information has been successfully used to abstract higher-level concepts such as important context-sensitive summaries and patterns over time periods [24-25], with the obtained knowledge then being applied to various clinical domains. Shahar et. al. proposed knowledge-based temporal abstraction methods to summarize a patient's condition during a time period, to support or modify current treatment plans, and to support recommendations of medical decision-support systems [24]. They also automated the entry process of temporal abstraction knowledge by physicians through the development of knowledge acquisition tools [25].

Temporal abstraction has also been combined with case-based reasoning [26] and used to predict future onset of diseases based on the present patient condition [19]. Temporal sequences of values are abstracted into states such as low, normal, high, etc. and are combined to form trends over time intervals. The trends are characterized and used to

retrieve similar previous cases, which in turn are used to understand and predict future onset of diseases.

Temporal information has been used by many medical information systems to answer time-based clinical queries[27] using TQuery, a specialized database query language to form and execute temporal and contextual based queries on an object-oriented medical record database. It has also been used to provide causal explanations for given clinical conditions in diagnostic reasoning [28].

Temporal mining of sequential clinical datasets provides opportunities for enhanced understanding of medical phenomena such as symptoms, progression and diagnosis, and prediction of future behavior of diseases. Moreover, there is an important need to understand and analyze such datasets since they contain clinically significant temporal knowledge. Evidence-based knowledge such as trends and irregularities extracted from such datasets is required to support clinical decision making [29].

In this thesis, a temporal mining-based framework is proposed to analyze chronic disease datasets with temporal components, with specific focus on extracting significant trends in patient measurements that occur before the development of a particular chronic disease. These trends, also called episodes, enable the identification of at-risk patients and may predict future onset of a disease. They also provide opportunities for early intervention in such patients and establish evidence-based results for personalized treatment decisions.

In order to obtain well structured mining results in relevant patient groups, patients with similar characteristics were grouped based on domain-concept mining (DCM) techniques [3] and mining was performed independently in each patient group. The DCM approach enables the discovery of episodes in under-represented patient groups. Without the use of DCM, such episodes may remain uncovered, particularly with high minimum support thresholds. On the other hand, lowering the minimum support threshold to retrieve less-frequent episodes will increase the computational complexity and return many trivial associations. DCM improves the efficiency of mining algorithm by reducing the size of data set and memory required for computation. The results obtained by using DCM are also naturally organized according to patient groups.

This study differs from previous temporal mining research in the clinical domain in that it systematically generates episodes from discretized measurement changes, utilizes the domain-concept mining approach to form patient groups, and performs temporal mining to improve the understanding of the chronic disease development and progression.

2.3 Lymphedema

The proposed temporal mining approach was applied to a dataset containing measurements from breast cancer survivors for better understanding of patient measurement patterns leading to lymphedema [30-32]. Lymphedema (LE), a chronic condition commonly observed as an after-effect of breast cancer treatment, causes significant swelling in limb areas due to the accumulation of protein-rich fluid [33]. It impacts functional abilities, impairs limb motion, and affects the physical and

psychosocial health of patients [34-35]. Recent statistics have shown that there are an estimated 2.5 million breast cancer survivors in the United States with over 192,000 American women developing breast cancer this year (2009 - 2010) alone [2]. With the advanced treatments available, the survival rate for breast cancer is the highest among all cancers (89% at 5 years after diagnosis) [1]. Unfortunately, although the precise percentage of breast cancer survivors developing LE is not known, up to 40% of the survivors are estimated to develop LE or limb swelling [31]. Moreover, studies have shown that LE is a life-time risk for all breast cancer survivors [36].

In order to raise the profile of LE and to further improve LE management, an international consensus document called 'Best Practice (BP) for the Management of Lymphoedema' document [37] was created, and the guidelines and recommendations from this document are regularly followed by the LE community. The document aims to improve the care provided to LE patients by presenting a model for best practice in LE and by providing useful risk reduction and treatment guidelines for managing LE. Although the BP document associates the recommendations with supporting research evidence where possible, there are still a few guidelines in the document that are based only on experienced judgment or limited evidence, and that require further support from evidence-based research [38-39]. The BP document as well as the LE community would benefit from an informatics framework capable of providing evidence-based treatment decisions for timely LE risk reduction.

The findings from this thesis will demonstrate the potential of temporal mining approach for LE risk detection and for providing opportunities for early intervention which are particularly important in LE analysis, since studies have shown that LE if not diagnosed and managed in its early stages may become irreversible in nature [40]. With the availability of large LE datasets, such findings can be better established and be used to propose new evidence-based recommendations or to further strengthen the existing recommendations for LE risk reduction in the Best Practice (BP) document [37]. Without an informatics framework, it is difficult for physicians and researchers to keep track of temporal trends and identify at-risk patients. The developed framework would take the LE community one step closer to achieving the goal of designing a robust decision support system for LE risk reduction and targeted treatment decisions.

CHAPTER 3

INFORMATICS FRAMEWORK

Temporal mining has been used to analyze time-series data in many different areas since the mid- 90's [20-21]. In this thesis, a temporal data mining framework is proposed for clinical trials and other studies in the medical domain. This framework is suitable for data analysis from a clinical setting where patient measurements are collected at regularly scheduled visits over a period of time. Mining clinical data is distinct from mining in other fields as clinical data often contains large volumes of data collected from different sources with different data structures [5]. Taking this issue into consideration, a generic framework was developed that can be used in a variety of clinical settings. Though there are many possible applications of such a framework, we focus on utilizing it to detect patients at risk of developing a particular disease/condition by identifying relevant changes in patient measurements before its onset. The approach can also be generalized and used for the temporal tracking of chronic diseases.

3.1 Domain concept mining

In order to determine and compare the temporal changes in measurement values across different patient groups, patients are categorized based on relevant attributes. This concept of grouping patients with common characteristics together and performing data mining in each patient group is based on domain-concept mining (DCM) [3]. An important rationale for using the DCM approach is that it is more efficient in identifying possible frequent episodes that lead to a disease in patients belonging to under-represented groups. When the complete patient dataset is considered for mining, even though a given episode frequently leads to a disease in patients having a particular common characteristic, the overall episode support may be lowered by other patient groups where the episode does not lead to disease as frequently. This low overall support for the given episode when all patients are considered may be less than the minimum support threshold, and ultimately such episodes containing useful and important information about disease prediction may not be retrieved. The results mined from smaller patient groups provide more opportunities for personalized disease management.

3.2 Process Flowchart

Figure 1 shows a flowchart for the proposed framework and describes the process for grouping patients and for generating and comparing the frequent temporal episodes in each patient group. First, patients are grouped based on grouping criteria L of the demographic attributes in set D. MCs are calculated by comparing the patient measurements at each visit (M) to the measurements at the baseline visit (B). The

calculated MCs are then discretized into one of the measurement change categories in X. Temporal sequences are generated by considering MCs of a patient across all scheduled visits. Episodes or subsequences are extracted from temporal sequences by taking windows of different widths and sliding the windows across temporal sequences. The episodes that occur at a given distance of *y* from the diagnosis visit z_i are extracted. The frequent episodes are then systematically compared and contrasted across different patient groups to better understand the disease risk differences between patient groups.

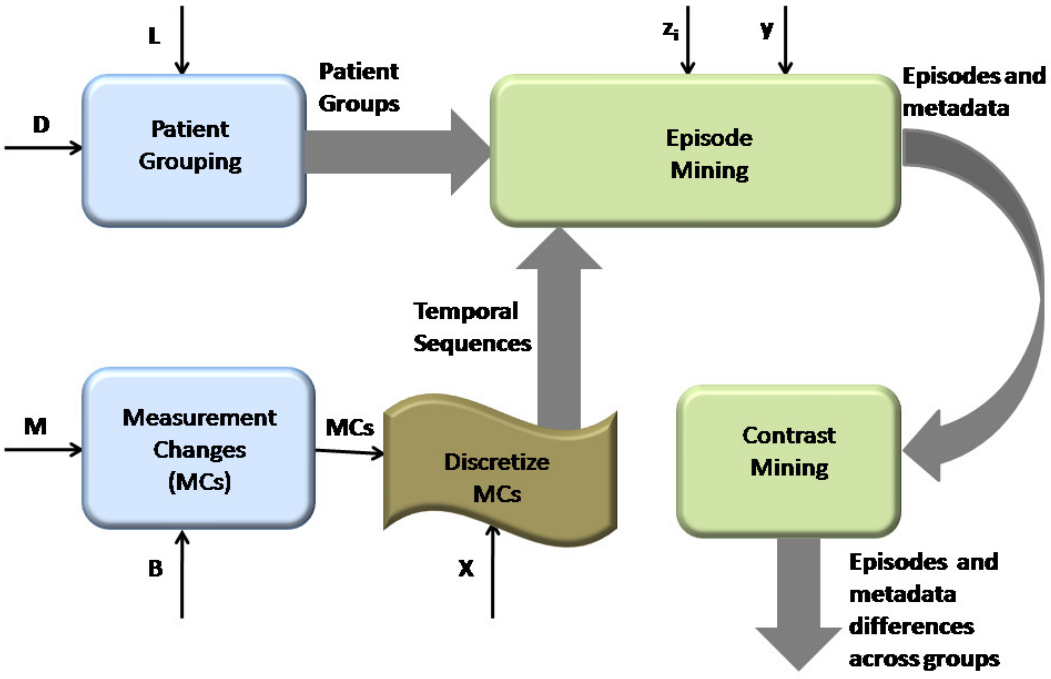

Figure 1: Process flowchart

3.3 Mathematical Notation

A summary of the notation and functions used in the thesis is provided below in Table 1.

Table 1: Notations used in the thesis

D	The set containing demographic attribute values for patients 1 to n_p $D = \{\vec{d}_1, \vec{d}_2 \ldots . \vec{d}_{n_p}\}$ where n_p is the number of patients in the study
$\vec{d}_i$	The vector of demographic attribute values for patient i $\vec{d}_i = (d_{i,1}, d_{i,2} \ldots . . d_{i,n_{da}})$ where n_{da} is the number of demographic attributes
T	The set of visits $T = \{\bigcup_{j\in[1,n_v]} t_j\}$ where t_j is the j^{th} visit of a patient and n_v is the total number of visits
z_i	The disease diagnosis visit for patient i
M	The set containing measurements of patients 1 to n_p $M = \{\vec{m}_1, \vec{m}_2 \ldots . \vec{m}_{n_p}\}$
$\vec{m}_i$	The vector of measurements for patient i $\vec{m}_i = (\vec{m}_{i,1}, \vec{m}_{i,2} \ldots \ldots . \vec{m}_{i,n_{ma}})$ where n_{ma} is the number of measurement attributes
$\vec{m}_{i,k}$	The vector of measurement k values for patient i for visits 1 to n_v $\vec{m}_{i,k} = (\vec{m}_{i,k}^{j} \mid 1 \le j \le n_v)$
L	The set of levels for the attributes used for dividing patients into groups $L = \{\vec{l}_1, \vec{l}_2 \ldots . \vec{l}_{n_{ma}+n_{da}}\}$
$\overrightarrow{l_k}$	The vector of values used for dividing patients into $n_{g,k}$ groups for a demographic attribute k $\overrightarrow{l_k} = (l_{k,1}, l_{k,2} \ldots . . l_{k,n_{l,k}})$ where $n_{l,k} = \begin{cases} n_{g,k} & if\ k\ is\ a\ discrete\ variable \\ n_{g,k} + 1 & if\ k\ is\ a\ continuos\ variable \end{cases}$
X	The set representing measurement change categories $X = \{x, x_2 \ldots \ldots . x_{n_x}\}$ where n_x is the number of measurement categories

3.4 Problem Definition

A formal representation of the framework, including problem definition, is described in this section. A variety of attribute values may be collected from a patient to study factors that lead to a disease or result in disease progression. The collected attributes may include both demographic data as well as measurement data. The demographic attributes are defined as variables that are collected a single time (typically at the initial visit) and assumed to be constant throughout the study (e.g. name, gender, date of birth, blood type). On the other hand, measurement attributes are defined as variables that are collected at defined intervals (typically at each visit) and assumed to change or fluctuate throughout the study (e.g. systolic blood pressure, hematocrit, cholesterol). While demographic attributes are used mainly to divide patients into groups, measurement attributes are used to study the conditions leading to a disease by analyzing changes in measurement attribute values before disease development.

Let $D = \{\vec{d}_1, \vec{d}_2 \ldots . \vec{d}_{n_p}\}$ contain the demographic data for all patients with $\vec{d}_i = (d_{i,1}, d_{i,2} \ldots . . d_{i,n_{da}})$ being the vector of demographic data for patient i, n_p being the total number of patients in the study, and n_{da} being the number of demographic attributes. A set of regularly scheduled visits is defined, $T = \{t_1, t_2, \ldots, t_{n_v}\}$ where n_v is the total number of visits. The first visit in which a disease diagnosis or abnormal level is reached is represented as z_i. Using T, the measurement data for all patients can be defined as $M = \{\vec{m}_1, \vec{m}_2 \ldots . \vec{m}_{n_p}\}$, where $\vec{m}_i = (\vec{m}_{i,1}, \vec{m}_{i,2} \ldots \ldots . \vec{m}_{i,n_{ma}})$ contains the

measurement data for patient i and $\vec{m}_{i,k} = (\vec{m}_{i,k}^{j} | \ 1 \leq j \leq n_v)$ is patient i's vector of data for measurement attribute k at a visit j and n_{ma} is the number of measurement attributes.

To better understand the notation described above, consider a sample clinical study of n_p patients in which the measurement attributes, weight and LDL (low-density lipoproteins) cholesterol ($n_{ma} = 2$), and demographic attributes, gender and age ($n_{da} = 2$) (see Table 2), are collected over four visits ($n_v = 4$). This sample clinical study will continue to be used in the next few chapters.

Table 2: Attributes collected in the sample clinical study

Attributes			
Measurement Attributes		Demographic Attributes	
Weight	LDL cholesterol	Gender	Age

Suppose further that patient $p_1$is a 30-year-old female who weighs 130 lbs and has a LDL cholesterol of 100 mg/dL during the t_1 visit; weighs 135 lbs and has a LDL cholesterol of 130 mg/dL during the t_2 visit; weighs 137 lbs and has a LDL cholesterol of 130 mg/dL during the t_3 visit; and weighs 132 lbs and has a LDL cholesterol of 125 mg/dL during the t_4 visit. The demographic vector for patient p_1 would be $\overrightarrow{d_1}$= ('Female', 30). Similarly, the measurement vector for patient $p_1$would be $\vec{m}_1$ = ((130, 135, 137, 132), (100, 130, 130, 125)) as shown in Table 3.

Table 3: Example measurements for patient $\boldsymbol{p_1}$

Measurement Attribute (k)	$\vec{m}_{1,k}$			
	t_1	t_2	t_3	t_4
Weight (in lbs)	130	135	137	132
LDL cholesterol (in mg/dL)	100	130	130	125

CHAPTER 4

DATA MINING

4.1 Temporal Modeling

Although data mining algorithms may be applied to the full set of data (or patients), generally the most interesting results are those mined from specific subsets of the data (or patient groups). In this framework, the demographic attributes are used to divide the patient data set into groups. In particular, let $L = \{\vec{l}_1, \vec{l}_2 \ldots . \vec{l}_{n_{da}}\}$ represent the criteria (levels) for patient grouping where $\vec{l}_k = (l_{k,1}, l_{k,2} \ldots . . l_{k,n_{l,k}})$ is the vector of values used for splitting demographic attribute *k* into groups and $n_{l,k}$ is the number of such dividing values for this attribute. On the basis of levels L of demographic attributes D, patients are divided into groups as shown in Figure 1. For discrete variables, $\vec{l}_k$ will allow for grouping into $n_{l,k}$ groups, one for each possible value of the variable; however, for continuous variables, $\vec{l}_k$ will allow grouping into $n_{l,k} - 1$ groups with each element of the vector forming a group boundary condition.

For example, for the discrete gender attribute, the levels vector $\overrightarrow{l_1} = ('Male', 'Female')$ can be used to divide patients into two groups as shown in Table 4. However, for the continuous age attribute, if one wanted to divide the patient into the four groups [0, 25), [25, 50), [50, 75), [75, 150) (see Table 5), the levels vector would need to contain the following five elements: $\overrightarrow{l_2} = \{0, 25, 50, 75, 150\}$.

Table 4: Example patient groups based on gender

Gender	
Male	Female

Table 5: Example patient groups based on age

Age (in years)			
< 25	25 – 49	50 – 74	75 – 150

It should be noted that patient groups can be formed using multiple attributes by recursively applying the grouping function. For example, a patient dataset grouped on the basis on gender can further be divided into groups based on age as shown in Table 6.

Table 6: Example patient groups formed by recursive grouping based on the gender and age criteria respectively

Patients							
Male				Female			
< 25 yrs	25 – 49 yrs	50 – 74 yrs	75 – 150 yrs	< 25 yrs	25 – 49 yrs	50 – 74 yrs	75 – 150 yrs

Temporal analysis of patient measurements can be performed based on the changes in measurement values over a monotonically increasing time frame. Measurements at the baseline visit are used as the basis for comparison to calculate measurement changes (MCs) as shown in Figure 1. The baseline, b_k for a measurement attribute *k* can either be specified absolutely as a particular visit (*e.g.* t_1, t_2) or relatively, as one visit with respect to the current visit being analyzed (*e.g.* $\Delta t(1)$, $\Delta t(2)$). For example, if the measurements at visit t_j are being analyzed to determine the measurement change episodes, the baseline could be specified absolutely as t_1, in which case the measurements at visit t_j are compared to the measurements at visit t_1, or relatively as $\Delta t(1)$, in which case the measurements at visit t_j are compared to the measurements at visit t_{j-1}. The visit at which patient p_i is diagnosed with the disease of interest or in which some measurement reaches an abnormal state is represented by z_i . For instance, if patient p_1 is diagnosed with heart disease at visit t_4, then $z_i = 4$.

For data mining, the values of the MCs should be discrete in nature. To facilitate the discretization of MCs (see Figure 1), $X = \{x_1, x_2 \ldots \ldots . x_{n_x}\}$ is introduced as the set of

discretized values for a MC, where n_x is the number of such categories. The general function used to translate MCs into discretized values based on the baseline b_k is shown as:

$$f\left(m_{i,k}^{j}, m_{i,k}^{b_k}\right) = \delta \mid \delta \in X \tag{1}$$

where i refers to the patient, j the visit, and k the measurement attribute.

For example, if X = { '↗', '↘', '→'} where '↗' represents an increase in a measurement value when compared to the measurement at baseline, '↘' represents a decrease in the measurement value, and '→' represents a stable measurement value, then one could define a discretizing function as:

$$f\left(m_{i,k}^{j}, m_{i,k}^{b_k}\right) = \begin{cases} \nearrow & if\ m_{i,k}^{j} > m_{i,k}^{b_k} \\ \searrow & if\ m_{i,k}^{j} < m_{i,k}^{b_k} \\ \rightarrow & if\ m_{i,k}^{j} = m_{i,k}^{b_k} \end{cases} \tag{2}$$

If this function is applied to an entire set of measurements for an attribute, it will generate a temporal sequence that can be mined. This sequence of MCs for patient i over visits 1 to n_v is represented by $F\left(\vec{m}_{i,k}, m_{i,k}^{b_k}\right)$ and is defined as

$$F\left(\vec{m}_{i,k}, m_{i,k}^{b_k}\right) = \left(f\left(m_{i,k}^{1}, m_{i,k}^{b_k}\right), f\left(m_{i,k}^{2}, m_{i,k}^{b_k}\right), \ldots, f\left(m_{i,k}^{n_v}, m_{i,k}^{b_k}\right)\right) \tag{3}$$

Continuing with the example clinical study introduced in chapter 3, if weight change is determined by comparing weight at visit $t_1 (b_1 = t_1)$ and the cholesterol change is determined by comparing cholesterol at the previous visit ($b_2 = \Delta t(1)$), then the resultant

temporal sequences for weight and LDL cholesterol values (see Table 7) generated from patient measurement values in Table 3 are:

$$F\left(\vec{m}_{1,1}, m_{1,1}^{b_1}\right) = (\nearrow \nearrow \searrow)$$

$$F\left(\vec{m}_{1,2}, m_{1,2}^{b_2}\right) = (\nearrow \rightarrow \searrow)$$

Table 7: MC sequences for weight and LDL cholesterol attributes for patient $\boldsymbol{p_1}$

Measurement (k)	$F\left(\vec{m}_{1,k}, m_{1,k}^{b_k}\right)$		
	$t_1 - t_2$	$t_2 - t_3$	$t_3 - t_4$
weight	↗	↗	↘
LDL cholesterol	↗	→	↘

4.2 Episode Mining

To extract usable MC episodes from temporal sequences, support and confidence are defined using windows of change. A MC episode is defined as a subsequence of the full length MC sequence of a patient from visits 1 to n_v, $F\left(\vec{m}_{i,k}, m_{i,k}^{b_k}\right)$. The frequency of a MC episode leading to a disease after *y* visits is defined as the fraction of same-size windows that end a distance of *y* before the diagnosis visit (*z*) in which the episode occurs. Given a window *w* of width *win*, and $W_y(win)$, the set of all windows with width *win* at a distance of *y* before the diagnosis visit *z*, the frequency of a MC episode α leading to a disease in *y* visits can be obtained as:

$$fr_y(\alpha,\ win) = \frac{|\{w \in W_y(win)|\ \alpha \text{ occurs in w}\}|}{|\ W_y(win)|} \quad (4)$$

The probability of a MC episode leading to a disease in *y* visits is measured using the concept of confidence in episode mining. The confidence of an episode α leading to a disease is calculated as:

$$conf_y(\alpha,\ win) = \frac{|\{w \in W_y(win)|\ \alpha \text{ occurs in w}\}|}{|\{w \in W(win)|\ \alpha \text{ occurs in w}\}|} \quad (5)$$

where $W(win)$ is the set of windows of width *win* irrespective of disease diagnosis.

Given, an attribute *k*, the baseline for that attribute b_k, and the measurement vectors $\vec{m}_{G_r,k}$ for all patients in the demographic group G_r, the episode mining function (see Figure 1) that returns all MC episodes with frequency greater than a threshold frequency fr_{min} and confidence greater than a threshold confidence $conf_{min}$ is given as:

$$epi\left(G_r, \vec{m}_{G_r,k}, \overrightarrow{b_k}, fr_{min}, conf_{min}\right) = \left\{\left(\alpha, fr_y(\alpha,\ win), conf_y(\alpha,\ win)\right)\right\} \quad (6)$$

where $fr_y(\alpha,\ win) > fr_{min}$ and $conf_y(\alpha,\ win) > conf_{min}$

The episode mining function essentially returns the episodes that occur frequently in a patient group and cause a disease with high probability.

4.3 Contrast Mining

The frequency and confidence values of the episodes returned from the episode mining function are then systematically compared across different patient groups (see Figure 1) to understand risk differences between patient groups. This mining in various patient groups is analogous to emerging pattern mining and contrast mining [41-42] which have become popular over the past few years. Contrast mining [41] is useful to detect

differentiating characteristics between non-overlapping groups. For example, it can be used to identify patient groups with a high probability of developing disease when compared to other patient groups. Emerging patterns [42] are described as episodes whose supports significantly increase across different groups. Emerging pattern mining can be used, for instance, to examine increase in disease risk with an increase in BMI of the patient.

4.4 Mining Algorithm

Given the measurements, baselines, and patient groups, Algorithm 1 describes the steps to find temporal MC episodes and to calculate the frequency and confidence of the episodes leading to the considered disease. The variable $seq_{k,b_{k,}\,p_i}$ in line 3 contains the temporal sequence of a patient from visits 1 to n_v. Episodes are extracted from $seq_{k,b_{k,}\,p_i}$ by taking windows of widths 1 to *len*($seq_{k,b_{k,}\,p_i}$) and moving the windows across the sequence $seq_{k,b_{k,}\,p_i}$.

For example, consider the temporal sequence of LDL cholesterol changes in the previous example from Table 7, $F\left(\vec{m}_{1,2}, m_{1,2}^{b_2}\right)$ = (↗ → ↘). Episodes (α) are extracted from this sequence as shown below, by considering windows of widths 1 to 3 and moving the window across the sequence.

S = {'↗','→','↘'} for *win* = 1

S = {'↗ →',' → ↘'} for *win* = 2

S = {'↗ → ↘'} for *win* = 3.

Thus, the set of all possible episodes that can be extracted from the sequence '↗ → ↘' is S = {'↗', '→', ' ↘', '↗ →',' → ↘', '↗ → ↘'}. The frequency and confidence of episodes leading to the disease in *y* visits are calculated. For example, if one wants to find the episodes that lead to a disease after one visit (*y* = 1) and the patient is diagnosed with a disease at t_4 visit (z = 4), then one would find the frequencies for the set of episodes that end at the ($z - y$) visit, i.e. the 3rd visit. The set of candidate episodes in this case that end in the 3rd visit is {' →' ,'↗ →'} as illustrated in Table 8.

Table 8: Episode generation from temporal sequence $\boldsymbol{F\left(\vec{m}_{1,2}, m_{1,2}^{b_2}\right)}$

Temporal sequence $F\left(\vec{m}_{1,2}, m_{1,2}^{b_2}\right)$ over visits $t_1 - t_4$	Episodes generated from temporal sequence						Episodes (α) ending at one visit before diagnosis (at t_3 visit)
	win = 1		*win* = 2		*win* = 3		
	α	ends at	α	ends at	α	ends at	
↗ → ↘	↗	t_2	↗ →	t_3	↗ → ↘	t_4	→ ↗ →
	→	t_3	→ ↘	t_4			
	↘	t_4					

Line 11 in the algorithm shows the output which contains the episode α, the frequency of the episode leading to the disease in *y* visits $fr_y(\alpha,\ len(\alpha))$, and the confidence of the episode leading to the disease $conf_y(\alpha,\ len(\alpha))$. All episodes with a frequency greater than a threshold frequency fr_{min} and a confidence greater than a threshold confidence $conf_{min}$ are considered as frequent episodes which lead to the disease with high probability and are stored in the frequent episode set, *FS*. The frequency and confidence

values of episodes in *FS* are compared across different patient groups as shown in line 16 to analyze and understand the variations in disease risks in different patient groups.

Algorithm 1.

1. for each measurement attribute $k \in A$

2. for each patient $p_i \in G_r$

3. $seq_{k,b_{k,}\,p_i} \leftarrow F\left(\vec{m}_{i,k}, m_{i,k}^{b_k}\right)$

4. find episodes from $seq_{k,b_{k,}\,p_i}$

5. store episodes in S

6. end for

7. for each episode $\alpha \in S$

8. calculate $fr_y(\alpha,\ len(\alpha))$

9. calculate $conf_y(\alpha,\ len(\alpha))$

10. if ($fr_y(\alpha,\ len(\alpha)) >\ fr_{min}$ and $conf_y\left(\alpha,\ len(\alpha)\right) > conf_{min}$)

11. output α, $fr_y(\alpha,\ len(\alpha))$, $conf_y(\alpha,\ len(\alpha))$

12. store episode α in *FS*

13. end if

14. end for

15. for each episode $\alpha \in FS$

16. compare $fr_y(\alpha,\ len(\alpha))$ and $conf_y(\alpha,\ len(\alpha))$ across groups $G_1, G_2, \ldots . G_{n_{g,k}}$

17. end for

18. end for

CHAPTER 5

CASE STUDY OF TEMPORAL MINING FOR LYMPHEDEMA DATASET

The temporal mining approach described in the previous chapters was evaluated by applying to a LE dataset to identify and analyze commonly occurring episodes in limb volume changes (LVCs) before the development of LE. The discovered frequent episodes can help to identify patient groups who are at a greater risk of developing LE in the near future. They are expected to be used to better predict the onset of LE, and therefore to provide more opportunities for early intervention, which has been shown to be important in the successful management of LE [43].

5.1 Subjects

The data required for this 30-month NIH-funded study were collected from 233 US Midwestern women who have been diagnosed with breast cancer and scheduled for

surgery (n_p= 233). Participants selected for the study had to be over 18 years of age with no prior history of LE or breast cancer. In addition, the ability to understand English and give informed consent was required. Participants with ages ranging from 26 to 95 years were recruited, and data including limb volume measurements as well as demographic characteristics such as body mass index (BMI), age, dominant-side, and cancer-affected side were collected at a US Midwestern university-affiliated state cancer center.

Of all the participants in the study, 221 were unilateral breast cancer survivors while the remaining 11 were affected by breast cancer in both left and right sides. Out of the 221 unilateral breast cancer survivors, 109 patients were affected by cancer in their dominant side while 112 patients had their cancer-affected side different from their dominant side. Also, 37 of the 233 patient study group experienced swelling after breast cancer surgery during the post-operative visit.

5.2 Data Collection

5.2.1 Data Collection Timeline

The data required for this study were collected by trained research staff members at 9 scheduled visits ($T = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9\};\ n_v$= 9): preoperative (before surgery), post-operative (at approximately 2 weeks after surgery), every three months post-op during the first year, and every six months thereafter, as shown in Figure 2. In the figure, t_j represents the j^{th} visit of a patient.

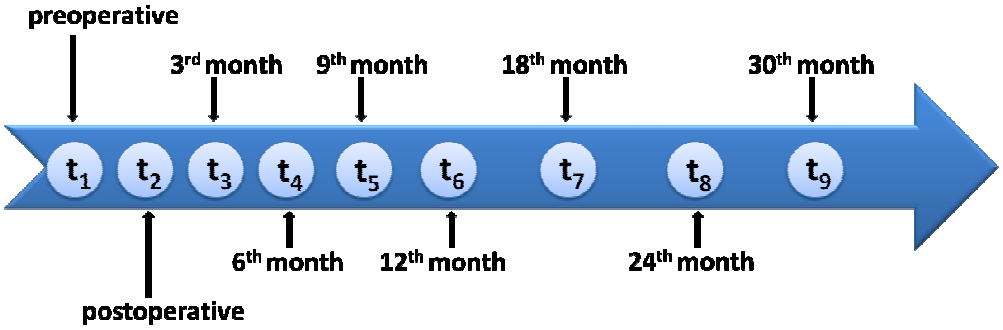

Figure 2: Data collection timeline

5.2.2 Data Collection Interface

The primary LE-related measurement used in this study was limb volume (LV) of the cancer affected side, which was measured each visit using two different methods: perometry and circumference.

The limb volume (LV) of each arm was first calculated by performing infra-red perometry with the arm in a horizontal position [44]. Figure 3 shows the user-interface that was developed to record the LV measured using an optoelectronic volumetry device called the perometer. The measurements are taken three times at each visit and an average value is calculated to minimize errors in measurement. In the below sample measurements shown in Figure 3, the LV of interest is the average LV of the left limb since the patient is affected by cancer in the left side.

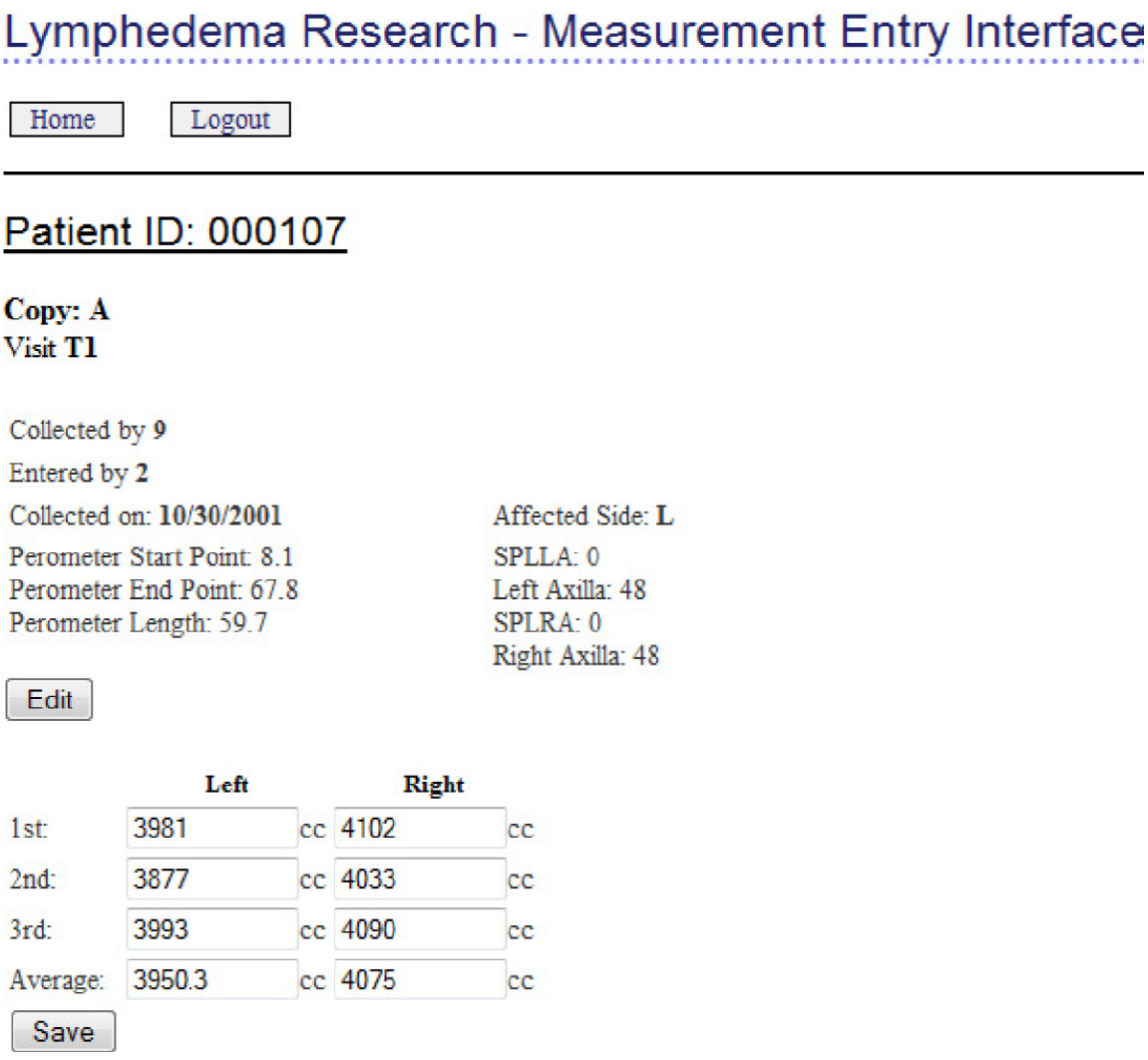

Figure 3: User interface for collecting perometer measurements

In addition to perometer readings, circumference measurements (in cm) were collected using a nonstretch, flexible tape measure at the hand, at the wrist, and every 4 cm thereafter to the axilla [44]. From the partial screenshot shown in Figure 4, it can be observed that circumference measurements were taken at 4cm from the wrist (CM4), at 8cm from the wrist (CM8), and so on.

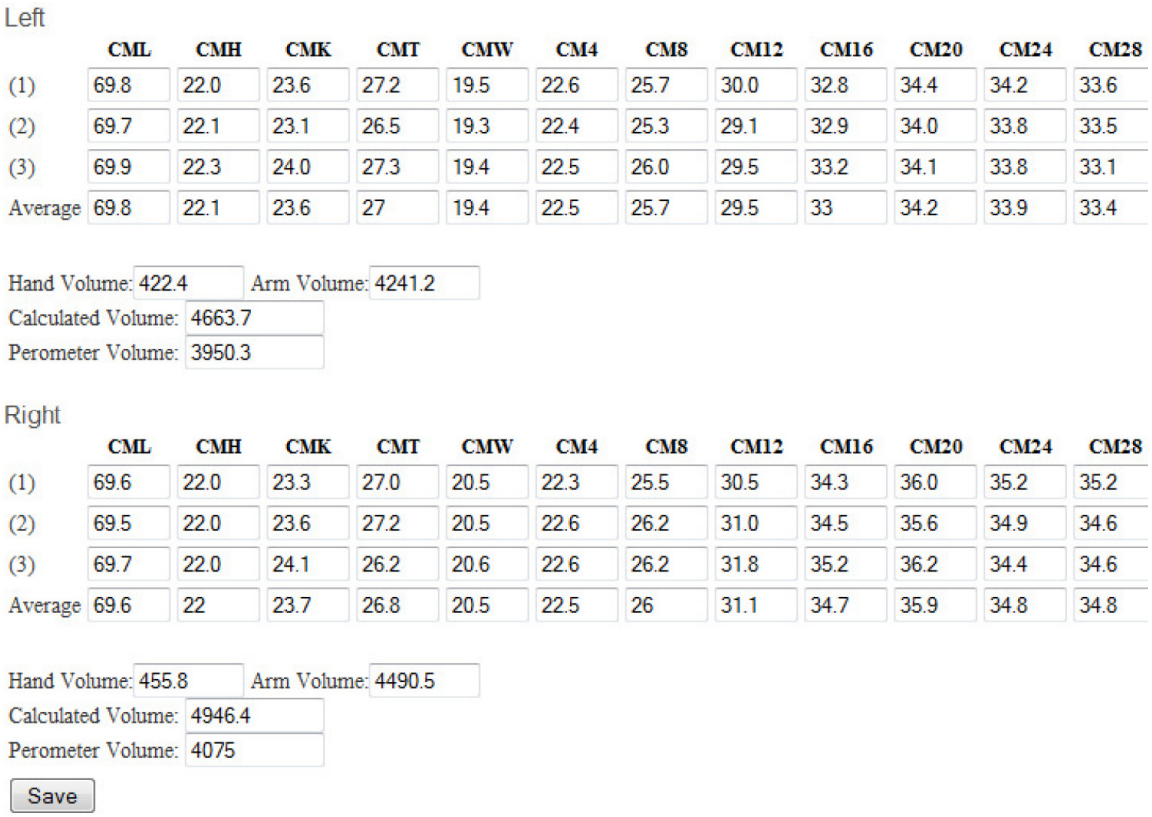

Figure 4: User interface for collecting circumference measurements

LV was calculated from circumference measurements as the sum of hand and arm volumes. Hand volume (vol_{hand}) was estimated using Eq. 7.

$$vol_{hand} = 123 + (0.036 * cmw^2 * cmh) \quad (7)$$

where,

cmw is the circumference of the wrist and,

cmh is the length of the hand measured from the longest fingertip to the wrist.

Arm volume was calculated by summing the volumes of each 4-cm segment of the arm using the following truncated cone formula:

$$vol_{arm} = \sum_i \frac{cml_i * (c_{1i}^2 + c_{1i}c_{2i} + c_{2i}^2)}{12\pi} \quad (8)$$

where,

cml_i is the length of the arm segment (4 cm in this case),

c_{1i} is the circumference at the lower end of the i^{th} arm segment, and

c_{2i} is the circumference at the upper end of the i^{th} arm segment.

In addition to measurement data, demographic data of the patients such as BMI, age, etc. were also collected. While clinically measurable demographic data such as BMI were measured and collected, certain other demographic/historical data such as age, cancer-affected side, and dominant side were self-reported by patients during the preoperative visit, t_1.

Figure 5 shows all the attributes collected from patients and illustrates the relationship between attributes via the Entity-Relationship Diagram (ERD).

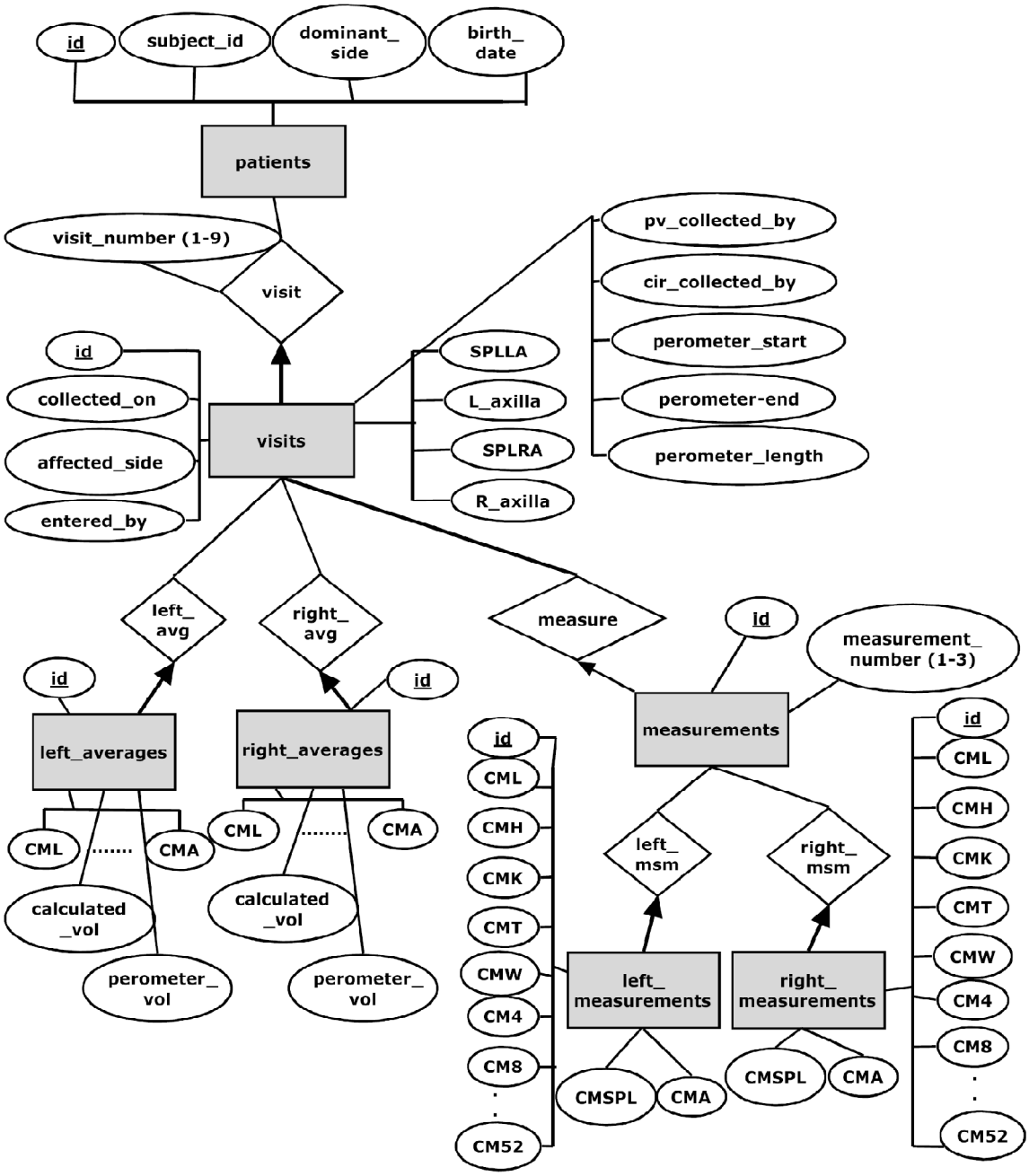

Figure 5: The entity-relationship diagram for LE dataset

5.3 Patient groups

In order to determine LE risk in different patient groups, patients were grouped based on: body mass index (BMI), occurrence of post-operative swelling, and age. BMI groupings were based on the guidelines from Centers for Disease Control and Prevention (2008) [45] , and the age variable was partitioned into four roughly equal sized groups.

The grouping criteria for BMI, post-operative swelling, and age demographic attributes are:

1. *BMI*: [0, 18.5), [18.5, 25.0), [25.0, 30.0), [30.0, ∞)
2. *Post-operative swelling*: patients with swelling equivalent to the subclinical LE criteria of 3% LVC at the post-operative (t_2) visit [46], patients without such swelling at the post-operative (t_2) visit
3. *Age*: [0, 55), [55, 65), [65, 75), [75, ∞)

In terms of the definitions described in Table 1,

For the BMI attribute, $\overrightarrow{l_1} = \{0, 18.5, 25.0, 30.0, \infty\};\ n_{g,1} = 4$

For the post-operative swelling attribute,

$\overrightarrow{l_2} = \{post - op\ swelling, no\ post - op\ swelling\};\ n_{g,2} = 2$

For the age attribute, $\overrightarrow{l_3} = \{0, 55, 65, 75, \infty\};\ n_{g,3} = 4$

The average time for the onset of LE was observed to be 6.9 months post-operative in a recent study by Stout et al. [46]. By considering the closest visit (visit t_4 at the 6th post-operative month) to the average LE onset time, LVC episodes were further analyzed within and after the first six months of surgery by dividing each of the BMI, post-operative swelling, and age groups into two sub groups. Patients who were diagnosed with LE in the first six months were categorized as one group, and those who were diagnosed after six months as another group.

5.4 Data Selection

Perometric LV was primarily used in this study due to the high accuracy of perometer in volume estimation [47-48]. For patients whose perometric measurements were unavailable (31 patients) due to equipment not being in operation or being serviced at the time of the patient visit, circumferential LV was used instead. This substitution is justified by the significant correlation (r = 0.89) between the perometric and circumferential LVs in this study.

5.5 Measurement Change Discretization

For the temporal analysis, sequences of limb volume change (LVC) were mined. At each visit t_j, the LV at the previous visit $t_{j-1}$was considered as the baseline (b_1= $\Delta t(1)$) and LVC was analyzed by comparing volume with the LV at previous visit. The set of measurement change categories used in this research is X = { '↗', '↘', '→'}. Each LVC was discretized into one of the three categories:

1. rise in LV (↗)
2. drop in LV (↘)
3. stable LV (→)

A rise in LV (↗) is defined as a 3% or greater increase in volume when compared to the LV at the previous visit; a drop in LV (↘) is defined as a 3% or greater decrease in LV; and stable LV (→) is defined as a LVC of less than 3%. As slight fluctuations in weight and fluid retention are natural, the 3% threshold was selected for discretization analogous to the study by Stout et al. [46]. If the LV was unavailable at either of the visits being considered due to the corresponding patient missing the visit, it was denoted by an 'x'; episodes containing an 'x' were not considered for the temporal analysis.

5.6 Episode selection

Episodes for mining were then defined by identifying the subsequences in each patient's full LVC sequence. For those patients who developed LE during the study, only subsequences occurring before developing LE were recorded, as LVCs after being diagnosed with LE can no longer provide useful information in estimating the risk of developing LE.

There are several LE diagnostic criteria available: 200 ml perometry LVC [44]; 10% perometry LVC [44]; 2cm circumferential increase [44]; self-reported signs and symptoms of heaviness and swelling [44]; and the 5% BMI-adjusted LVC criterion [49].

Some criteria (e.g., 10% perometry LVC) are more conservative in the definition of LE as compared to some others (e.g., 2cm circumferential increase); however, there is no standardized method for diagnosing LE [44]. In this paper, the 5% BMI-adjusted LVC criterion [49] was used as a proxy for development of LE due to its consideration of commonly experienced weight fluctuations following breast cancer treatment while using LV for LE assessment.

To illustrate the conversion of LVs to temporal sequences, examples of LVC sequences and LE diagnoses are provided in Figure 6 for three patients (0177, 0251, 0266 with ages 70, 44, 54 years, and BMIs 35.5, 27, 20.7 kg/m^2 respectively) affected by unilateral LE. The visit at which LE diagnosis condition according to the 5% BMI adjusted criterion was met is indicated by 'LE'. For patient 0177, LV was stable between the t_1 and t_2 visits, the t_2 and t_3 visits, and the t_3 and t_4 visits; it then increased between the t_4 and t_5 visits with the LE criterion being met at the t_5 visit ($z = 5$). Similarly, patient 0266 had stable LV from t_1 to t_6, but experienced an increase in LV between t_6 and t_7 visits, with the LE criterion being met at t_7 ($z = 7$). On the other hand, patient 0251 had two stable periods between t_1 and t_3, and subsequently had two consecutive increases between t_3 to t_4 and t_4 to t_5 and at t_5 the LE criterion was met ($z = 5$).

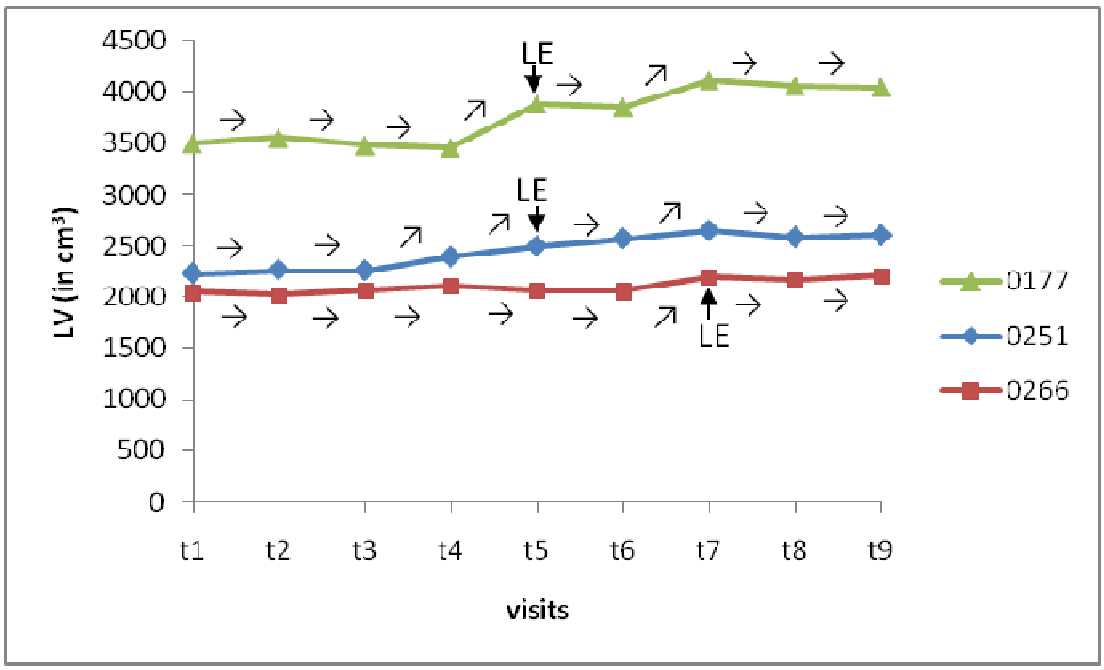

Figure 6: LVCs over t_1 – t_9 visits

It is to be noted that only the subsequence occurring before the development of LE was considered for episode generation. For example for patient 0177, the full LV sequence is $F(\vec{m}_{0177,1}, m^{b_1}_{0177,1})$ = (→ → → ↗ → ↗ → →), where $b_1 = \Delta t(1)$; $z_{0177} = 5$. Since the patient developed LE at visit t_5, only the subsequence (→ → → ↗) occurring before the t_5 visit was used for generating episodes. The set of episodes generated from the subsequence (→ → → ↗) is

S = {'→',' ↗',' → →',' → ↗',' → → →',' → → ↗',' → → → ↗'}.

5.7 Frequent episodes

In this study, only LVC episodes in windows with at least a width of 2 (*win* ≥ 2) occurring *immediately* before LE (*y* = 0) were considered for identifying the frequent

episodes. Because of this, each patient meeting the LE criterion can only have at most one episode for each window width. For example, for the patient 0177, only '→ ↗', '→ → ↗', and '→ → → ↗' episodes for window widths 2, 3, and 4 respectively are considered for generating frequent episodes. The episode '→ → →' corresponding to changes from t_1 to t_4 would not be included for patient 0177 as the episode ends at a distance $y = 1$ from the LE diagnosis visit, $z = 5$.

The frequency and confidence values of a LVC episode immediately leading to LE are calculated using equation 4 and equation 5 respectively with $y = 0$. Frequency and confidence thresholds of 15% (fr_{min} = 15%; $conf_{min}$= 15%) were used, meaning that LVC episodes that occur with a frequency of at least 15% were considered frequent and those that occur with a confidence greater than 15% were considered as episodes leading to LE with significant probability. It is important to note that if an episode α (such as '→ → → ↗') is frequent, then all sub-episodes of α '→ → ↗' and '→ ↗' are also frequent [20].

CHAPTER 6

RESULTS AND DISCUSSION

6.1 Applications and Discussion of the Temporal Mining Framework

The temporal mining model developed in this research can be used in various clinical settings to monitor the progression of chronic diseases, to analyze trends in patient data, to identify patients at risk, and to provide information for early interventions. To accommodate the diversely structured clinical data collected from different sources, we attempted to make the temporal mining model as generic as possible. The temporal framework provides flexibility in choosing baseline visits. The visit relative to which measurement changes provide an accurate representation of disease progression can be considered as the baseline visit. The number of levels used to discretize measurement changes is also flexible. Any number of levels can be used based on clinically significant differences in patient measurements depending on the study. In our case study, three levels - increase (↗), decrease (↘), and stable (→) were used based on a 3% LV change compared to the previous visit. However, as the number of levels increases, the number of possible episodes also increases. With the large number of possible episodes that could

lead to a disease, it may become difficult to establish frequent and statistically significant results in small datasets.

The temporal mining model can be used to identify MC episodes that have a high probability of leading to disease in the future. For example, for a given MC episode observed in a patient, the model can be used to predict probability of the patient developing the disease in 3 months, in 6 months, and so on. Furthermore, the model can also be adjusted to study temporal sequences after events such as treatment administration to assess treatment outcomes.

6.2 Results of the Lymphedema Case Study

The 233 breast cancer affected patients were categorized into discrete groups based on BMI, post-operative swelling, and age as defined in section 5.3. For better understanding of the results, the distribution of patient population in each group is summarized prior to reporting the frequent LVC patterns leading to LE in each patient group. Each patient group is further divided into two sub-groups based on the time-frame at which the LE criterion was met (within and after the first six months after surgery) and similar results are reported. The frequent LVC patterns leading to LE can serve as guidelines for identifying future patients at risk of developing LE. Some of the guidelines useful in making evidence-based treatment decisions are shown in the subsections below and can either be added to the Best Practices (BP) document [37] or used as evidence to strengthen the existing recommendations.

6.2.1 BMI Groups

The percentage of women in the study per BMI group and the percentage of women in each BMI group affected by LE (using the 5% BMI-adjusted LVC criterion) are shown in Table 9. It should be noted that the underweight group was not studied due its small sample size (0.86%)

Table 9: BMI-based group compositions

BMI	Weight Status	Number of patients	Percentage of patients	Percentage of the group affected by LE
Below 18.5	Underweight	2	0.86%	-
18.5 - 24.9	Normal	53	22.74%	33.96%
25.0 - 29.9	Overweight	78	33.47%	51.28%
30 and above	Obese	98	42.06%	55.1%

- *Overweight and obese patients are at a greater risk of meeting the LE criterion*

  It was observed that higher percentages of overweight and obese patients met the LE criterion when compared to normal weight patients (51.28% versus 33.96%; p = 0.051 and 55.1% versus 33.96%; p = 0.013 respectively in Table 9). This evidence supports that obesity is an important risk factor for LE, as stated in the BP document.

LVC patterns leading to LE with frequency and confidence thresholds of 15% in at least one of the patient groups were identified and a subset of the results obtained are shown in

Table 10. It should be noted that the window width *win* is equal to the length of the LVC pattern and henceforth will not be specified separately.

Table 10: Frequent LVC patterns associated with LE by BMI groups

LVC Pattern (α)	$conf_0(\alpha,\ \text{win})$ (in %)			$fr_0(\alpha,\ \text{win})$ (in %)		
	Normal	Overweight	Obese	Normal	Overweight	Obese
→ ↗	14.28	**27.27***	**33.33**	35.71	**39.47**	**37.25**
↗ ↗	18.18	**35.39**	**40.90**	14.28	**15.78**	**17.64**
→ → ↗	**23.52**	**21.73**	**29.41**	**30.76**	**18.51**	**17.85**
→ ↗ ↗	-**	**54.54**	**41.67**	-	**22.22**	**17.85**
→ → → ↗	**28.50**	10.00	**37.50**	**22.22**	7.14	**17.64**
→ → ↗ ↗	-	**100.00**	50.00	-	**28.57**	5.88

*confidence and frequency values > 15% are in bold face

** indicates that the corresponding patterns did not occur in the patient group

- *The probability of a LVC pattern resulting in LE varies with BMI*

  From Table 10, it can be seen that similar LVC patterns have varying probabilities of leading to LE based on the BMI group in which the patterns occur. For example, the probability of consecutive increases in LV (↗ ↗) resulting in LE varies across BMI groups (18.18% in normal BMI group, 35.39% in overweight BMI group, and 40.9% in obese BMI group).

- *The risk of a LVC pattern leading to LE increases with BMI and is higher for overweight and obese patients*

From Table 10, it can be observed that when the LV remained stable between two visits and then increased over the next two visits (→ ↗), the confidence of the LVC pattern '→ ↗' leading to LE by the next visit was 14.28% in a normal BMI patient, while it was significantly higher (33.33% versus 14.28%; p = 0.044) in obese BMI patients.

- *Patients with consecutive LV increase have a greater probability of meeting the LE criterion*

  From Table 10, it can be observed that LVC patterns in which LV increased in at least two consecutive visits ('→ ↗ ↗') were observed to have higher confidences of resulting in LE when compared to the patterns with a single increase in LV ('→ → ↗') (54.54% versus 21.73%; p = 0.058 in overweight patients).

- *A variety of LVC patterns lead to meeting the LE criterion with high confidence values in overweight and obese patients*

  From Table 10, it can be observed that in overweight and obese patients there are a greater number of LVC patterns that could lead to LE with high confidence values. For example, the LVC pattern '→ ↗ ↗' leads to LE with confidence values of 54.54% and 41.67% in overweight and obese patients respectively. This further strengthens the recommendation in the BP document that patients should be encouraged to maintain an optimal healthy body weight.

Patients in each BMI category were further divided into two groups to analyze the frequent LVC patterns associated with LE within and after the first six months of surgery,

and the results are shown in Table 11. It should be noted that the same patient can have different patterns of varying lengths. For example, a patient can have the '→ → ↗ ↗' pattern as well as its sub-patterns '→ ↗ ↗', and '↗ ↗'. Since the overweight and obese patients are at a greater risk of LE, only results for such patients are shown in Table 11 for the purpose of clarity.

Table 11: LVC patterns associated with LE within and after six months of surgery by BMI groups

LVC Pattern (α)	$conf_0(\alpha,\ \text{win})$ (in %)				$fr_0(\alpha,\ \text{win})$ (in %)			
	Overweight		Obese		Overweight		Obese	
	LE in first 6 months	LE after 6 months	LE in first 6 months	LE after 6 months	LE in first 6 months	LE after 6 months	LE in first 6 months	LE after 6 months
↘ ↗	25.00	8.33	12.50	**16.00**	8.33	7.14	3.12	**21.05**
→ ↗	**40.00***	7.50	**40.63**	13.95	**50.00**	21.43	**40.63**	31.58
↗ ↗	18.18	**30.77**	**46.15**	**13.64**	8.33	**28.57**	**18.75**	**15.79**
→ → →	**25.00**	1.75	11.11	4.54	**23.08**	7.14	20.00	16.67
→ → ↗	**50.00**	8.69	**40.00**	**16.67**	**23.08**	14.20	**20.00**	**16.67**
→ ↗ ↗	**28.50**	**44.44**	**33.33**	**37.50**	**15.38**	**28.57**	**20.00**	**16.67**
↗ ↘ ↗	**-****	20.00	50.00	**37.50**	**-**	7.14	10.00	**16.67**
↗ → ↗	**100.00**	14.28	**-**	14.28	**15.38**	7.14	**-**	11.11
→ → → ↗	-	7.69	-	**27.27**	-	7.14	-	**17.65**
→ → ↗ ↗	-	**80.00**	-	25.00	-	**28.57**	-	5.88

*confidence and frequency values > 15% are in bold face

** indicates that the corresponding patterns did not occur in the patient group

- *Stable LV does not rule out the chances of meeting the LE criterion in overweight and obese patients, particularly in the first six months after surgery*

  From Table 11, it was observed that there was a possibility of developing LE in the first six months after surgery even when LV remained stable ('→ → →' in overweight and obese patients with confidences 25% and 11.11%), whereas, such incidents did not occur often after six months (see Table 11). Recall that a pattern of consecutively stable LVs by our criterion may still reflect slow gradual increases in LV.

- *In general, a single LV increase has high confidence of leading to LE in the first six months of surgery. After the first six months of surgery, continuous increase in LV has a greater confidence of meeting the LE criterion when compared to a single LV increase.*

  It was observed that in general, a single time increase in LV resulted in LE with high confidence within the first six months after surgery ('→ → ↗' with confidences 50% and 40% in overweight and obese patients respectively). After the first six months of surgery LVC patterns with continuous increase were observed to have higher confidences of leading to LE when compared to single time LV increase (44.44% for '→ ↗ ↗' versus 8.69% for '→ → ↗'; $p = 0.019$ in overweight patients in Table 11).

6.2.2 Postoperative Swelling Groups

The percentage of patients with and without swelling criteria in the postoperative visit (T_1), and the percentage of LE-affected patients in each group are shown in Table 12.

Table 12: Postoperative swelling-based group compositions

Postoperative swelling	Number of patients	Percentage of patients	Percentage of the group affected by LE
No	196	84.12%	45.41%
Yes	37	15.88%	64.86%

- *Postoperative swelling is an important risk factor for LE development*

  From Table 12, about 64.86% of patients experiencing postoperative swelling met the LE criterion at a later point, while only 45.41% of patients without postoperative swelling met the criterion (see also Mahamaneerat et al. [49]). This demonstrates the association between postoperative swelling and LE ($p = 0.030$) and supports the inclusion of postoperative swelling as one of the risk factors for LE development.

The confidence and frequency values of LVC patterns leading to LE (with confidence and frequency thresholds > 15% in at least one of the groups) in patients with and without postoperative swelling are shown in Table 13.

Table 13: Frequent LVC patterns associated with LE by postoperative swelling groups

LVC Pattern (α)	$conf_0(\alpha,\ \text{win})$ (in %)		$fr_0(\alpha,\ \text{win})$ (in %)	
	Without post-op swelling	With post-op swelling	Without post-op swelling	With post-op swelling
↘ ↗	17.50	**42.85**	8.33	**15.00**
→ ↗	**26.27***	**36.36**	**42.85**	**20.00**
↗ ↗	28.94	**50.00**	13.09	**30.00**
→ → ↗	**24.49**	**25.00**	**21.05**	**18.18**
→ ↗ ↗	**40.00**	50.00	**17.54**	9.09
↗ ↘ ↗	50.00	**40.00**	7.02	**18.18**
↗ → ↗	23.08	**66.67**	5.26	**18.18**
→ → → ↗	22.22	**28.57**	11.43	**40.00**
→ → ↗ ↗	57.14	**100.00**	11.43	**20.00**

*confidence and frequency values > 15% are in bold face

- *A variety of LVC patterns lead to meeting the LE criterion with high confidence values in patients with postoperative swelling*

  From Table 13, a single instance of LV increase in postoperative swelling patients was found to lead to LE with high confidence (42.85%, 36.36%, and 50% for LV patterns '↘ ↗', '→ ↗', and '↗ ↗' respectively) irrespective of whether the volume decreased, remained stable or increased in the previous visit.

The frequent LVC patterns associated with LE within and after the first six months of surgery in each of the patient groups based on postoperative swelling are shown in Table 14.

Table 14: LVC patterns associated with LE within and after six months of surgery by postoperative swelling groups

LVC Pattern (α)	$conf_0(\alpha,\ \text{win})$ (in %)				$fr_0(\alpha,\ \text{win})$ (in %)			
	Without post-op swelling		With post-op swelling		Without post-op swelling		With post-op swelling	
	LE in first 6 months	LE after 6 months	LE in first 6 months	LE after 6 months	LE in first 6 months	LE after 6 months	LE in first 6 months	LE after 6 months
→ ↗	**36.00***	9.68	100.00	**15.38**	**56.25**	25.00	14.28	**33.33**
↗ ↗	20.00	**17.94**	**50.00**	**22.22**	8.33	**19.44**	**28.57**	**33.33**
→ → ↗	**38.89**	11.90	-	**20.00**	**33.33**	13.89	-	**40.00**
→ ↗ ↗	**26.67**	**33.33**	-	**33.33**	**19.04**	**16.67**	-	**20.00**
↗ ↘ ↗	-**	28.57	**50.00**	-	-	11.11	**33.33**	-
↗ → ↗	-	12.50	**100.00**	-	-	8.33	**33.33**	-
→ → → ↗	-	16.67	**-**	**28.57**	-	11.42	-	**40.00**
→ → ↗ ↗	-	40.00	**-**	**100.00**	-	11.14	-	**20.00**

*confidence and frequency values > 15% are in bold face

** indicates that the corresponding patterns did not occur in the patient group

- *In general, a given LVC pattern has a greater confidence of meeting the LE criterion within the first six months after surgery when compared to after six months.*

  For example, when LV remains stable between two visits and then increases at the next visit ('→ ↗'), there is a greater chance of resulting in LE in the first six months of surgery when compared to later (36% versus 9.68%; $p < 0.0001$ in patients without postoperative swelling in Table 14). Such evidence-based findings indicate a need for patient vigilance and careful monitoring for LE

symptoms in postoperative swelling patients, especially the first several months after surgery.

6.2.3 Age Groups

Patients were partitioned into four roughly equal-sized groups based on their age. The categorization details and the percentages of patients in each group affected by LE in our data sample are shown in Table 15.

Table 15: Age-based group compositions

Age	Number of patients	Percentage of patients	Percentage of the group affected by LE
Below 55	55	23.6%	56.36%
55 - 64	66	28.32%	50.00%
65 - 74	65	27.89%	43.07%
75 and above	46	19.74%	45.65%

- *The risk of developing LE decreases with age*

  It was observed that in general, the chances of meeting the LE criterion tended to decrease with age ($p = 0.29$), contrary to the common assumption that older patients are at a higher risk of developing LE (see also Armer et al. [50]). While 56.36% of young patients (below 55 years) met the LE criterion, 45.65% of old patients (above 75 years) met the criterion (see Table 15).

The frequent LVC patterns associated with meeting the LE criterion and their frequency and confidence levels across different age groups are shown in Table 16.

Table 16: Frequent LVC patterns associated with LE by age groups

LVC Pattern (α)	$conf_0(\alpha,\ \text{win})$ (in %)				$fr_0(\alpha,\ \text{win})$ (in %)			
	< 55 years	55 – 64 years	65 – 74 years	≥ 75 Years	< 55 years	55 – 64 years	65 – 74 years	≥ 75 years
↘ ↗	12.50	26.67	**30.76**	9.09	3.33	12.90	**16.00**	5.56
→ ↗	**36.11**	**23.40**	**29.73**	**17.86**	**43.33**	**35.48**	**44.00**	**27.78**
↗ ↗	**33.33**	**26.31**	28.57	**66.67**	**20.00**	**16.13**	8.00	**22.22**
→ → ↗	**36.84**	18.18	**21.05**	12.50	**35.00**	10.00	**26.67**	7.69
→ ↗ ↗	33.33	**30.77**	50.00	**100.00**	15.00	**20.00**	6.67	**23.08**
→ → → ↗	**62.50**	-**	12.50	-	**33.33**	-	10.00	-
→ → ↗ ↗	**60.00**	100.00	50.00	-	**20.00**	10.00	10.00	-

*confidence and frequency values > 15% are in bold face

** indicates that the corresponding patterns did not occur in the patient group

- *As a general phenomenon, the frequency of a LVC pattern meeting the LE criterion is higher in young patients when compared to old patients*

  Although there were certain cases to support the contrary, the general observation was that the same LVC pattern had a greater chance of being associated with LE in young patients (< 55 years) when compared to old patients (>75 years). For example, when the LV was stable between two visits and then increased between the next two visits ('→ ↗'), there was a 36.11% probability of developing LE in patients younger than 55 years while the probability dropped to 17.86% in patients older than 75 years; p = 0.11 (see Table 16).

- *LE risk is lifelong*

  It was observed that even when LV remained stable for a long time ('→ → → ↗'), young patients are still at a significant risk of LE (62.5% in patients less than 55 years age in Table 16).

The above results obtained by dividing patients into groups based on age were not found to be as statistically significant as the results obtained by dividing patients on the basis of BMI and postoperative swelling. Further partitioning of patients in different age groups in our dataset depending on whether the LE criterion was met within or after the first six months of surgery resulted in an insufficient patient sample size and no significant results. Hence, those results have not been included in the paper. We would like to note here that we also partitioned the patients based on dominant side and breast cancer-affected side, but our analysis did not yield any statistically significant results for such groups and hence those results have not been included in the thesis.

6.3 Study Limitations

There are a number of admitted limitations to this study. First, though the temporal mining approach attempts to capture a wide variety of patient data, it cannot capture natural language data used in a clinical setting. While commonly used natural language information such as 'The patient had a high fever can be formulated as questions with 'Yes' and 'No' options, it is not possible to capture all natural language information. Second, the best criteria for dividing patients into groups are not always known

beforehand. While dividing patients into groups based on some criteria can give a good understanding of the risk factors and group specific MC episode information that lead to a disease, some other criteria may not yield significant results. Finally, we discuss the data set. Though the LE dataset used in the study has the advantages of having baseline, preoperative measurements and several timestamps, the sample is relatively small, especially considering that the dataset is partitioned so that episodes can be mined in individual groups. There are also missing values in several patients' sequences of LV measurements which affect the frequency of mined episodes.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

In the paper, we proposed a temporal mining framework to identify and study frequent temporal changes in measurements that lead to a disease in different groups of patients. The approach can be used to identify patients at risk of developing a disease, track disparities in disease progression in different patient groups, monitor the effect of treatment plans in patient groups, and plan appropriate early interventions based on the measurement change episodes. The analysis of measurement change episodes in different patient groups will increase the understanding of the various risk factors for a disease and will enable clinicians to make more targeted and personalized treatment decisions.

To demonstrate the usefulness of the framework, we applied the temporal mining approach to a LE dataset in order to identify breast cancer survivors at a risk of developing LE by examining the limb volume changes before the LE diagnosis in different groups of patients. Our temporal analysis of the LE dataset shows that in

general, limb volume change episodes have a higher probability of being associated with LE in obese patients than do the same episodes in normal weight patients. Similarly, the same trend holds for patients with post-operative swelling as compared to patients without post-operative swelling. Furthermore, our data revealed that a given limb volume change episode has a greater chance of resulting in LE within the first six months after surgery when compared to the same episode occurring after six months. Such evidence-based findings of interest can either be added to the Best Practices document or used as evidence to strengthen existing recommendations.

The findings reported in the case study demonstrate the potential of the proposed informatics framework for understanding the risk factors of a disease and with the availability of large datasets, such findings can be well-established and can help design a robust decision support system to establish evidence-based intervention decisions.

7.2 Future Work

The temporal mining model developed in this thesis can be generalized in the future to be used in clinical settings where patient visits may not be regularly scheduled. The current model considers the time period between two successive visits to be the same for all patients. An interesting extension would be to adjust the model to accommodate patient information that has not been collected at pre-scheduled visits, but was collected whenever the patient experienced symptoms of a disease and visited the clinic. The temporal framework can also be extended to accommodate missing values in the patient dataset which may occur due to the corresponding patient missing a scheduled visit. This

can be explored, for example, through the imputation of missing data based on previous visit and next visit patient information.

The model can be further evaluated by applying to large chronic disease datasets such as datasets containing records of heart disease or diabetic patients. Such large datasets will better establish the significance of temporal mining results. The results obtained can be compared to known clinical information to estimate efficiency of the model in recognizing measurement change patterns leading to a disease. Previously unknown results can be used to propose new evidence-based guidelines for disease control.

The current temporal mining model concentrates on studying disease leading measurement change (MC) patterns generated from a single measurement. However, some diseases may depend on many factors and may require multiple measurements to be analyzed simultaneously. For such cases, the temporal mining model can be enhanced to study combinations of MC patterns resulting from multiple measurements.

Future work can also include integration of the temporal mining framework with electronic medical records of patients to connect to large patient datasets with access to all the information stored in the medical records. This would provide disease progression and risk information for individual patients and would ultimately help in building a robust decision support system.

BIBLIOGRAPHY

[1] Centers for Disease Control and Prevention (CDC). (2009, October). *Chronic Disease Prevention and Health Promotion*. Available: http://www.cdc.gov/nccdphp/

[2] American Cancer Society. (2009, October). *Breast Cancer Facts & Figures 2009-2010.* Available: http://www.cancer.org/downloads/STT/F861009_final 9-08-09.pdf

[3] W. Mahamaneerat, C-R. Shyu, S. Ho, and C. Chang, "Domain-Concept Association Rules Mining for Large Scale and Complex Cellular Manufacturing Tasks," *Journal of Manufacturing Technology Management,* vol. 18, pp. 787-806, 2007.

[4] M. H. Dunham, *Data Mining: Introductory and Advanced Topics*, 1st ed., New Jersey: Prentice Hall/Pearson Education, 2003.

[5] K. J. Cios and G. W. Moore, "Uniqueness of medical data mining," *Artif Intell Med,* vol. 26, pp. 1-24, 2002.

[6] I. M. Mullins, M. S. Siadaty, J. Lyman, K. Scully, C. T. Garrett, W. G. Miller, R. Muller, B. Robson, C. Apte, S. Weiss, I. Rigoutsos, D. Platt, S. Cohen, and W. A. Knaus, "Data mining and clinical data repositories: Insights from a 667,000 patient data set," *Comput Biol Med,* vol. 36, pp. 1351-77, 2006.

[7] A. Oztekin, D. Delen, and Z. J. Kong, "Predicting the graft survival for heart-lung transplantation patients: an integrated data mining methodology," *Int J Med Inform,* vol. 78, pp. 84-96, 2009.

[8] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artif Intell Med,* vol. 34, pp. 113-27, 2005.

[9] S. Haykin, *Neural networks: a comprehensive foundation*. New Jersey: Prentice Hall, 1998.

[10] J. R. Quinlan, "Induction of decision trees," *Machine Learning,* vol. 1, pp. 81-106, 1986.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning* New York: Springer-Verlag, 2001.

[12] A. Kusiak, B. Dixon, and S. Shah, "Predicting survival time for kidney dialysis patients: a data mining approach," *Comput Biol Med,* vol. 35, pp. 311-27, 2005.

[13] S. C. Shah, A. Kusiak, and M. A. O'Donnell, "Patient-recognition data-mining model for BCG-plus interferon immunotherapy bladder cancer treatment," *Comput Biol Med,* vol. 36, pp. 634-55, 2006.

[14] G. Richards, V. J. Rayward-Smith, P. H. Sonksen, S. Carey, and C. Weng, "Data mining for indicators of early mortality in a database of clinical records," *Artif Intell Med,* vol. 22, pp. 215-31, 2001.

[15] S. P. Imbermana, B. Domanskia, and H. W. Thompsonb, "Using dependency/association rules to find indications for computed tomography in a head trauma dataset, " *Artificial Intelligence in Medicine,* vol. 26, pp. 55-68, 2002.

[16] M. Toussi, J.-B. Lamy, P. Toumelin, and A. Venot, "Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes," *BMC Medical Informatics and Decision Making,* vol. 9, 2009.

[17] B. Honigman, P. Light, R. M. Pulling, and D. W. Bates, "A computerized method for identifying incidents associated with adverse drug events in outpatients," *Int J Med Inform,* vol. 61, pp. 21-32, 2001.

[18] M. S. Siadaty and W. A. Knaus, "Locating previously unknown patterns in data-mining results: a dual data- and knowledge-mining method," *BMC Med Inform Decis Mak,* vol. 6(13), 2006.

[19] R. Schmidt and L. Gierl, "A prognostic model for temporal courses that combines temporal abstraction and case-based reasoning," *Int J Med Inform,* vol. 74, pp. 307-15, 2005.

[20] H. Mannila, H. Toivonen, and A. Verkamo, "Discovery of Frequent Episodes in Event Sequences," *Data Mining and Knowledge Discovery,* vol. 1, pp. 259-289, 1997.

[21] R. Agrawal and R. Srikant, "Mining Sequential Episodes," in *Proceedings of the Eleventh International Conference on Data Engineering*, Taipei, 1995, pp. 3-14.

[22] H. Ning, H. Yuan, and S. Chen, "Temporal Association Rules in Mining Method," in *Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences*, Hangzhou, 2006, pp. 739-742.

[23] J. C. Augusto, "Temporal reasoning for decision support in medicine," *Artif Intell Med,* vol. 33, pp. 1-24, 2005.

[24] Y. Shahar and M. A. Musen, "Knowledge-based temporal abstraction in clinical domains," *Artif Intell Med,* vol. 8, pp. 267-98, 1996.

[25] Y. Shahar, H. Chen, D. P. Stites, L. V. Basso, H. Kaizer, D. M. Wilson, and M. A. Musen, "Semi-automated entry of clinical temporal-abstraction knowledge," *J Am Med Inform Assoc,* vol. 6, pp. 494-511, 1999.

[26] A. Aamodt and E. Plaza, "Case-based reasoning: foundation issues. Methodological variation and system approaches," *AI Commun,* vol. 7, pp. 39-59, 1994.

[27] M. G. Kahn, S. Tu, and L. M. Fagan, "TQuery: a context-sensitive temporal query language," *Comput Biomed Res,* vol. 24, pp. 401-19, 1991.

[28] W. Long, "Temporal reasoning for diagnosis in a causal probabilistic knowledge base," *Artif Intell Med,* vol. 8, pp. 193-215, 1996.

[29] M. Berlingerio, F. Bonchi, F. Giannotti, and F. Turini, "Mining Clinical Data with a Temporal Dimension: A Case Study," in *IEEE International Conference on Bioinformatics and Biomedicine*, California, 2007, pp. 429-436.

[30] J. M. Armer, "The problem of post-breast cancer lymphedema: impact and measurement issues," *Cancer Invest,* vol. 23, pp. 76-83, 2005.

[31] J. M. Armer, B. R. Stewart, and R. P. Shook, "Occurrence of Lymphoedema Post-Breast Cancer Treatment," *Journal of Lymphoedema,* vol. 4, pp. 14-18, 2009.

[32] J. A. Petrek and M. C. Heelan, "Incidence of breast cancer-related lymphedema," *Cancer,* vol. 83, pp. 2276-2781, 1998.

[33] P. S. Mortimer, "The pathophysiology of lymphedema," *Cancer,* vol. 83, pp. 2798-802, 1998.

[34] M. M. Hull, "Functional and psychosocial aspects of lymphedema in women treated for breast cancer," *Innovations in Breast Cancer Care,* vol. 3, pp. 97 - 100, 1998.

[35] M. E. Radina and J. M. Armer, "Surviving breast cancer and living with lymphedema: Resiliency among women in the context of their families," *Journal of Family Nursing,* vol. 10, 2004.

[36] J. Ferlay, F. Bray, P. Pisani, and D. Parkin, *Cancer Incidence, Mortality and Prevalence Worldwide.* IARC Cancer Base No. 5, version 2.0. Lyon, France: IARC Press, 2004.

[37] Lymphedema Framework, *Best Practice for the Management of Lymphoedema*, London: MEP Ltd, 2006.

[38] E. Muscari, "Lymphedema: responding to our patients' needs," *Oncol Nurs Forum,* vol. 31, pp. 905-12, 2004.

[39] S. H. Ridner, "Breast cancer lymphedema: pathophysiology and risk reduction guidelines," *Oncol Nurs Forum,* vol. 29, pp. 1285-93, 2002.

[40] B. D. Lawenda, T. E. Mondry, and P. A. Johnstone, "Lymphedema: a primer on the identification and management of a chronic condition in oncologic treatment," *CA Cancer J Clin,* vol. 59, pp. 8-24, 2009.

[41] S. D. Bay and M. J. Pazzani, "Detecting Group Differences: Mining Contrast Sets," *Data Min. Knowl. Discov,* vol. 5, pp. 213-246, 2001.

[42] G. Dong and J. Li, "Efficient Mining of Emerging Episodes: Discovering Trends and Differences," in *Knowledge Discovery and Data Mining*, 1999, pp. 43-52.

[43] J. R. Casley-Smith, "Alterations of untreated lymphedema and it's grades over time," *Lymphology,* vol. 28, pp. 174-85, 1995.

[44] J. M. Armer and B. R. Stewart, "A Comparison of Four Diagnostic Criteria for Lymphedema in a Post-Breast Cancer Population," *Lymphat Res and Biol,* vol. 3, pp. 208-217, 2005.

[45] Centers for Disease Control and Prevention (CDC). (2009 October). *Body Mass Index: BMI for Adults* Available: http://www.cdc.gov/healthyweight/assessing/bmi/

[46] N. L. S. Gergich, L. A. Pflazer, C. McGarvey, B. Springer, L. H. Gerber, and P. Soballe, "Preoperative Assessment Enables the Early Diagnosis and Successful Treatment of Lymphedema," *Cancer,* vol. 112, pp. 2809-19, 2008.

[47] H. N. Mayrovitz, N. Sims, and J. Macdonald, "Assessment of limb volume by manual and automated methods in patients with limb edema or lymphedema," *Adv Skin Wound Care,* vol. 13, pp. 272-6, 2000.

[48] S. Tierney, M. Aslam, K. Rennie, and P. Grace, "Infrared optoelectronic volumetry, the ideal way to measure limb volume," *Eur J Vasc Endovasc Surg,* vol. 12, pp. 412-7, 1996.

[49] W. K. Mahamaneerat, C. R. Shyu, B. R. Stewart, and J. M. Armer, "Post-Op Swelling and Lymphoedema Following Breast Cancer Treatment: A Baseline-Comparison BMI-Adjusted Approach," *Journal of Lymphoedema,* vol. 3, pp. 20-25, 2008.

[50] J. M. Armer and M. R. Fu, "Age Differences in Post-Breast Cancer Lymphedema Signs and Symptoms," *Cancer Nursing,* vol. 28, pp. 200-209, 2005.