Temporal Data Mining in Electronic Medical Records from Patients with Acute Coronary Syndrome

Wynona Black

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Meliha Yetisgen, Co-Chair

John H. Gennari, Co-Chair

Eric Horvitz

Program Authorized to Offer Degree:

Biomedical and Health Informatics

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

University of Washington

Abstract

Temporal Data Mining in Electronic Medical Records from Patients with Acute Coronary Syndrome

Wynona Black

Chair of the Supervisory Committee:

Meliha Yetisgen

John H. Gennari

Department of Biomedical Informatics and Medical Education

Every 25 seconds someone in the US has a cardiac event and one person per minute will die from it. ST-elevated myocardial infarction (STEMI), non ST-elevated myocardial infarction and unstable angina are caused by ischemia and referred to as acute coronary syndrome (ACS). STEMI is the most severe and accounts for a quarter of ACS cases. Research in STEMI treatment tends to focus on a single event and the risks/benefits thereof. The interaction between events during an encounter is especially important in STEMI, where the timing of treatments is crucial for positive patient outcomes. However there is a dearth of research into the temporal relationships between events.

To explore temporal relationships between events, I created a sequential pattern mining algorithm (SPM) and a temporal association rule mining algorithm (TARM) to mine the Acute Coronary Syndrome Patient Database (ACSPD). The ACSPD is a large, 9 year EMR database derived from 128 healthcare institutions across the US. I simulated data to examine SPM performance and found that it is well-suited to extract

patterns from noisy data. The TARM is designed to discover rules comprised of 3 temporally ordered events, i.e. clinical practice patterns (CPP).

Using the SPM in the ACSPD, I discovered 39 order sets. Not all order sets are present for the 9 year span and overall order set use drops in 2004. I postulate that this denotes a shift in medical practice. In late 2004, the American Heart Association (AHA) published new STEMI treatment guidelines. I condensed the ACSPD sequences using the order sets then applied TARM. Using support, confidence, Bayes' factor, lift, likelihood, and Zhang's measures, I found substantial variation, rarity and weak antecedent-consequent pairing in the CPPs. To explore the interaction between clinical decisions and patient outcomes, I compared the CPPs with AHA STEMI performance measures for compliance and analyzed the risk of bleeding and mortality. CPP compliance with STEMI performance measures decreases mortality and bleeding risk, but there is evidence of complex interactions between measures that augments/masks the effect. The contributions of this work are 1) exploring CPPs and their effect on patient outcomes using EMR big data and 2) algorithm performance evaluation using simulation.

Acknowledgements

To my family and friends

Table of Contents

Chapter 1 Temporal Data Mining and Acute Coronary Syndrome 1
1.1 Acute Coronary Syndrome 1
1.2 Temporality in Data Mining 2
1.3 Research Statement 3
1.4 Contributions 3
Chapter 2 Sequential Pattern Mining 5
2.1 Mining Sequential Patterns, a Review of Related Works 5
2.2 Sequential Pattern Mining Algorithm 8
2.3 Sequential Pattern Mining Thresholds 12
2.4 Simulating Event Sequence Data 12
2.5 Evaluation Metrics: Precision, Recall and F-measure 14
2.6 Edit Distance and Embedded Patterns 15
2.7 Evaluation of the SPM Algorithm Using Various Simulated Data Parameters 16
2.8 Evaluation of the SPM Algorithm Using Various User-defined Parameters 20
2.9 Embedded Known Patterns and Edit Distance 23
2.10 Static vs. Dynamic Confidence Thresholds 26
2.11 Conclusion 26
Chapter 3 Temporal Association Rule Mining 29
3.1 Association Rule Data Mining 29
3.2 Temporal Association Rule Data Mining 31
3.3 Temporal Association Rule Mining Algorithm 32
3.4 Conclusion 35
Chapter 4 Interestingness Measures for Temporal Association Rules 36
4.1 Objective and Subjective Measures 36
4.2 Properties and Selection Criteria for Objective Measures 38
4.3 Subjective Measures and Incorporating Expert Knowledge 42
4.3.1 Performance Measures as a Knowledge Base 45
4.4 Conclusion 47
Chapter 5 SPM and TARM in ACSPD 50
5.1 Acute Coronary Syndrome Patient Database 50
5.1.1 Events 52
5.1.2 Encounters 53
5.1.3 Outcomes 53
5.2 Study Design 54
5.3 Sequential Pattern Mining to Find Order Sets 55
5.4 Discover Clinical Practice Patterns Using Temporal Association Rule Mining Algorithm 56
5.5 Clinical Performance Measures Represented as Temporal Association Rules 57
5.6 Identifying Clinically Relevant Results using Interestingness Measures 59
5.7 Related Works 60
5.8 Ethical Considerations 61
5.9 Findings 62
5.9.1 Order Set Discovery in ACSPD Using Sequential Pattern Mining 62
5.9.2 TARM in ACSPD 68
5.9.3 TAR Compliance with AHA/ACC STEMI Performance Measures and Outcomes 73
5.10 Verification of Findings 83
5.11 Conclusion 86
Chapter 6 Future Work 88
6.1 Dissertation Summary 88
6.2 Strengths, Limitations, and Assumptions 90
6.3 Implications for Clinical Practice 92
6.4 Future Work 94
6.5 Contributions 95
Appendix 96
References 97

List of Tables

Table 2.1 Sequential Pattern Mining Algorithm 12
Table 2.2 Variable Parameters for Simulated Data and Sequential Pattern Mining Algorithm 13
Table 2.3 Comparison of Known Patterns and Patterns Discovered by the SPM Algorithm 15
Table 2.4 Evaluation Metrics 15
Table 2.5 Number of Events and Sequential Pattern Mining Algorithm Performance 19
Table 2.6 Number of Patterns and Sequential Pattern Mining Algorithm Performance 19
Table 2.7 Pattern Probability and Sequential Pattern Mining Algorithm Performance 19
Table 2.8 Frequency Threshold and Sequential Pattern Mining Algorithm Performance 22
Table 2.9 Confidence Threshold and Sequential Pattern Mining Algorithm Performance 22
Table 2.10 Maximum Discovered Pattern Length and Sequential Pattern Mining Algorithm Performance 22
Table 2.11 Events and SPM Discovered Patterns, Embedded Known Patterns, and Minimum Edit Distance 24
Table 2.12 Patterns and SPM Discovered Patterns, Embedded Known Patterns, and Minimum Edit Distance 25
Table 2.13 Pattern Probability and SPM Discovered Patterns, Embedded Known Patterns, and Minimum Edit Distance 25
Table 2.14 Frequency Threshold and SPM Discovered Patterns, Embedded Known Patterns, and Minimum Edit Distance 25
Table 2.15 Events and SPM Discovered Patterns, Embedded Known Patterns, and Minimum Edit Distance 25
Table 2.16 Maximum Pattern Length and SPM Discovered Patterns, Embedded Known Patterns, and Minimum Edit Distance 26
Table 2.17 Comparison of Dynamic and Static Constraints Using Identical Simulation and User-Specified Parameters 28
Table 3.1 Temporal Association Rule Mining Algorithm 34
Table 4.1 Interestingness Measure Attributes 37
Table 4.2 Antecedent and Consequent Contingency Table 38
Table 4.3 Temporal Association Rule Interestingness Measures 41
Table 4.4 Interestingness Measures and Confirmatory Measure Properties 42
Table 4.5 AHA/ACC STEMI Performance Measures Published in 2006 and 2008 48
Table 4.6 Rule Representation of AHA STEMI Performance Measures 49
Table 5.1 Event Type and Time Granularity in the ACSPD 52
Table 5.2 Order Sets by Time from Admission 64
Table 5.3 Order Set ID and Constituent Events 65
Table 5.4 Summary of Interestingness Measures for Practice Patterns Discovered Using TARM 68
Table 5.5 Top 5 Clinical Practice Patterns Ranked by Interestingness Measure 69
Table 5.6 AHA/ACC STEMI Performance Measures 75
Table 5.7 Risk of Mortality for Noncompliance with the AHA/ACC Performance Measures 80
Table 5.8 TAR Compliance with AHA/ACC STEMI Performance Measures and Mortality 82
Table 5.9 TAR Compliance with AHA/ACC STEMI Performance Measures and Bleed Event 82
Table 5.10 Encounter Compliance with Performance Measures and Outcomes 84

List of Figures

Figure 2.1 Sequential Pattern Mining Algorithm....9
Figure 2.2 Removing Redundant Subpatterns from Discovered Patterns in the SPM Algorithm....10
Figure 3.1 Types of Temporal Association Rules....32
Figure 3.2 Temporal Association Rule Mining Algorithm....35
Figure 5.1 Study Design....55
Figure 5.2 Study Data....55
Figure 5.3 Number of Order Sets by Frequency Using Dynamic Confidence Thresholds in the ACSPD....66
Figure 5.4 Number of Order Sets by Frequency Using Static Confidence Threshold in the ACSPD....66
Figure 5.5 ACSPD Discovered Order Set Frequency by Year....67
Figure 5.6 Discovered Order Sets by Appearance and Duration in the ACSPD....67
Figure 5.7 Performance Measure Compliance in STEMI Encounters....83
Figure 5.8 Mortality and Performance Measure Compliance in STEMI Encounters....85
Figure 5.9 Bleeding Events and Performance Measure Compliance in STEMI Encounters....85

Chapter 1 Temporal Data Mining and Acute Coronary Syndrome

Clinical medicine is poised to make significant advances as a result of the wide spread implementation of electronic medical records (EMRs). The confluence of evidence-based medicine and "big data" promises deeper insight into patient treatment and optimal outcomes. The promises of EMR data are improvements in the areas of public health, adverse events, personalized care and more. The confluence of information technology, EMRs, natural language processing and data mining will forge the tools necessary to fulfill those promises. And the integration of disparate data types will forge the techniques necessary to fulfill those promises. As EMR use increases in the US (especially with the passage of the HITECH Act), there is an opportunity to analyze EMR patient data to better understand the correlations between practice choices and outcomes.

In this chapter, I introduce acute coronary syndrome and temporal data mining, and discuss the purpose and the contributions of this dissertation.

1.1 Acute Coronary Syndrome

Every 25 seconds someone in the US has a cardiac event and one person per minute will die from it. (Roger, Go et al.) Each year approximately 4.5% of the US population will have a myocardial infarction (MI) (CDC, 2011) resulting in over a million hospitalizations(Writing Group Members, Rosamond et al. 2008) and over 170,000 deaths (CDC, 2011). An MI is caused by damage to the cardiac muscle of the heart as a result of ischemia.(Torpy, Burke et al. 2010) Ischemia is insufficient blood flow because the arteries that supply the muscle are blocked by thrombus formation.(Davies 2000, Hamm, Bertrand et al. , Libby) Plaques are lipid deposits in the arteries. Plaque disruption triggers a platelet response. The aggregation of platelets can form a thrombus that occludes the blood vessel. (Epstein, Rosenberg et al. , Dahlbäck) Formation of lipid deposits is caused by a combination of diet, lifestyle and genetic factors.

Not all MI's are the same. There are two main types, ST-elevated (STEMI) and non-ST elevated (NSTEMI) myocardial infarction. The ST refers to the ST segment of the electrocardiogram (ECG) trace. The heart beat is the contraction of cardiac muscle as a result of electrical stimulation controlled by cardiac nerves. This pulse is measured using an ECG. When there are changes to the electrochemical properties of the heart it corresponds to distinct changes in the ECG trace. The ECG trace has features

labeled P through U. Elevation of the segment identified as ST is the distinguishing characteristic of the types of myocardial infarctions. This elevation is consistent with more severe damage to the cardiac tissue. The more severe damage causes increased morbidity and mortality in STEMI patients compared to NSTEMI patients.

STEMI, NSTEMI and a third cardiac event, unstable angina (UA), are all caused by acute myocardial ischemia and are collectively referred to as acute coronary syndrome (ACS).(Torpy, Burke et al. 2010) STEMI accounts for 25%-40% of all ACS cases.(Yeh, Sidney et al. , Mehta, Parsons et al.) There was significant progress in the last decade to decrease the STEMI mortality rate. (Fox, Steg et al. , Yeh, Sidney et al.) The timing of treatment for STEMI, specifically minimization of the amount of time from the infraction to hospitalization and treatment is central to optimal patient outcomes.(Braunwald, Antman et al. , Cheitlin, Armstrong et al. , Antman, Anbe et al. , Eagle, Guyton et al. , Smith Jr, Feldman et al. , Anderson, Adams et al. , Antman, Hand et al. , King III, Smith Jr et al. , Kushner, Hand et al. 2009) The mortality rate of a STEMI patient increases for every 30 minute delay in treatment.(Krumholz, Merrill et al.) Therefore the American College of Cardiologists (ACC) and the American Medical Association (AHA), in the treatment guidelines for ACS emphasize the importance of timing of events.

1.2 Temporality in Data Mining

Data mining is the application of artificial intelligence and/or statistical methods to large amounts of data. Temporal data mining then is the inclusion of temporality in the data mining process.(Roddick and Spiliopoulou 2002) Lin et al. defines temporal data mining as part of a knowledge discovery process that "enumerates structures (temporal patterns or models) over the temporal data".(Lin, Orgun et al. 2002) Temporal data mining goes beyond the analysis of static events, to uncover hidden knowledge that without the inclusion of temporality would not be known.(Chen and Petrounias 1998) It can elucidate the importance of the arrangement of events and/or the relationship between events.

Although conceptually facile, the inclusion of temporality in data mining is challenging. Currently there are relatively few algorithms specifically designed for temporal data mining. Researchers utilize methods such as time-series, longitudinal analysis and state change. For events of the same type, and differing only in time, time series is the approach often used. Time series usually requires periodicity in at least

one dimension. EMR data are asynchronous in events and timing, it has many different events and those events may or may not display periodicity. A longitudinal approach is best suited for understanding which features or characteristics influence the occurrence of an outcome. Longitudinal models are useful in classifying risk, such as having an MI. A research focus on a change in a patient's condition from critical to serious or a patient is transferred from ICU to inpatient lends itself to state change models based on the Markov assumption. Alternatively researchers can utilize existing data mining algorithms, treating temporal data like all other data. Each of these approaches has its own assumptions and limitations. They are generally beneficial for prediction of when or if an event will happen but not for discovering the *relationship* between events. Sequential pattern mining and temporal association rule mining do just that. They explore the temporal relationship between events. STEMI treatment is highly time-dependent, and understanding clinical practice and its effect on patients necessitates incorporating temporality in data mining. The focus of this research is the temporal relationship between events in STEMI treatment. To the best of my knowledge, sequential pattern mining and temporal association rule mining have not been used together in EMR data to explore STEMI treatment.

1.3 Research Statement

The ultimate goal of this research is to uncover time-based relationships between events, i.e. clinical practice patterns that are novel and beneficial to achieving optimal patient outcomes. In order to accomplish this I have done the following:

1. Create a sequential pattern mining algorithm (SPM) that incorporates user-defined parameters for adaptation to specific applications and removes redundancy in the discovered patterns.
2. Use simulation to establish SPM performance in various scenarios. The SPM is well-suited to discover patterns in noisy data.
3. Create a temporal association rule data mining algorithm (TARM) that is able to discover temporal association rules that can mine a large number of sequences, does not use candidates, is able to process intervals and can include gaps of arbitrary length.
4. Mine EMR data using SPM and TARM to find order sets (sequential patterns) and time-based relationships between events (temporal association rules).
5. Assess patient outcomes as they relate to clinical practice patterns (temporal association rules) and evidence-based knowledge (STEMI treatment performance measures).

1.4 Contributions

The primary contribution of this work is in exploring clinical practice patterns and their effect on patient outcomes. The focus of this dissertation is an exploratory analysis of the temporal relationship between events during STEMI treatment. Clinical practice patterns are the representation of the temporal

relationships, i.e. the temporal association rules that describe STEMI treatment. This work is the first to mine clinical practice patterns and assess those practice patterns for compliance with treatment guidelines. Furthermore, to the best of my knowledge, this is the first work to analyze clinical practice pattern compliance and patient outcomes.

Novel methods for temporal data mining in big data are the second area of contribution. Gaining insights into clinical practice using EMR big data is challenging because of the large amount of data and often the data are incomplete. Existing methods are not designed to work with large incomplete data. Furthermore, existing methods are not designed for exploratory analysis of temporal relationships. Therefore I created two algorithms for temporal data mining in EMR big data, the sequential pattern mining algorithm (SPM) and the temporal association rule mining algorithm (TARM). To the best of my knowledge, no research exists on the combined used of sequential pattern mining and temporal association rule data mining in EMR data.

The third area of contribution is in using simulation. Simulation of data is a powerful technique that can inform researchers about the properties of an algorithm or system. In this dissertation, I use simulated data to evaluate the performance of the SPM algorithm. The purpose of the SPM is to find sequential patterns in data. However large amounts of labeled data for algorithm performance evaluation are not available. Therefore simulated data are generated to mimic real world encounter data. In EMR data, events do not have the same probability of occurrence, some events happen more often than others. Likewise sequential patterns, i.e. order sets, do not have the same probability. By varying a set of data simulation parameters, I was able to evaluate the performance of the SPM algorithm in different conditions. It is my hope that future work will unlock more insights into STEMI treatment and reduce the morbidity and mortality it causes.

Chapter 2 Sequential Pattern Mining

Sequential pattern mining is the search for important patterns in sequences of elements. It can be exploratory, such as finding patterns in DNA, or predictive, using known patterns to anticipate stock prices. Sequential pattern mining can be computationally complex when there are a large number of possible elements. Sequential pattern mining requires a strong ordering of elements. In this research, the ordering of events is governed by time. To decrease the complexity of the data mining task, the number of possible elements is constrained. Applying sequential pattern mining to clinical medicine is challenging because the ordering of events can be asynchronous and simultaneous. Furthermore there are a large number of possible events which include medications, lab tests, procedures and other activities that are part of patient care. The data from human decision making and actions in the clinical setting can cause uncertainty in the ordering of events and a large number of events. Thus the goal is to create a sequential pattern mining algorithm that can deal with the vagaries of human decisions with minimal impact to model accuracy.

In this chapter, I review the relevant literature on sequential pattern mining, introduce a novel sequential pattern mining algorithm and use simulation to evaluate its performance.

2.1 Mining Sequential Patterns, a Review of Related Works

The challenges of sequential pattern mining are related to effectiveness and efficiency. Effectiveness refers to the ability of an algorithm to find patterns that are interesting and/or useful to the user. (Giannella, Han et al. 2003) Efficiency addresses the computational complexity. A more efficient algorithm is able to find sequential patterns with less computational overhead. The evolution of sequential patterning mining algorithms is driven by improving effectiveness and efficiency.

There are two generally accepted categories of sequential pattern mining algorithms, apriori-like and pattern growth.(Mooney and Roddick 2013) Srikant and Agrawal (Srikant and Agrawal 1996) introduced the apriori-based method in their General Sequential Pattern (GSP) algorithm. In the apriori-based method, a database of transactions (sequences) is scanned for frequent items and candidate sequences are constructed based on the frequent items. The transaction database is scanned for the frequency of the candidate sequence. Srikant defined the frequency of a pattern as its support. If the support is above a

minimum level, the candidate frequency is then used to generate new candidate sequences and the process is repeated until there are no more candidates that exceed the minimum support.

The great contribution of the apriori approach is the fundamental lemma that a super pattern of a non-frequent item cannot be frequent. In other words, if a sequential pattern contains an infrequent item, then that pattern cannot be frequent. The implication for SPM algorithm development is to focus only on frequent items and ignore all other items. This reduces complexity and in conjunction with the support level may produce interesting results for the user. (Agrawal, Imieliński et al. 1993) GSP is more appropriate for small databases and short sequences. The practical limitations of GSP are that it requires multiple scans of the database, can produce a large number of candidate sequences, and has trouble with longer patterns. Although no single variation of the apriori algorithms addresses all of these limitations, there are many that target a single limitation.

The original GSP algorithm and its successors are unable to handle longer sequences. To search for longer patterns, the frequent-pattern (FP) tree algorithms were developed. (Han, Pei et al. 2000) The FP-tree based algorithms allow for longer patterns by creating an efficient structure for candidate tree generation and support storage. There are several variations of FP-tree algorithms. Generally speaking, the nodes are the elements and the branches are the relationships between elements. The tree is pruned when the support for a branch is below a user-defined threshold. FP-Tree algorithms are intended to run in memory and therefore are challenged by a large sequence database.

Apriori-based algorithms can generate a large number of candidate sequences and reducing the number of sequences would increase efficiency. Constraints are used to limit the number of sequences. These constraints can assume various forms. Garofalakis et al propose regular expression-like constraints (Garofalakis, Rastogi et al. 1999), whereas Zaki et al. use user-defined constraints (Zaki 2001). Apriori-based algorithms traditionally used the support (frequency) as a constraint threshold. The assumption is that more frequent patterns will be most interesting to users and the constraint threshold is highly dependent upon user-needs. This may or may not be the case depending on the application. Pei (Pei, Han et al. 2007) examined the effect of several user-defined constraints on mined patterns, including the confidence constraint. Confidence is the conditional probability of the final element in a pattern given the

previous elements. In Pei's work, the use of a confidence constraint instead of support shifts the importance from the frequency to the sequential relationship. In a strong pattern, the conditional probability of is high. A weaker pattern will have a lower confidence.

The incorporation of constraints other than support into sequential pattern mining algorithms paved the way for a different type of algorithm, the pattern growth algorithm. Han (Han, Pei et al. 2000) created FreeSpan, an algorithm that reduces the number of scans necessary by leveraging all two-item relationships. It then uses a database of projections of the 2-item relationships to limit the search space for longer sequential patterns. The advantage of Han's approach is a more efficient method of mining frequent patterns requiring fewer database scans. The disadvantage is that it is designed for short patterns. The pattern growth algorithm neither generates candidate sequences nor scans a database for their support. Instead, the pattern growth algorithm uses the sequences in the database to create frequent patterns as opposed to the apriori-based frequent items. The sequence database is scanned multiple times, but only those sequences that exist in the database are sought. These algorithms perform better than traditional apriori-based; however existing implementations run in memory.

Apriori-like and pattern growth SPM algorithms are primarily suited for exploratory research. When the mining task is beyond exploration and into the realms of classification and prediction, probabilistic approaches are more appropriate. A probabilistic approach to sequential pattern mining is used in applications in bioinformatics, clinical informatics, and finance. The GenBank sequence alignment is based on sequential pattern mining (Ewens and Grant 2001). Probabilistic approaches are most useful when the task is more than finding the frequent patterns but assigning classifications to those patterns or based on those patterns. Although this is an active area of research, it is outside the scope of the current research.

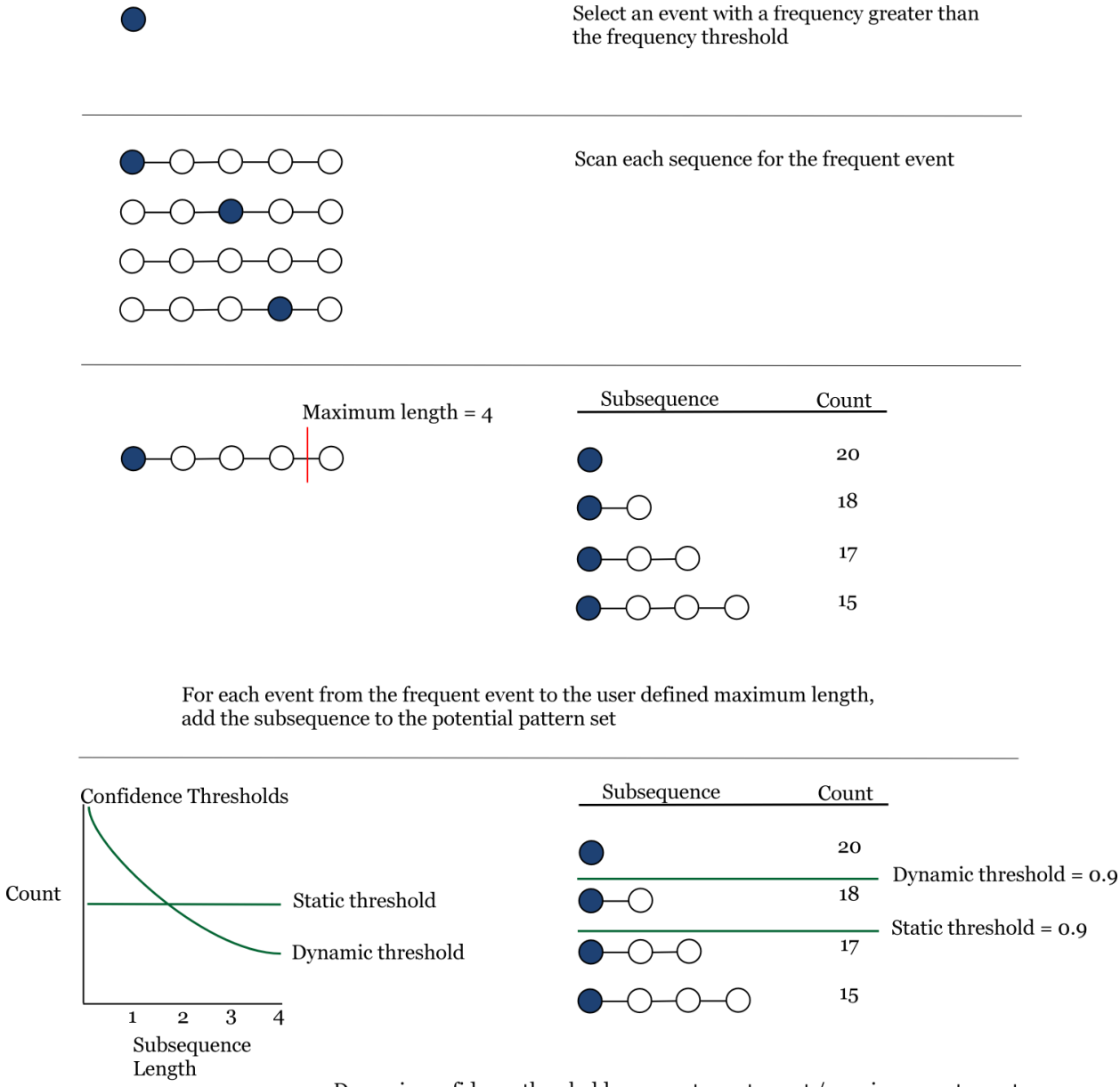
In this research, I focus on clinical medicine using data from electronic medical records (EMR). EMRs are a collection of events that occur during a patient encounter. These events can be ordered sequentially by time. The use of sequential pattern mining is relatively rare in clinical medicine. However recently, there is an increase in interest. (Reis, Kohane et al. 2009, Denny, Peterson et al. 2010, Patnaik, Butler et al. 2011) Patnaik et al, for example used sequential pattern mining to mine serial and parallel episodes

(patterns). Serial patterns are equivalent to sequential patterns and parallel episodes are groups of events that are significant but with no intrinsic order. Application of existing sequential pattern mining methods to clinical data presents several challenges. Clinical data has a large number of events, has considerable variation in sequences, the support assumption may not be adequate for the application and patterns are of unknown length. Furthermore, clinical data repositories can be very large, and therefore algorithms designed to run in memory are not feasible. Existing SPM algorithms are not designed to find patterns in EMR data. Therefore, I created a novel variation of SPM to find patterns in EMR data.

2.2 Sequential Pattern Mining Algorithm

The sequential pattern mining algorithm I developed for the task of mining frequent sequences in EMR event data combines the advances of previous research and includes novel approaches to improve effectiveness and efficiency. Specifically this algorithm 1) exploits the apriori assumption, 2) includes user-defined thresholds, 3) refines patterns by removing redundancy, and 4) retains only the longest unique patterns. These 4 attributes combine to create a novel and accurate sequential pattern mining algorithm. (See Figure 2.1 and Table 2.1). This sequential pattern mining algorithm exploits the apriori assumption that a pattern cannot be frequent unless the constituent events are frequent. Therefore it is important to first define and then identify all of the frequent events in the data. A frequent event is defined as an event with count above a user-defined threshold. The frequent events are the basis for pattern discovery and the assumption is that interesting patterns will occur more often. Therefore, using a frequent event as the start of the pattern influences the pattern effectiveness. To identify the frequent events the data are scanned once. The frequent event is the beginning of the pattern, but pattern growth is governed by another user-defined parameter, the maximum pattern length. Pattern growth is the addition of events to a sequential pattern until a termination condition is met. The maximum pattern length serves as the termination condition in this algorithm. The maximum pattern length parameter directly influences the effectiveness and efficiency. Should the user select a long maximum pattern length, the computational complexity of subsequent steps would increase. However, selecting a short maximum pattern length can exclude interesting patterns from the final results. Therefore the maximum length should be selected to increase effectiveness but not impede efficiency. To discover the patterns, each encounter (sequence of events) in the data is scanned again, to find the location of frequent events.

Figure 2.1 Sequential Pattern Mining Algorithm

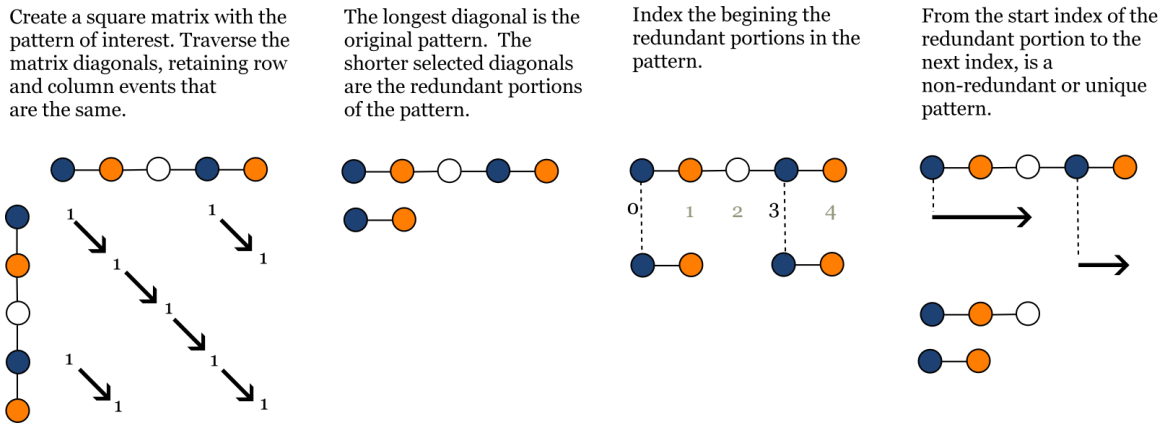


When a frequent event is found, the algorithm initiates pattern growth. Beginning with the frequent event, the algorithm will create a new pattern by continuing to add the next event until reaching the

maximum pattern length. The algorithm adds each pattern to potential pattern set storage. The in-memory temporary storage is used to track the potential patterns and pattern frequency. The only data that is required to be in memory, are the events of the specific encounter being processed and the temporary storage. This means that the algorithm can process very long sequences of event data to find potential patterns. By using this temporary in-memory storage, the data are only scanned once for all frequent events and potential patterns. Compared to previous approaches that require multiple scans of the data and retention of all the event data in-memory, this approach is more efficient.

The next 3 steps refine the potential patterns and boost algorithm effectiveness in EMR data. A conditional probability threshold is used to select interesting patterns, redundant portions of the patterns are removed and patterns that are subsequences of other patterns are eliminated. The user-defined confidence threshold is incorporated into the algorithm because the frequency threshold alone is inadequate. In other words, it's not only the number of occurrences of a pattern that makes it interesting but also the strength of the conditional relationship between events. An interesting pattern is one with a strong conditional relationship between the constituent events. The confidence is the conditional probability of the final event given the previous events. The algorithm is able to mine interesting patterns even in noisy data situations by altering the confidence level.

Figure 2.2 Removing Redundant Subpatterns from Discovered Patterns in the SPM Algorithm



When discovering sequential patterns with a maximum length parameter, it is possible to discover a pattern that is actually a shorter sub-pattern repeated in succession, such as the same order entered twice. For the purpose of this explanation, a sub-pattern is a pattern that is subsumed or a part of a larger

pattern. When a sub-pattern is repeated, the found pattern will have internal redundancy. For example, the pattern can be “ABCABC” or “ABC**ABC” where ** are any number of events. In the first example, the shorter redundant sub-pattern is “ABC”. The second pattern is more difficult, in that the pattern could have an arbitrary number of events between the redundant sub-patterns, “ABC**ABC”. Of interest in this implementation are the non-redundant as well as the redundant events in the discovered pattern, therefore “ABC**ABC” becomes “ABC**”, and “ABC”. To find redundant sub-patterns I used the diagonals of the sequence matrix (see Figure 2.2), the pattern is used to construct the rows and columns of a square matrix. If the row event and the column event are equivalent, the cell is assigned a value of 1, if the row and column event are different, the cell is 0. A redundant sub-pattern will have a diagonal of 1’s in succession. A diagonal of “1, 1” is an event entered twice and therefore not considered a redundant sub-pattern. A diagonal of 3 or more 1’s is considered a redundant sub-pattern. When a pattern is found to have shorter redundant sub-patterns, the location of the redundant sub-pattern is indexed in the original pattern. The indices are used to produce the non-redundant sequential patterns. A non-redundant sequential pattern begins with the first event of the redundant sub-pattern and ends with the last event prior to the next redundant sub-pattern. If the pattern with redundancy is “ABC**ABC” then the redundant section is “ABC” and the non-redundant patterns derived from the pattern with redundancy are “ABC**” and “ABC”. All non-redundant patterns are added to the potential order temporary in-memory storage. The inclusion of redundancy removal from the patterns has not, to the best of my knowledge, been included as part of a sequential pattern mining algorithm.

The final step in discovering sequential patterns from event data is to find the longest unique patterns by removing sub-patterns from the results set. In keeping with the examples used thus far, the pattern “ABC” is included in the pattern “ABC**”. Therefore, “ABC” is removed from the results and “ABC**” is retained as the longest unique sequence. After the sub-patterns are removed, the remaining patterns are returned as the results of the sequential pattern mining algorithm. This final step is another novel approach included in this sequential pattern mining algorithm that differentiates it from previous algorithms for pattern discovery.

Table 2.1 Sequential Pattern Mining Algorithm

Input: Database, D, of sequences, $s_i = e_o, e_1 \ldots e_n$ where e is an event and n denotes the temporal ordering.
Frequency threshold: the user-defined level for frequent events
Confidence constraint: user-defined level for conditional probability of events
Maximum pattern length: the maximum length of a pattern

Output: Set of interesting patterns

```
Create frequent event set
    For each event, count the occurrences
    If the frequency is ≥ user-defined frequency threshold add to frequent event set
Find sequential patterns
    For each sequence in the database
        For each event in the sequence
            For each frequent event in frequent event set
                If event = frequent event then return the event position, m
                    For each position in i = m to i = (m + maximum pattern length)
                        Add subsequence(m, i) to temporary pattern storage
                    Increase count of i by 1
    For each pattern in the temporary pattern storage
        If confidence ≥ user-defined confidence threshold then retain, else remove
        If a pattern has redundancy, create non-redundant patterns
        If the pattern is a subsequence of another pattern, then remove
  Return all interesting patterns
```

2.3 Sequential Pattern Mining Thresholds

Currently there is no consensus on how to select the frequent event or the confidence constraint thresholds. The frequency threshold and confidence constraint work in concert to limit the number of patterns identified and increasing the usefulness of the discovered patterns. Selecting a frequency threshold that is too low will cause the number of patterns identified to increase. Setting the frequency threshold too high potentially means that patterns will be missed resulting in false positives for interesting rules. A confidence constraint that is too lenient will lead to discovery of patterns with weak relationships whereas a confidence constraint that is too stringent can eliminate interesting results. The choice of threshold is left to the user in the SPM algorithm. I evaluated the algorithms performance using many different thresholds.

2.4 Simulating Event Sequence Data

Event sequence data varies in complexity and a sequential pattern mining algorithm needs to handle challenging data environments. At one end of the spectrum a sequence can have a small number of events that are repeated often. At the other end of the spectrum a sequence can have a large number of events that may or may not reoccur. Inherent in this is that the probability of any event occurring in a particular position in a sequence is not uniform for all events. Also the number and length of the interesting patterns is unknown. A sequential pattern mining algorithm should be able to find interesting patterns of various lengths. In order to evaluate the performance of the sequential pattern mining algorithm in different conditions, I simulated sequence data with known interesting patterns.

Table 2.2 Variable Parameters for Simulated Data and Sequential Pattern Mining Algorithm

Scope	Parameter	Explanation
Simulated data	Sequences	Number of sequences generated
	Maximum sequence length	Maximum number of events in a sequence
	Patterns	Number of patterns generated
	Pattern length	Number of events in a pattern
	Pattern proportion	Proportion of patterns in sequence
SPM Algorithm	Frequency threshold	Lower bound threshold for event frequency
	Confidence threshold	Lower bound threshold for discovered pattern confidence
	Maximum pattern length	Maximum number of events in a discovered pattern
	Confidence constraint type	Dynamic threshold or static threshold

To simulate sequence data the user sets 5 parameters; 1) the number of sequences to be generated, 2) the maximum length for the sequences, 3) the number of events in a pattern, 4) the number of patterns available to include in the sequences, and 5) the probability that each position in the sequence is an independent event or a pattern. (See Table 2.2) By adjusting these parameters, I can simulate different event data environments and measure how the SPM algorithm performs.

When creating simulated data, the first task is to make the set of events available to build a sequence. The user sets the number of events available and the data simulator randomly assigns each event a numeric identifier. The second task is to create the known patterns of interest. The events that make up the patterns are randomly selected using a Gaussian distribution over the set of available events. The Gaussian distribution for event selection ensures that the probability is not the same for all events. Similar to real data, some events will occur more often than others. Using the available events and patterns, the sequences are generated. The user sets the prevalence of patterns in the generated data. This means that for each position in the sequence, there is a fixed probability that it is a pattern or an independent event. Now suppose the user decides to set the probability of a pattern at 0.30. We start with an empty sequence. The first position has a 30% chance of being a pattern and a 70% chance of being an independent event. If a pattern is selected, then all of the events that constitute the pattern are inserted into the sequence. A Gaussian distribution of patterns ensures that some patterns will be selected more often than others. The algorithm continues to build the sequence until the length is reached. In real data, the event sequences are of different lengths. To mimic this, the length of the sequence set to a random number between 1 and the user-defined maximum sequence length. This process is repeated until the user-specified number of sequences is generated. For each sequence

generated, the count and location of each pattern is recorded. This information is used to calculate the algorithm performance metrics.

In real event data, sequences contain patterns of events and independent events. The probability of occurrence is not uniform for all independent events, some events will occur more often than others. Similarly, the occurrence of event patterns is not uniform for all patterns, some patterns occur more often than others. To further complicate matters, the events that comprise a pattern, "ABC", can occur independently of the pattern, "B", or as part of another pattern "ACD". Simulating sequences of event patterns and independent events requires defining the distribution of the probabilities, and for this application the Gaussian distribution is most appropriate. The system created to generate the simulated sequences relies heavily on Gaussian random number generators. The use of Gaussian random number generators for the events that comprise the patterns and for the selection of events or patterns to include in the sequence is to mimic real data. The Gaussian number generator also ensures that the distribution of events in the patterns is similar to the distribution of events in the surrounding the sequence. The similar distribution means that the algorithm is finding interesting patterns based on conditional relationships. However it could prove difficult to distinguish known patterns from spontaneously occurring ones that exceed the frequent event and confidence thresholds.

2.5 Evaluation Metrics: Precision, Recall and F-measure

The purpose of the sequential pattern mining algorithm is to discover sequential patterns in ordered data. To evaluate the performance of the sequential pattern mining algorithm I used precision, recall and F-measure. Precision and recall are measures which compare the results, in this case the discovered patterns with a gold standard, the known (generated) patterns. The precision is the number of known patterns that were included in the discovered patterns. The recall is the number of discovered patterns that are in the known patterns. The F-measure is the harmonic mean between the precision and recall. The harmonic mean can be weighted to capture the relative importance placed on the precision or recall. Of interest in this evaluation is discovering known patterns, therefore the recall is more important and I will use a weight of 2.

Table 2.3 Comparison of Known Patterns and Patterns Discovered by the SPM Algorithm

		Known patterns	
		Known pattern	Unknown pattern
Discovered patterns	Discovered pattern	*True positive*	*False positive*
	Undiscovered pattern	*False negative*	*True negative*

The comparison of the generated patterns and the discovered patterns can be thought of as a two-by-two table. (See table 2.3) The cells of the table are designated as the true positive, false positive, true negative and false negative. A true positive is a known pattern that is discovered by the SPM. A false positive is a discovered pattern that is not a known pattern. A false negative is a known pattern that is not discovered by the SPM algorithm. A true negative would be a pattern that is neither known nor discovered. In this context there will be no true negatives. Precision, recall and F-measure can be calculated from the true positives, false negatives, and false positive values. (See Table 2.4)

Table 2.4 Evaluation Metrics

Precision	True positive / (True positive + False positive)
Recall	True positive / (True positive + False negative)
F-measure	$((1 + \beta^2)$ (precision * recall)) / (β^2 * (precision + recall))

2.6 Edit Distance and Embedded Patterns

The edit distance is a method of measuring similarity. The edit distance measure used in this study is based on the Levenshtein distance. (Levenshtein 1965) The Levenshtein distance is a measure for the similarity between strings and is used in natural language processing and computer science applications. For this application I adjusted the granularity of the Levenshtein distance because the original was too fine for sequential patterns. Here I am not interested in the character differences between words, but the word difference between sentences. The patterns are groups of events, the names of those events are unique integers of varying lengths. It is the differences between patterns of events that are of interest, not the characters in the names of the events. Therefore I modified the Levenshtein distance to measure the similarity between patterns (words). In this version of the edit distance, all changes were assessed the same penalty. Therefore the resulting edit distance is a quantification of the number of changes between two patterns. An edit distance equal to one means that the discovered pattern and generated pattern differ by one event. It does not comment on the type of change, insertion, deletion, or substitution between two patterns. It also cannot indicate if the same generated pattern is part of multiple discovered

patterns. To capture this information quantification of the number of generated patterns that are embedded (contained) in the discovered patterns was also included.

2.7 Evaluation of the SPM Algorithm Using Various Simulated Data Parameters

To evaluate the performance of the sequential pattern mining algorithm under different data environments, I varied the data parameters while holding the user-defined parameters constant. The user-defined parameters were frequency threshold = 20, static confidence threshold = 0.9, and maximum pattern length = 50. For the sake of simplicity I also held the number of events in a pattern (n=15) and the number of sequences (n=500) constant. I varied the number of events (n = 25-100), the number of patterns (n=5-50), and the probability of pattern (n=0.1-0.4). An experiment requires a value for each parameter for the number of events, the number of patterns and the probability of a pattern, e.g. events = 25, patterns = 5, and the probability of a pattern = 0.1. For each experiment I used the constant parameters and selected values for the variable data parameters. Due to the use of random number generators to create the simulated data, each experiment was repeated 10 times to diminish the influence of outliers.

The number of events can refer to the distinct identity of an event, i.e. "A" or the frequency of events en masse, n= 100. In the context of simulated data parameters the number of events refers to the unique event types or names. For instance, if we have the events, "A" and "B", the number of events is two. The parameter governs the number of events, but not the frequency. The frequency is the result of random processes in data generation. When the number of events is low (events= 25), the number of known patterns discovered by the SPM algorithm is low (true positive: range = 0-5, mean ± SD = 0.5±0.81) (See Table 2.5). As the number of events increases, the ability of the SPM algorithm to discover known patterns remains relatively constant. Generally, the SPM algorithm discovers more patterns than there are known patterns regardless of the number of events. When we compare the mean number of false positives for events = 25 (mean ± SD 19.4±3.87) to all other categories of events, the number of discovered patterns in excess of the known patterns is slightly greater (all p-values ≥ 0.05, data not shown). The number of false negatives is the same regardless of the number of events in the simulated data (all p-values > 0.05, data not shown). Given the propensity of the SPM algorithm to fail to discover the known patterns, the precision is low. The recall is also low due to the number of discovered patterns

that are not known patterns. The F-measure, the weighted harmonic mean of the precision and recall was biased toward recall. The F-measure is also low.

The number of events in data can vary widely, from relatively few events with high frequencies to a large number of events with low frequencies. When the number of events is low, we would anticipate that the known patterns and the surrounding independent events would look very similar. There are, after all, a small number of events from which to choose. When the number of events is large we would anticipate that the events in the known patterns would be similarly diverse as the surrounding data. In both of these extremes it would be difficult to discover the known patterns. I designed the SPM algorithm to discover patterns in noisy data where the number of events could be low, high or moderate. I evaluated the performance of the SPM algorithm using a range of the number of events and found that regardless of the number of events, the SPM algorithm is successful in discovering patterns in data. Unfortunately the SPM algorithm is unable to discover known patterns in data. Its performance, as measured by precision and recall, is low. The poor performance seems consistent and unaffected by variation in the number of events in the underlying data, indicating that the performance of the SPM algorithm is governed by other factor(s) than the number of events.

Patterns, like events, in this context refer to the number of unique patterns and the frequency of those patterns is the result of random processes. When the number of patterns is low (n=5), the mean number of discovered known patterns (mean ± SD = 0.07±0.26) is also low. (See Table 2.6) When we compare the mean true positives when there are few patterns with the mean number of discovered known patterns when there are many patterns (n=50, mean ± SD = 0.98±1.11), we see that there are more true positives when the patterns are high (p-value ≤ 0.05). The false positives, like the true positives, when there are few known patterns is significantly less than then the number of known patterns is high (n= 5, mean ± SD = 16.34±3.8 vs. n=50, mean ± SD =19.25±3.65; p-value ≤ 0.05). In other words, when the patterns are few, the SPM algorithm is better at identifying those known patterns in the simulated data. Congruent with this finding is that as the known patterns increase, the false negatives or undiscovered known patterns also increases (all p-values ≤ 0.05, data not shown).

The probability of a pattern is akin to the concept of prevalence. In data, the known patterns will comprise a percentage of the dataset. That percentage is the prevalence. When we build a sequence in the simulated data, each position has the same probability of being either an independent event or a pattern. That probability is the probability of a pattern. We would anticipate that as the pattern probability increases that the number of true positives, known patterns that are discovered, would increase accordingly. There is a significant difference between the true positives when the pattern probability is low (n=0.10, mean ± SD = 0.43±0.75) and when the probability is moderately high (n=0.30, mean ± SD = 0.51±0.82, p-value ≤ 0.05) or high (n=0.40, mean ± SD =0.51±0.82, p-value ≤ 0.05). (See Table 2.7) Discovering patterns that are not known has a significant difference between the extreme pattern probabilities much like the true positives. But the false positives are greater for the lower probability compared to high pattern probability mean ± SD. The known patterns that were not discovered by the SPM algorithm remain constant regardless of the pattern probability (all p-value > 0.05). Given the significant difference in true positives between low and high pattern probability, it is not surprising that the precision follows the same dynamic (n=0.10, 0.02±0.04 vs. n=0.40, 0.03±0.05; p-value ≤ 0.05). Recall and F-measure also demonstrate the difference between the extreme probabilities. When the pattern probability is low, the precision is lower, the recall is lower, and the F-measure is lower. Simply, the SPM algorithm is discovering fewer patterns when there are fewer patterns to discover.

When we vary the parameters that characterize the simulated data, the number of events, the number of patterns and the pattern probability, we find that if there is a difference in performance it is between the extremely low and extremely high values. Otherwise the performance is constant regardless of the simulated data parameter. Therefore, except for extreme values, the performance of the SPM algorithm is driven by attributes other than the underlying data, such as the user-defined parameters.

Table 2.5 Number of Events and Sequential Pattern Mining Algorithm Performance

Events	True Positive		False Positive		False Negative		Precision		Recall		F-Measure	
	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD
25	0-5	0.50±0.81	9-32	19.4±3.87	2-38	18.62±9.40	0-0.31	0.03±0.04	0-0.33	0.02±0.05	0-0.14	0.01±0.02
50	0-5	0.49±0.84	9-33	18.63±3.89	2-36	18.62±9.43	0-0.29	0.03±0.05	0-0.33	0.02±0.05	0-0.14	0.01±0.02
75	0-8	0.49±0.83	6-31	18.53±3.88	2-39	18.63±9.43	0-0.42	0.03±0.05	0-0.33	0.02±0.05	0-0.19	0.01±0.02
100	0-6	0.48±0.81	6-31	17.99±3.76	2-37	18.60±9.31	0-0.33	0.03±0.05	0-0.33	0.02±0.04	0-0.16	0.01±0.02
125	0-5	0.49±0.81	6-32	17.90±3.93	3-39	18.56±9.39	0-0.29	0.03±0.05	0-0.25	0.02±0.04	0-0.14	0.01±0.02
150	0-4	0.42±0.73	6-29	17.54±3.86	2-35	16.88±8.80	0-0.27	0.02±0.05	0-0.33	0.02±0.05	0-0.16	0.01±0.02
175	0-5	0.52±0.83	7-31	17.74±3.69	2-38	18.63±9.47	0-0.28	0.03±0.05	0-0.33	0.02±0.04	0-0.12	0.02±0.02
200	0-4	0.48±0.77	8-30	17.52±3.90	2-36	18.66±9.39	0-0.29	0.03±0.05	0-0.33	0.02±0.04	0-0.14	0.01±0.02
225	0-4	0.45±0.81	8-30	17.69±3.72	2-36	18.64±9.31	0-0.29	0.03±0.05	0-0.33	0.02±0.04	0-0.13	0.01±0.02
250	0-6	0.45±0.79	6-31	17.78±3.85	2-36	18.66±9.35	0-0.33	0.03±0.05	0-0.33	0.02±0.05	0-0.13	0.01±0.02
275	0-4	0.48±0.75	8-31	17.35±4.14	2-36	18.60±9.44	0-0.29	0.03±0.05	0-0.33	0.02±0.04	0-0.11	0.02±0.02
300	0-4	0.52±0.82	7-31	17.54±3.83	2-37	18.59±9.43	0-0.23	0.03±0.05	0-0.33	0.02±0.04	0-0.11	0.02±0.02
325	0-4	0.47±0.75	7-31	18.02±4.34	2-36	18.63±9.43	0-0.30	0.03±0.05	0-0.33	0.02±0.04	0-0.15	0.01±0.02
350	0-7	0.46±0.80	8-28	17.64±3.88	2-36	18.56±9.42	0-0.39	0.03±0.05	0-0.33	0.02±0.04	0-0.17	0.01±0.02
375	0-5	0.55±0.80	8-31	17.46±3.90	2-37	18.54±9.38	0-0.38	0.03±0.05	0-0.33	0.03±0.05	0-0.16	0.02±0.03
400	0-5	0.53±0.80	6-29	17.33±3.87	3-37	18.57±9.48	0-0.27	0.03±0.05	0-0.17	0.02±0.03	0-0.12	0.02±0.02
425	0-5	0.52±0.82	7-31	17.58±4.07	2-37	18.58±9.46	0-0.33	0.03±0.05	0-0.33	0.02±0.04	0-0.17	0.02±0.03
450	0-6	0.52±0.88	8-30	17.45±3.84	2-38	18.50±9.27	0-0.35	0.03±0.05	0-0.33	0.02±0.05	0-0.15	0.02±0.03
475	0-5	0.44±0.77	8-29	17.60±3.50	2-37	18.58±9.36	0-0.33	0.03±0.05	0-0.33	0.02±0.04	0-0.14	0.01±0.02
500	0-5	0.46±0.76	7-31	17.64±3.86	2-39	18.59±9.39	0-0.31	0.03±0.05	0-0.33	0.02±0.05	0-0.14	0.01±0.02

Table 2.6 Number of Patterns and Sequential Pattern Mining Algorithm Performance

Patterns	True Positive		False Positive		False Negative		Precision		Recall		F-Measure	
	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD
5	0-1	0.07±0.26	6-29	16.34±3.80	2-4	3.22±0.53	0-0.11	0.00±0.02	0-0.33	0.02±0.08	0-0.10	0.00±0.02
10	0-1	0.11±0.31	7-31	17.23±3.86	5-9	7.16±0.74	0-0.11	0.01±0.02	0-0.17	0.01±0.04	0-0.08	0.01±0.02
15	0-3	0.17±0.42	6-31	17.67±4.02	8-13	10.79±0.90	0-0.27	0.01±0.03	0-0.25	0.02±0.04	0-0.16	0.01±0.02
20	0-3	0.26±0.51	6-31	17.49±3.91	10-18	14.21±1.09	0-0.3	0.02±0.04	0-0.23	0.02±0.04	0-0.15	0.01±0.02
25	0-4	0.37±0.62	7-33	17.89±3.98	13-21	17.46±1.34	0-0.33	0.02±0.04	0-0.24	0.02±0.03	0-0.17	0.01±0.02
30	0-4	0.56±0.80	6-31	17.92±3.75	16-25	20.58±1.52	0-0.29	0.03±0.05	0-0.19	0.03±0.04	0-0.14	0.02±0.03
35	0-5	0.71±0.90	8-31	18.16±3.81	17-28	23.66±1.68	0-0.28	0.04±0.05	0-0.23	0.03±0.04	0-0.15	0.02±0.03
40	0-6	0.79±0.92	8-31	18.59±3.71	20-33	26.70±1.85	0-0.38	0.04±0.05	0-0.23	0.03±0.03	0-0.16	0.02±0.03
45	0-5	0.87±0.99	8-32	18.94±3.85	22-36	29.65±1.90	0-0.33	0.05±0.06	0-0.17	0.03±0.03	0-0.14	0.02±0.03
50	0-8	0.98±1.11	9-31	19.25±3.65	24-39	32.54±1.99	0-0.42	0.05±0.06	0-0.23	0.03±0.03	0-0.19	0.02±0.03

Table 2.7 Pattern Probability and Sequential Pattern Mining Algorithm Performance

Pattern Probability	True Positive		False Positive		False Negative		Precision		Recall		F-Measure	
	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD
0.10	0-5	0.43±0.75	6-31	18.16±4.06	2-38	18.31±9.25	0-0.33	0.02±0.04	0-0.33	0.02±0.04	0-0.17	0.01±0.02
0.20	0-6	0.49±0.79	6-33	17.95±3.92	2-37	18.49±9.36	0-0.29	0.03±0.05	0-0.33	0.02±0.04	0-0.16	0.01±0.02
0.30	0-6	0.51±0.82	6-32	17.78±3.83	2-39	18.57±9.42	0-0.38	0.03±0.05	0-0.33	0.02±0.04	0-0.16	0.02±0.02
0.40	0-8	0.51±0.84	6-32	17.86±3.85	2-39	18.56±9.41	0-0.42	0.03±0.05	0-0.33	0.03±0.05	0-0.19	0.02±0.03

2.8 Evaluation of the SPM Algorithm Using Various User-defined Parameters

To evaluate the performance of the SPM algorithm under different user-defined constraints I varied the frequency threshold, confidence threshold, and maximum pattern length while the data parameters remained constant. The data parameters were events (n = 100), number of patterns (n = 15), pattern length (n = 15), the number of sequences (n=500), the maximum sequence length (n = 100), and the pattern probability (n = 0.2). The varied user-defined parameters were the frequency threshold (n = 10 – 100), the confidence threshold (n= 0.6 – 1.0) and the maximum pattern length (n = 10 – 80).

The frequency threshold is the lower bound of the frequent events. The events with a frequency greater than the user-defined threshold become the first event in the potential patterns. Using a low frequency threshold allows for more potential patterns to be evaluated by the algorithm, but doesn't necessarily ensure increased accuracy. As the frequency threshold increases from n = 10 to n = 100, the known patterns that are discovered by the SPM algorithm peaks at a threshold of n=20, and decreases from n= 30 to n=100. (See Table 2.8) Discovered patterns that are unknown is maximum at the lowest frequency threshold (n=10, mean ± SD = 34.65±30.34), decreases from there (n=20, 8.14±7.28) and is constant from n= 30 to n = 100 and false positives have 3 groupings, n = 10, n=20, and n = 30 to n=100. The known patterns that were not discovered by the SPM algorithm, except for the lowest frequency threshold, increases as the frequency threshold increases. The precision is lowest for the lowest frequency threshold (n = 10, 0.12±0.08). It increases as the frequency threshold increases and peaks at n = 40 (0.64±0.23), and decreases to n = 100 (0.24±0.19). The recall is lowest for the highest frequency threshold (n = 100, 0.13±0.1) and highest at n = 20 (0.48±0.14). The F-measure follows the lead of the precision, with a peak at n = 40 (0.32±0.13) and decreases to n = 100. SPM performance is a balance between maximizing the number of known patterns discovered and minimizing the number of unknown patterns discovered. This balance is better when the frequency threshold is moderately low (n = 40). As the frequency threshold increases, the performance of the algorithm diminishes.

The confidence threshold governs the strength of the conditional relationship between the last event in pattern and the preceding events. A low confidence threshold means that the SPM algorithm will evaluate potential patterns with weak conditional relationships whereas a high confidence threshold will only include potential patterns with strong relationships. For this evaluation of the impact of confidence

threshold on algorithm performance, I used a static threshold. (A more thorough discussion of static and dynamic confidence thresholds in included later in this chapter.) Generally, as the confidence threshold increases from n = 0.6 to n = 1.0 the discovery of known patterns decreases and the discovery of unknown patterns increases. (See Table 2.9)

The maximum pattern length is a user-defined parameter that sets the longest discovered pattern. (See Table 2.10) The true positives for the shortest and longest maximum pattern lengths are significantly different from each other and the rest of the parameter values in between (maximum pattern length, ml = 10 vs. ml = 80, p-value ≤ 0.05; ml = 10 vs. ml = 20-70 all p-values ≤ 0.05; ml = 80 vs. ml 20-70, p-value ≤ 0.05). For unknown patterns, the shortest length is the highest and it decreases as the maximum pattern length increases. However, there is no significant difference in the false positives from n = 40 to n = 80 (p-values > 0.05). There is also no significant difference in the false negatives from maximum pattern lengths n = 30 to n= 70. The false negatives are highest when the maximum pattern length is shortest. There is a difference between the false negatives of the longest maximum pattern length and the other lengths (all p-values ≤ 0.05). The precision, recall and F-measure show the same significant differences for the shortest and longest maximum pattern length. (See Table 2.10) The maximum pattern length between the extremes is not significantly different (all p-values > 0.05). The maximum pattern length of the discovered pattern directly effects algorithm performance. When the user selects a maximum pattern length that is short, the SPM performance is poor.

The user-defined parameters of the SPM algorithm allow the user to control pattern discovery. As the frequency and confidence thresholds increase the performance of the SPM algorithm diminishes. The maximum pattern length affects algorithm performance at the extremes, but plays a minor role for moderate lengths. The performance of the algorithm is driven by user-defined parameter selection. The frequency and confidence threshold seem to be the major determining factors in algorithm performance but the user should use care in selection of all parameters.

Table 2.8 Frequency Threshold and Sequential Pattern Mining Algorithm Performance

Frequency Threshold	True Positive		False Positive		False Negative		Precision		Recall		F-Measure	
	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD
10	0-7	3.44±1.49	7-210	34.65±30.34	3-13	7.46±1.66	0-0.44	0.12±0.08	0-0.7	0.32±0.14	0-0.32	0.11±0.06
20	0-8	5.21±1.47	0-43	8.14±7.28	2-11	5.71±1.69	0-1	0.47±0.23	0-0.78	0.48±0.14	0-0.51	0.29±0.11
30	0-7	5.17±1.85	0-35	5.55±8.45	2-13	5.80±2.05	0-1	0.63±0.26	0-0.78	0.47±0.17	0-0.51	0.33±0.13
40	0-7	4.62±1.90	0-38	5.71±9.84	3-12	6.37±2.11	0-1	0.64±0.27	0-0.70	0.42±0.18	0-0.51	0.32±0.13
50	0-6	4.09±1.85	0-38	5.83±9.77	3-13	6.85±2.06	0-1	0.60±0.27	0-0.67	0.38±0.17	0-0.47	0.29±0.13
60	0-6	3.46±1.71	0-38	5.83±8.82	4-13	7.50±1.90	0-1	0.53±0.26	0-0.60	0.32±0.16	0-0.47	0.25±0.12
70	0-5	2.88±1.53	0-37	5.92±7.78	4-13	8.05±1.75	0-1	0.45±0.24	0-0.56	0.27±0.14	0-0.42	0.21±0.11
80	0-6	2.35±1.33	0-32	5.89±6.71	4-13	8.57±1.60	0-1	0.39±0.22	0-0.56	0.22±0.12	0-0.42	0.17±0.10
90	0-5	1.82±1.19	0-30	5.72±5.67	5-13	9.14±1.43	0-1	0.32±0.21	0-0.50	0.17±0.11	0-0.39	0.13±0.09
100	0-4	1.40±1.06	1-30	5.71±4.81	6-13	9.52±1.38	0-0.8	0.24±0.19	0-0.36	0.13±0.10	0-0.31	0.10±0.08

Table 2.9 Confidence Threshold and Sequential Pattern Mining Algorithm Performance

Confidence Threshold	True Positive		False Positive		False Negative		Precision		Recall		F-Measure	
	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD
0.6	0-7	3.84±2.06	0-210	14.14±30.63	3-13	7.11±2.19	0-1	0.49±0.29	0-0.70	0.35±0.19	0-0.51	0.25±0.14
0.65	0-7	3.95±1.86	0-94	9.50±13.79	2-13	7.01±2.02	0-1	0.48±0.28	0-0.78	0.36±0.17	0-0.51	0.25±0.13
0.7	0-7	3.76±1.91	0-99	9.51±14.48	3-13	7.20±2.06	0-1	0.47±0.28	0-0.70	0.34±0.18	0-0.51	0.24±0.13
0.75	0-7	3.63±1.90	0-85	8.55±11.68	3-13	7.30±2.09	0-1	0.46±0.28	0-0.70	0.33±0.18	0-0.51	0.23±0.13
0.8	0-7	3.47±1.92	0-91	8.39±11.7	3-13	7.44±2.15	0-1	0.45±0.27	0-0.70	0.32±0.18	0-0.49	0.22±0.13
0.85	0-8	3.29±1.99	0-63	7.77±9.78	3-13	7.67±2.13	0-1	0.42±0.27	0-0.70	0.30±0.18	0-0.51	0.21±0.13
0.9	0-7	3.20±2.02	0-80	7.84±10.24	3-13	7.74±2.17	0-1	0.42±0.28	0-0.70	0.29±0.19	0-0.51	0.21±0.13
0.95	0-7	3.00±2.07	0-58	7.15±8.57	2-13	7.91±2.30	0-1	0.39±0.28	0-0.78	0.28±0.19	0-0.51	0.20±0.14
1	0-7	2.86±2.12	0-52	7.20±8.62	2-13	8.09±2.29	0-1	0.38±0.28	0-0.78	0.26±0.20	0-0.51	0.18±0.14

Table 2.10 Maximum Discovered Pattern Length and Sequential Pattern Mining Algorithm Performance

Max Length	True Positive		False Positive		False Negative		Precision		Recall		F-Measure	
	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD
10	0-7	0.88±1.72	8-210	30.43±21.59	3-13	10.09±1.91	0-0.26	0.02±0.04	0-0.70	0.08±0.16	0-0.22	0.02±0.04
20	0-7	3.92±1.73	0-177	8.18±19.69	3-13	7.04±1.96	0-1	0.50±0.25	0-0.70	0.36±0.16	0-0.51	0.25±0.12
30	0-8	3.91±1.74	0-115	6.08±12.03	2-12	7.01±1.91	0-1	0.51±0.25	0-0.78	0.36±0.16	0-0.51	0.26±0.12
40	0-7	3.96±1.76	0-63	5.40±7.61	2-13	6.95±1.97	0-1	0.52±0.25	0-0.78	0.37±0.17	0-0.51	0.26±0.12
50	0-7	3.91±1.69	0-40	5.21±6.71	2-12	7.00±1.84	0-1	0.51±0.24	0-0.78	0.36±0.16	0-0.51	0.26±0.11
60	0-7	3.94±1.69	0-44	5.28±6.70	2-13	6.97±1.88	0-1	0.51±0.24	0-0.78	0.36±0.16	0-0.51	0.26±0.12
70	0-7	3.76±1.77	0-39	5.33±6.78	2-13	7.22±1.97	0-1	0.49±0.25	0-0.78	0.34±0.17	0-0.51	0.24±0.12
80	0-7	3.26±1.91	0-44	5.26±6.69	2-13	7.70±2.07	0-1	0.44±0.26	0-0.78	0.30±0.18	0-0.51	0.21±0.13

2.9 Embedded Known Patterns and Edit Distance

The SPM algorithm performs poorly in discovering known patterns. One reason could be that the known patterns are part of the discovered patterns but due to the noisy data other events are included. To test this I used two measures, the embedded patterns and the minimum edit distance. The embedded known patterns are known patterns that are contained within a discovered pattern. The number of embedded patterns is the number of known patterns that are contained in at least one discovered pattern. As the number of events increases, the mean embedded patterns remains constant. (See Table 2.11) Conversely, as the number of patterns or the pattern probability increases, generally the number of embedded patterns increases. (See Table 2.12 and Table 2.13) However the results for pattern probability are not statistically significant. If we compared the number of discovered patterns with the mean number of embedded patterns, the embedded patterns are greater than the discovered patterns which seem to indicate that the discovered patterns contain more than one known pattern.

The effect of frequency and confidence threshold on the embedded patterns is similar. The number of embedded patterns decreases as the value of the parameter increases (all p-values ≤ 0.05). The user-defined parameter directly affects SPM algorithm performance. The maximum discovered pattern length has no association with the embedded patterns (all p-values ≥ 0.05). Selecting more stringent frequency and confidence thresholds has a detrimental effect of the number of embedded patterns. This could be a result of a decrease in discovered patterns as the thresholds increase.

The edit distance is the number of events in the discovered patterns that need to change to be equivalent to the known pattern. I calculated the edit distance between the discovered patterns and each known pattern, and retained the minimum. The minimum edit distance is a measure of the similarity. Here I report the mean and standard deviation for the minimum edit distance. As the number of events increases, the difference between the known pattern and the closest discovered pattern is 37 events except for when the number of patterns is low. This means that the difference between the discovered pattern and its most similar known pattern is 37 changes. It should be noted that the cost of all changes including insertion of additional events was weighted equally. The value of 37 reflects all changes, insertions, deletions, and mutations. The relatively high edit distance values could be because the found patterns were considerably longer than the known patterns. The number of known patterns that were discovered

without additional changes (i.e. insertions) was extremely low, 0.5%. The percentage of known patterns that were embedded in one or more of the discovered patterns was close to 100%. The SPM algorithm was able to discover the known patterns with a high degree of precision but not a high degree of accuracy in a noisy data environment despite changes in the simulated data parameters.

For the user-defined parameter of frequency threshold the edit distance between the discovered pattern and the most similar known pattern was between 1.88 and 5.21. (See Table 2.14) Therefore the most similar known pattern only differed from the discovered pattern by 2 -5 changes. The mean edit distance is unaffected by changes in the confidence threshold (all p-values > 0.05) However the range and variance are wide. The performance fluctuates from exact match to highly error prone. There is a trend of increasing minimum edit distance as the confidence threshold increases. The maximum length of the discovered pattern has a U-shaped minimum edit distance, where the edit distance is greatest at the extremes. By adjusting the user-defined parameters, I was able to increase the accuracy of the SPM algorithm without sacrificing precision. The SPM algorithm performance is governed more by the user choice of settings than the characteristics of the underlying data.

Table 2.11 Events and SPM Discovered Patterns, Embedded Known Patterns, and Minimum Edit Distance

Events	True Positive		Discovered		Embedded		Edit Distance	
	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD
25	0-5	0.50±0.81	9-33	19.90±3.82	2-14	8.15±2.79	11.87-42.21	34.67±4.98
50	0-5	0.49±0.84	9-33	19.12±3.77	2-13	8.03±2.76	11.57-43.11	35.58±4.67
75	0-8	0.49±0.83	8-31	19.01±3.74	2-15	8.07±2.81	14.53-45.36	36.29±4.32
100	0-6	0.48±0.81	8-31	18.47±3.64	2-14	8.05±2.73	14.00.-43.00	36.78±3.77
125	0-5	0.49±0.81	6-32	18.39±3.82	2-14	8.06±2.73	19.05-43.38	36.98±3.64
150	0-4	0.42±0.73	6-29	17.97±3.77	2-14	7.80±2.77	18.67-43.00	37.48±3.43
175	0-5	0.52±0.83	7-32	18.26±3.55	2-14	8.03±2.76	21.8-42.71	37.16±3.49
200	0-4	0.48±0.77	9-30	18.00±3.77	2-14	8.08±2.81	20.64-43.08	37.48±3.09
225	0-4	0.45±0.81	9-30	18.14±3.60	2-13	7.98±2.66	22.54-43.73	37.51±3.43
250	0-6	0.45±0.79	7-31	18.23±3.71	2-14	7.99±2.72	23.38-43.56	37.25±3.30
275	0-4	0.48±0.75	8-31	17.83±3.99	2-13	8.06±2.71	23.71-43.63	37.30±3.33
300	0-4	0.52±0.82	7-31	18.06±3.69	2-13	8.13±2.72	17.00-44.36	37.31±3.50
325	0-4	0.47±0.75	8-32	18.48±4.16	2-14	8.05±2.74	21.33-43.00	37.41±3.45
350	0-7	0.46±0.80	8-28	18.10±3.80	2-14	8.12±2.77	20.22-45.10	37.64±3.15
375	0-5	0.55±0.80	10-31	18.01±3.71	2-14	8.13±2.79	25.54-42.31	37.32±3.19
400	0-5	0.53±0.80	6-29	17.86±3.75	2-13	8.15±2.88	25.38-43.86	37.36±3.26
425	0-5	0.52±0.82	8-31	18.10±3.90	2-13	8.07±2.75	19.08-43.04	37.21±3.39
450	0-6	0.52±0.88	8-30	17.97±3.64	2-14	8.10±2.73	24.36-44.33	37.41±3.34
475	0-5	0.44±0.77	8-29	18.03±3.40	2-13	8.01±2.74	23.61-47.35	37.91±2.94
500	0-5	0.46±0.76	9-31	18.11±3.71	2-13	8.07±2.79	22.00-43.00	37.56±3.18

Table 2.12 Patterns and SPM Discovered Patterns, Embedded Known Patterns, and Minimum Edit Distance

Patterns	True Positive		Discovered		Embedded		Edit distance	
	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD
5	0-1	0.07±0.26	6-29	16.41±3.77	2-3	2.48±0.50	20.43-45.1	38.66±2.50
10	0-1	0.11±0.31	7-32	17.33±3.81	4-6	4.81±0.40	22.18-43.00	37.58±3.08
15	0-3	0.17±0.42	6-31	17.85±3.93	5-8	6.49±0.52	24.33-42.94	38.15±2.76
20	0-3	0.26±0.51	7-31	17.75±3.75	6-10	7.92±0.64	19.89-43.27	38.08±2.73
25	0-4	0.37±0.62	9-33	18.26±3.77	6-11	9.05±0.79	19.08-43.21	37.55±3.15
30	0-4	0.56±0.80	8-31	18.48±3.49	6-13	9.81±1.02	15.54-43.20	36.89±3.51
35	0-5	0.71±0.90	10-31	18.87±3.54	7-14	10.23±1.18	11.57-43.63	36.29±3.95
40	0-6	0.79±0.92	10-32	19.37±3.39	6-14	10.25±1.41	14.00-43.86	35.66±4.08
45	0-5	0.87±0.99	11-33	19.81±3.49	5-14	10.08±1.56	13.23-43.11	35.10±4.57
50	0-8	0.98±1.11	11-31	20.23±3.27	4-15	9.58±1.83	11.87-47.35	34.80±4.94

Table 2.13 Pattern Probability and SPM Discovered Patterns, Embedded Known Patterns, and Minimum Edit Distance

Pattern Probability	True Positive		Discovered		Embedded		Edit distance	
	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD
0.10	0-5	0.43±0.75	7-31	18.59±3.97	2-13	7.30±2.38	11.57-47.35	36.82±4.21
0.20	0-6	0.49±0.79	6-33	18.43±3.78	2-14	8.12±2.72	14.00-43.56	36.94±3.67
0.30	0-6	0.51±0.82	6-33	18.29±3.71	2-15	8.34±2.87	13.23-43.63	36.95±3.67
0.40	0-8	0.51±0.84	6-32	18.37±3.70	2-14	8.45±2.89	11.87-44.36	36.87±3.71

Table 2.14 Frequency Threshold and SPM Discovered Patterns, Embedded Known Patterns, and Minimum Edit Distance

Frequency Threshold	True Positives		Discovered		Embedded		Edit distance	
	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD
10	0-7	3.44±1.49	11-210	38.09±29.85	3-11	7.57±1.01	1.95-29.66	5.76±4.21
20	0-8	5.21±1.47	6-46	13.36±6.40	1-8	6.27±1.38	0-24.00	2.82±2.58
30	0-7	5.17±1.85	4-36	10.72±6.85	0-7	5.41±1.80	0-31.40	2.37±2.98
40	0-7	4.62±1.90	3-38	10.33±8.18	0-7	4.73±1.83	0-38.00	3.13±4.20
50	0-6	4.09±1.85	2-38	9.92±8.26	0-6	4.20±1.74	0-66.67	3.75±5.48
60	0-6	3.46±1.71	2-38	9.29±7.53	0-6	3.50±1.60	0-55.25	4.55±5.60
70	0-5	2.88±1.53	2-37	8.80±6.70	0-5	2.94±1.47	0-68.50	4.86±5.22
80	0-6	2.35±1.33	2-32	8.24±5.83	0-6	2.37±1.28	0-70.50	5.51±5.67
90	0-5	1.82±1.19	2-30	7.53±5.00	0-5	1.88±1.15	0-72.00	6.51±7.58
100	0-4	1.40±1.06	1-30	7.11±4.33	0-4	1.40±1.02	1.20-71.50	6.75±7.00

Table 2.15 Events and SPM Discovered Patterns, Embedded Known Patterns, and Minimum Edit Distance

Confidence Threshold	True Positives		Discovered		Embedded		Edit distance	
	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD
0.6	0-7	3.84±2.06	3-210	17.99±29.49	0-11	4.86±2.28	0-44.67	44.67±4.77
0.65	0-7	3.95±1.86	2-100	13.45±13.14	0-10	4.67±2.26	0-52.67	52.67±3.91
0.7	0-7	3.76±1.91	2-104	13.27±13.94	0-9	4.41±2.25	0-72.00	72.00±4.28
0.75	0-7	3.63±1.90	2-90	12.17±11.11	0-9	4.24±2.30	0-68.50	68.50±4.16
0.8	0-7	3.47±1.92	2-96	11.85±11.24	0-9	4.00±2.30	0-67.50	67.50±4.23
0.85	0-8	3.29±1.99	2-68	11.06±9.30	0-9	3.80±2.34	0-69.00	69.00±4.94
0.9	0-7	3.20±2.02	2-86	11.04±9.87	0-9	3.62±2.36	0-49.33	49.33±4.63
0.95	0-7	3.00±2.07	2-62	10.15±8.18	0-8	3.41±2.40	0-71.50	71.50±5.09
1	0-7	2.86±2.12	1-57	10.06±8.33	0-8	3.24±2.45	0-68.00	68.00±5.38

Table 2.16 Maximum Pattern Length and SPM Discovered Patterns, Embedded Known Patterns, and Minimum Edit Distance

Maximum Length	True positives		Discovered		Embedded		Edit distance	
	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD	Range	Mean±SD
10	0-7	0.88±1.72	8-210	31.31±22.19	0-11	1.01±1.99	0-11	1.01±1.99
20	0-7	3.92±1.73	3-177	12.10±19.3	0-9	4.46±2.10	0-9	4.46±2.10
30	0-8	3.91±1.74	3-115	10.00±11.62	0-9	4.47±2.11	0-9	4.47±2.11
40	0-7	3.96±1.76	3-63	9.36±7.30	0-10	4.46±2.14	0-10	4.46±2.14
50	0-7	3.91±1.69	3-42	9.12±6.48	0-9	4.46±2.10	0-9	4.46±2.10
60	0-7	3.94±1.69	4-46	9.23±6.50	0-9	4.46±2.09	0-9	4.46±2.09
70	0-7	3.76±1.77	2-42	9.08±6.68	0-9	4.45±2.12	0-9	4.45±2.12
80	0-7	3.26±1.91	1-47	8.52±6.79	0-9	4.46±2.13	0-9	4.46±2.13

2.10 Static vs. Dynamic Confidence Thresholds

The SPM algorithm can use two types of confidence thresholds 1) the dynamic constraint and 2) the static constraint. The dynamic constraint changes as the length of the potential discovered potential increases. The static constraint is constant as the length of the potential pattern increases. The type of constraint influences algorithm performance, but the choice of constraint is left to the user. To compare the static and dynamic constraints I used the same user-defined parameters and data parameters for both. (See Table 2.17) In my research, the static constraint is more efficient than the dynamic constraint. Efficiency, in this context refers to the number of patterns discovered compared to the number of known patterns. Efficiency is not a strict measure but a relative comparison. The static constraint type produces fewer discovered patterns compared to the dynamic. The static constraint is also more likely to discover the generated pattern without additional events. Under the comparison parameters, the static constraint correctly discovered 1 pattern out of 150 (15 patterns per iteration x 10 iterations) and the dynamic failed to correctly discover any pattern. It should be noted that the percentage of true positives varies with the user-defined parameters (see Figure 2.1). Despite the inability to correctly discover the generated pattern, the mean edit distance for the dynamic constraint is lower than the static. The choice of a static or dynamic constraint becomes a trade-off between efficiency and effectiveness. I used a static constraint to evaluate the performance of the sequential pattern mining algorithm using various data and user-defined parameters.

2.11 Conclusion

The SPM algorithm is well-suited to extract patterns from noisy data. It exploits the apriori lemma to discover patterns, incorporates a static or dynamic confidence constraint, and removes redundancy. The

ability of the SPM to discover patterns is governed more by the selection of the user-defined parameters than the characteristics of the underlying data. This flexibility in the algorithm makes it useful for many applications with complicated data, such as healthcare and electronic medical records.

Table 2.17 Comparison of Dynamic and Static Constraints Using Identical Simulation and User-Specified Parameters

Constraint	Events (n)	Sequences (n)	Maximum sequence length (n)	Patterns (n)	Pattern length (n)	Pattern probability	Frequency threshold (n)	Confidence threshold (n)	Maximum pattern length (n)	Discovered patterns (Mean± SD)	True positives (Mean ± SD)	Embedded (percent)	Edit distance (Mean± SD)
Static	50	500	100	15	15	0.2	20	0.9	50	18.30 ± 2.06	0.10 ± 0.32	100	37.50±3.00
Dynamic	50	500	100	15	15	0.2	20	0.9	50	244.30 ± 30.71	0.0±0.0	100	28.6±0.88
p-value										<0.001	0.34	na	<0.001

Chapter 3 Temporal Association Rule Mining

In the treatment of STEMI, the timing of events is critical for achieving optimal patient outcomes. Understanding the associations between events is important in providing the best care for a patient. The purpose of this research is to discover temporal relationships between events.

In this chapter, I review the relevant works on association rule and temporal association rule data mining, and I present my temporal association rule algorithm.

3.1 Association Rule Data Mining

The progenitor of temporal association rule data mining (TARM) is association rule data mining (ARM) and to understand the abilities and limitations of TARM, it is necessary to review its roots. ARM was first introduced as a way of exploring consumer purchasing behavior. Since its inception, it has been used in a wide variety of fields from web log analysis, text mining, and atmospheric science. It is used to make recommendations, classification, and find relationships in structured data and natural language. ARM is popular for many reasons, most notably its ease of use and potential usefulness. It is easy for end users to interpret, even if users are not experts in the data mining process. The representation of the rule is easy for researchers and programmers to express. The ability to discover associations previously unknown to users also makes association rule data mining a compelling approach to data exploration. (Zhang and Wu 2011) Despite its considerable attributes, association rule data mining is challenging to implement. It suffers from computational complexity, memory limitations, and difficulty mining subtle or complex rules.

Throughout the history of knowledge discovery from data there are pivotal innovations in associations, particularly in the area of statistics, i.e. χ^2, odds ratio, or binomial test. Although adapted to exploratory analyses, the purpose of most of these statistical methods is to test a hypothesis. Hypothesis testing assumes a specific relationship between two variables, i.e. a causal relationship and can fail to recognize weaker relationships. Weaker relationships exist between causal and independence. They are an implication of a relationship but not a conviction (causal). Weaker relationships are referred to as associations in this research.

An association is represented as a rule. A rule is composed of an antecedent and a consequent. The antecedent implies the consequent, but does not cause the consequent; thus the term association. The general form of a rule is

$$Antecedent \rightarrow Consequent \quad \textbf{Equation 3.1}$$

In theory, the number of events in the antecedent or the consequent is only limited by the length of the sequence of events from which they are derived. For practical computational purposes the number is constrained.

The first widely implemented method of mining associations was introduced by Agrawal and Srikant. (Agrawal, Imieliński et al. 1993, Agrawal and Srikant 1994) Agrawal and Srikant's method is a two phase process.

1. Find frequent items and form candidate rules using these items
2. Scan database for frequent item rules.

Recounting the history of innovation for association rule data mining and sequential pattern mining (See 2.1) is muddled because the genesis of much of the foundational ideas is derived from the same body of work. To be clear, they are distinct data mining tasks; sequential pattern mining is about *pattern* discovery and association rule data mining is for *rule* discovery. Agrawal and Srikant created the APRIORI algorithm for association rule mining and the GSP algorithm for sequential pattern mining. Both are based on the APRIORI heuristic. The APRIORI heuristic describes the anti-monotonic attribute of frequent item sets. An item set cannot be frequent unless all of constituent items are frequent. The definition of frequent is left to the user. Also both the APRIORI and GSP algorithms use the same two phase structure. Therefore many of the critiques of the APRIORI or the two phase implementation are applicable to both sequential pattern mining and association rule mining. Consequently the enhancements and innovations are applicable to both mining tasks. A deep bifurcation has not occurred and cross-pollination continues to be common but there are off-shoots of interesting research. The future direction of sequential pattern mining and association rule mining is the inclusion of probabilistic models (Chu, Tseng et al. 2008), sampling (Raissi and Poncelet 2007), and domain knowledge (Denny, Peterson et al. 2010, Hanauer and Ramakrishnan 2013) in mining task.

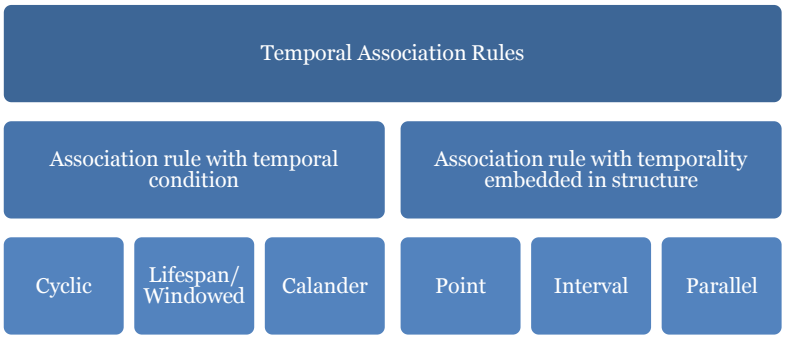
3.2 Temporal Association Rule Data Mining

Traditionally temporal association rule data mining explores time dependent data for rules that describe the occurrence of events in time. Temporal association rule mining grew out of the work of Agrawal and Srikant. They proposed a data mining approach that found co-occurring items in transactions with frequency (support) and/or conditional probability (confidence) greater than a user-defined minimum. Association rule mining algorithms, although similar in mining task, cannot be applied as is to temporal association rule mining. Strictly speaking, TAR are ordered where AR are not. When researchers began to incorporate order into AR mining, it wasn't considered TAR mining; rather it was grouped with sequential patterning mining because it did not consider gaps, intervals or complex temporal relationships. (Roddick and Spiliopoulou 2002)

In TARM there are two types of temporal rules. (Schluter and Conrad 2010) (See Figure 3.1) The first is an association rule that has a temporal constraint. The temporal constraint is vital to assessing the interestingness of the rule but is considered part of the rule and not incorporated directly into the interestingness measure. The association rule aspirin <--> ECG without a temporal constraint, when considered with all the association rules that can describe a hospitalization, this rule is not very interesting. But when the temporal constraint of "within 24 hours of admission" is included, this association rule becomes interesting. (For more discussion on interestingness see chapter 4.) A temporal constraint can be cyclic, lifespan/windowed, or calendar. A temporal rule with a cyclic constraint is a rule that repeats periodically, such as every two hours. (Ozden, Ramaswamy et al. 1998) A rule that describes patient monitoring or dispensing of medication doses would be an example of a cyclic rule. The lifespan/windowed rule is important during a specific period as in the rule expires after a given amount of time or a specific date.(Lee, Lin et al. 2001) A temporal rule derived from a policy is in effect as long as the policy is active. This rule is said to have a lifespan. A windowed rule is closely related to the lifespan rule in that it is only interesting in a specific time period. The previous example of "aspirin <--> ECG within 24 hours of admission" is a windowed rule. At 25 hours it is no longer interesting. The calendar rule is tied to a calendar date "December 2", rather than using a relative temporal representation "in 3 years".(Ramaswamy, Mahajan et al. 1998, Li, Shen et al. 2002 2002, Zimbrão, de Souza et al. 2002)

The second type of rule is one that incorporates the temporal relationship between events into the structure of the association rule. The rule A -> B, when used in the temporal mining context, is *A precedes B* and *A implies B*. The strength of the temporal relationship (A precedes B) and the association (A implies B) affect the interestingness of the temporal association rule. Embedding the temporal order in the structure of the rule means that the time of the event becomes more important. In a temporally structured rule, time of an event can be point or interval. A point is an event that occurs during a specific time and is represented as a time stamp or date stamp.(Ceglar and Roddick 2006) An interval event is an event that occurs over a period of time, an operation is an example of an interval event.(Ale and Rossi 2000) Having time as an integral part of the structure of the rule, means that we can mine rules where all the events happen in series, parallel or a hybrid.(Patnaik, Butler et al. 2011) A temporal rule with events in series means that the time of the events do not overlap, $t_{antecedent} < t_{consequent}$. A temporal rule with events in parallel allow for simultaneous events in the antecedent, consequent, or both. A rule with a group of events in parallel for the antecedent and the consequent is in series is a hybrid. To the best of my knowledge there is no research targeting the mining of hybrid rules.

Figure 3.1 Types of Temporal Association Rules



3.3 Temporal Association Rule Mining Algorithm

My TARM algorithm was designed for use with healthcare records which can have a large number of sequences and a large number of events per sequence. Previous studies that use association rules in medical data have a small number of events and generate rules with a single event on the right hand side and 0-1 event on the left hand side. (McCormick, Rudin et al. 2011) While it is important to capture complexity, I recognize that temporal association rule data mining can generate an astronomical number of rules. To address the potential "rule explosion" I preceded the rule mining with sequential pattern

mining. (See chapter 2) I used the discovered patterns, i.e. order sets, as a single event in the rule versus creating rules using all the events in the discovered pattern and the rest of the events in the encounter. Due to the complexity of clinical data, the number of sequences, and the length of the sequences, a memory expensive approach like APRIORI or pattern growth is inadequate to meet the demands of clinical data. My approach is a three- step process:

1) Condense sequential patterns using SPM (See chapter 2)
2) Discover rules using TARM (See Table 3.1)
3) Post-process discovered rules using interestingness measures and performance measures (See chapter 4)

The post-processing of discovered rules is to reduce the number of rules to those most useful for the user. This is accomplished by using interestingness measures. (See chapter 4) In addition, this step can include other user constraints such as lifespan or calendar constraints.

In this study, I created rules of three events. More formally an event, E<x,t> is a tuple of the event name attribute, x, and the event time stamp, t. An encounter is a sequence of events, $s_i = E_1, E_2, E_3 \ldots E_n$, ordered by increasing time, $t_{n-1} < t_n$. A rule, in this context, is 3 events, two in the antecedent and one in the consequent and it maintains the temporal order of the events in the sequence,

$$E_1 + E_2 \rightarrow E_3 \qquad \textbf{Equation 3.2}$$

where $t_1 < t_2 < t_3$. Using rules with three events as compared to the two (one antecedent and one consequent) in previous research allows for discovery of more complex rules.

To mine rules with the temporal order embedded in the structure, the goal of the algorithm is to discover all rules that satisfy the constraint, $t_1 \leq t_2 \leq t_3$. The constraint maintains the temporal structure while allowing for discovery of series and parallel events. As mentioned earlier, parallel events occur simultaneously whereas serial events do not. Related to the concept of series and parallel is gap. A gap is the duration between two events. In this implementation I selected a constraint that does not include an upper bound for the duration between events, because I wanted to mine proximal and distal relationships.

To discover the temporal association rules, I begin with a database of sequences. To increase the efficiency of the algorithm I scanned the database to find the distribution of the number of events in a sequence. In the TARM algorithm there is an efficiency constraint. This constraint sets the maximum

number of events used to create rules in a given sequence. This constraint is optional. I selected maximum number of events such that only 5% of the sequences were excluded. The longer the sequence, the more likely the TARM algorithm would be mining rules that occur only once. It is not meaningful or practical to compute interestingness measures using a single rule. Therefore the assumption is that missing these rules will have little impact on the findings of the mining task.

Table 3.1 Temporal Association Rule Mining Algorithm

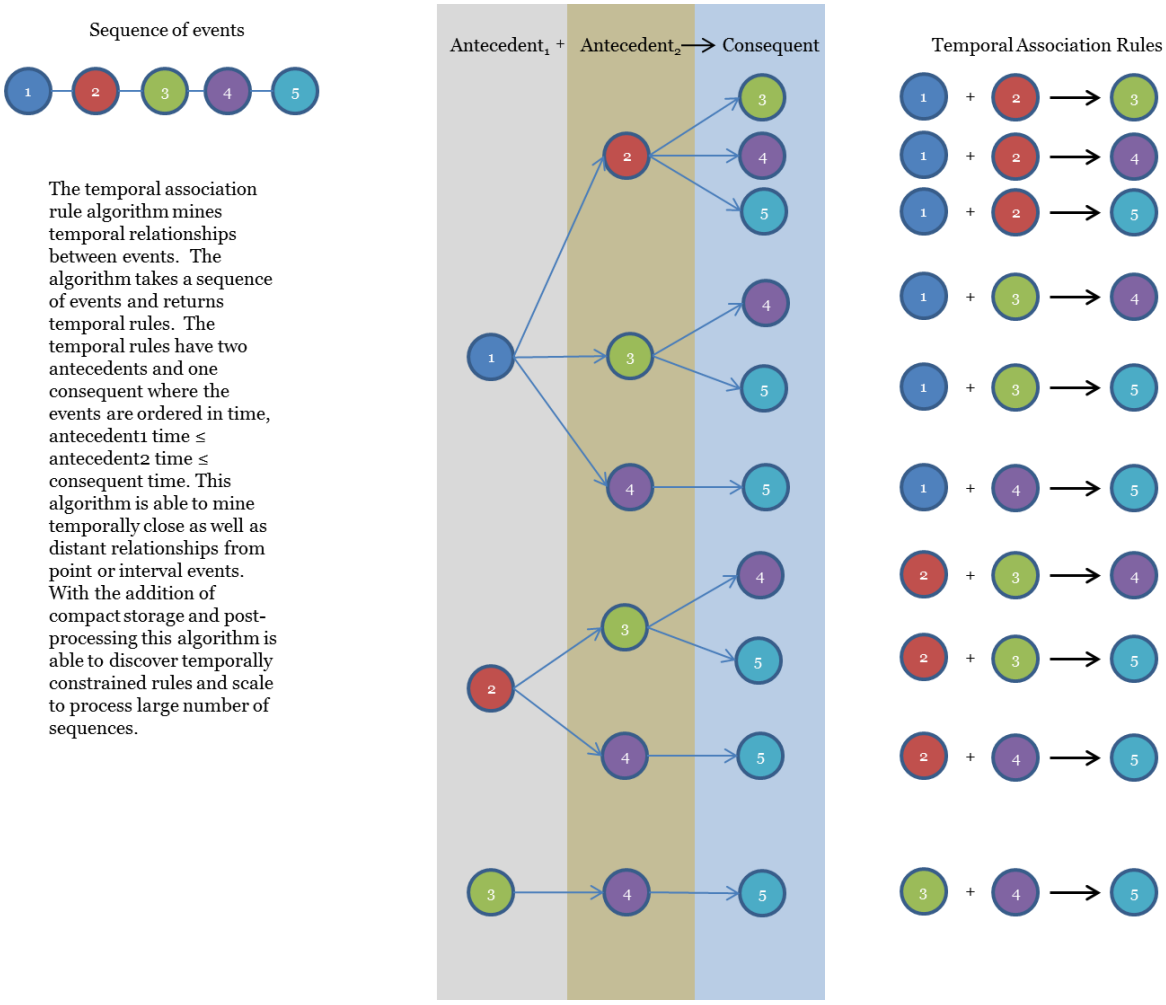
Input: Database, D, of sequences, $s_d = <e_o, t_o>, <e_1, t_1>...<e_n, t_n>$ where e is an event, t is the time in minutes from admission and n denotes the temporal ordering. Patterns, ID numbers and constituent events, discovered from Sequential Pattern Mining
Output: Set of temporal association rules, antecedent ($<e_i, t_i>, <e_j, t_j>$) -> consequent ($<e_k,t_k>$), where $i \leq j \leq k$.
Condense sequences using SPM discovered patterns For each sequence in the database For each pattern If the pattern is in the sequence, remove the events in the pattern from the sequence Insert the pattern ID into the sequence Assign the pattern the time of the last event removed from the sequence
Return condensed sequence
Create the temporal rules For each condensed sequence in the database For each event in the sequence, $<e_i, t_i>$ $<e_i, t_i>$ becomes the first event in the antecedent For each event $<e_j, t_j>$ where $i \leq j$ $<e_j, t_j>$ becomes the second event in the antecedent For each event $<e_k,t_k>$ where $j \leq k$ $<e_k,t_k>$ becomes the consequent
Return all temporal association rules
Post-processing Use user selected metrics to identify interesting rules

To create the rules I begin with the first event in the sequence, s_1. (See Table 3.1 and Figure 3.1) The first event, s_1, becomes E_1 in the temporal rule. The next event in the rule, E_2, is selected from the remaining events in the sequence that occur at the same time or after E_1, i.e. $s_2...s_{(n-1)}$. For simplicity, let's say I selected the second event in the sequence, s_2, to be E_2 in the rule. Now I have the antecedent events $E_1 + E_2$ as $s_1 + s_2$. The consequent event, E_3 can be any event that occurs at the same time or after E_2, $s_3...s_n$. If I select the first three events in the sequence, the temporal rule would be $s_1 + s_2 \rightarrow s_3$. To increase the efficacy of this step, I created a matrix of all three-item positions from 0 (the first position) to the value of the efficiency constraint. Creating the temporal association rules becomes a simple matching exercise. The user can use compact rule storage either in memory (not recommended) or on disk to maintain the rules for post processing.

3.4 Conclusion

The temporal association rule mining (TARM) algorithm I created is designed to discover rules comprised of temporally ordered events. It mines a large number of events and sequences but does not use candidates, can process intervals and includes gaps of arbitrary duration. I decided to decouple the identification of important rules from the rule mining and instead use user-specified measures in the post-processing of the rules. This means that this algorithm can capture low frequency rules and allows user flexibility in identifying interesting rules.

Figure 3.2 Temporal Association Rule Mining Algorithm



Chapter 4 Interestingness Measures for Temporal Association Rules

Temporal association rule data mining produces numerous rules making selection of the most interesting rules a difficult but necessary task. To automate or semi-automate the interesting rules selection process there are several metrics, called interestingness measures, that are used to rank the rules according to specific properties. A threshold or cutoff value of the interestingness measure is then used to determine the most interesting rules. The properties of the rules that make some rules more interesting than others depend on the application. To address the application-specific nature ranking rules, there is no universally appropriate measure and many interestingness measures have been proposed. Just as a universally appropriate measure does not exist, there also is no perfect measure; each one is a trade-off between aspects of the data it highlights, ease of interpretation and its reaction in certain conditions. Selection of the most appropriate measure is critical to the association rule data mining task.

In this chapter, I introduce interestingness measures and performance measures to evaluate temporal association rules.

4.1 Objective and Subjective Measures

Interestingness measures can be categorized as objective or subjective. Objective measures are data driven and are differentiated by their behavior in different conditions. Subjective measures incorporate a priori knowledge or expertise.[1] A subjective measure often does have an object component but an objective measure does not include external knowledge or information to rank rules. Geng further describes interestingness measures based on the structure and content of the rule; conciseness, coverage (generality), reliability, peculiarity, diversity, novelty, surprisingness, utility, or actionability. Conciseness, coverage, reliability, peculiarity and diversity are objective attributes because they characterize the structure of the rule. Novelty, surprisingness, utility, and actionability are subjective attributes and reference expert or domain knowledge. Table 4.1 provides a summary of attributes with short definitions.

Conciseness describes the length of the rule. The shorter the rule the more concise the rule is compared to other rules. The fraction or proportion of items in a set of rules that satisfy a specific example rule is the

[1] Here a priori refers to prior knowledge that is neither deduced nor derived from the current data and should not be confused with APRIORI the association rule data mining algorithm. For the clarification and consistency, the algorithm name will be written in small caps.

Table 4.1 Interestingness Measure Attributes

	Measure Attribute*	Definition
Objective Measures	Concise	A concise rule is a short rule
	Coverage (Generality)	The proportion of the rules that fit a specific pattern, e.g. relative support
	Reliability	The proportion of occurrences of the antecedent/ consequent relationship versus the antecedent, i.e. relative confidence.
	Peculiarity	Rules derived from outlier data. Rules that are far from other rules based on a given distance measure
	Diversity	The difference between the items in a rule, a highly diverse rule will have items that are very different from each other
Subjective Measures	Novelty	A rule that is new to the domain expert and could not be deduced from other rules.
	Surprisingness	A rule that contradicts expert knowledge
	Utility	A rule that contributes to a user's ultimate goal for data mining
	Actionability	A rule that contributes to decision making about future events.

*The contents of this table are based on the work of Geng et al. (Geng and Hamilton 2006)

coverage. It is the number of times a rule occurs in a rules database. The larger the proportion of the database that satisfies the rule then the greater the coverage of the rule is. The reliability of a rule is the consistency of the consequent occurring with the antecedent. This is also referred to as the confidence or the conditional probability of the consequent given the antecedent. The similarity or difference between two or more rules is quantified using a distance measure. The larger that distance is the more peculiar are rule is said to be. When the items that comprise a rule are similar to each other the rule lacks diversity. If the items are dissimilar, then the rule is considered diverse. When the result of the data mining process is a rule that is not contained in the apriori knowledge of the domain, it is novel. A rule that contradicts a priori knowledge is unexpected or surprising. A surprising rule may or may not be novel. The utility and actionability refer to the user's ability to use the rule or incorporate the rule into the decision making process.

The combination of data and domain knowledge is thought to lead to better and more useful data mining results. The context of temporal association rule data mining, a subjective measure such as novelty or surprise would be preferable for identifying interesting rules. However, incorporating domain knowledge in the data mining process, even in a specific task like selecting the interestingness measures and corresponding thresholds, is nontrivial. They rely on frequency, basically a popularity contest. The commonly used interestingness measures of support and confidence embody this myopic view and ignore rules with a low frequency and the probability of the consequent. Lift, likelihood, and Bayes factor were developed to address these shortcomings.

4.2 Properties and Selection Criteria for Objective Measures

One set of selection criteria for interestingness measures are the content and structure of the rule, an alternative approach to measure selection is how the rule behaves in various situations that could arise in the data. For example, a measure could increase in value with the frequency of the rule (coverage) or assume a maximum value when reliability is high.

A rule is composed of the antecedent and the consequent and to date all interestingness measures are based on the 2 x 2 contingency table where the rows and columns represent the presence/absence of the rule components. For the following discussion of interestingness measure properties the referent 2 x 2 table is composed of *a, b, c,* and *d* which are the number of rules in the set of all rules that satisfies the row/column conditions. The number of rules that contain the antecedent is $a + b$ and $a + c$ is the number that contain the consequent. The total number of rules is the sum *a, b, c,* and *d*. For a given rule $A \rightarrow B$ the value of *a* represents the confirmatory examples and *d* is the counter example of the rule $\neg A \rightarrow \neg B$.

Table 4.2 Antecedent and Consequent Contingency Table

		Consequent		
		Consequent	No Consequent	
		B	$\neg B$	
Antecedent	Antecedent	a	b	**a + b**
	A	$A \rightarrow B$	$A \rightarrow \neg B$	***A***
	No Antecedent	c	d	**c + d**
	$\neg A$	$\neg A \rightarrow B$	$\neg A \rightarrow \neg B$	$\neg$***A***
		a + c	b + d	
		B	$\neg B$	***T***

There are 5 properties used to differentiate the many interestingness measures. (See Table 4.4) Measures are differentiated on how they address the confirmatory and counter examples of a rule. If a measure assumes the same value for the rule $A \rightarrow B$ and its inverse $B \rightarrow A$, meaning it treats both the same then the measure is symmetrical. If a measure processes the rule and the inverse differently, then the measure is asymmetrical. For temporal association rules the antecedent precedes the consequent making the rule and its inverse distinct relationships. It is desirable, therefore, for the interestingness measure to be asymmetrical and process the rule without consideration for the inverse.

When selecting a measure to rank rules based on interestingness, it is important to understand how a particular measure behaves when the consequent frequency changes. In the contingency table, the frequency of the consequent is *c*, the number of rules that support the consequent but not the antecedent.

(See Table 4.2) As the frequency of the consequent (c) increases we can see from the contingency table that the reliability of the rule decreases. We want to know what the value of the measure does as c increases. Does the measure increase, decrease, or remain unchanged? For the current application, a measure whose value decreases as the consequent increases, i.e. as reliability decreases, making the rule less interesting is appropriate.

The increase in c is closely related to the concept of independence. Independence is when the occurrence of one event does not affect the presence or absence of another. This does not mean that the frequencies of the antecedent and consequent are equal but rather the probability of the consequent, $\frac{a+c}{T}$,and the probability of consequent given the antecedent, $\frac{a}{a+b}$,are equivalent. When the consequent is independent of the antecedent, how does the measure behave? Does the measure return a constant value or does it vary with the marginal frequencies? Independence in temporal association rules mining means that no temporal relationship exists between the antecedent and consequent. Therefore, a measure for ranking rules describing a temporal relationship should assume a constant value under independence.

When there are no counter examples of a rule, then $d = 0$. The $\lim_{d\to 0} measure = x$ and x could be a constant, ±∞, or a value dependent upon the frequency of the antecedent or consequent. The constant value, in this context should be the maximum value of the measure. This means that when there are no counter examples, the rule is more interesting. Not only are we concerned with the value of the measure when there are no counter examples, we are also interested in the shape of the measure values curve as the number of counter examples approaches 0. A curve can slowly approach the maximum value (concave), quickly approach the value (convex) or remain linear. In an extreme illustration, a convex curve reaches the maximum value while there are counter examples present whereas the concave curve only reaches the maximum when no counter examples exist. In temporal association rule data mining, the more desirable behavior of a measure in the context of decreasing counter examples is to approach a constant and maximal value. The tolerance of counter examples is minimal requiring a concave approach to the constant.

The final property by which measures are evaluated is their behavior with respect to T or the sensitivity of the measure to the size of the rule set. A measure that grows or shrinks with the total number of rules is

referred to as a statistical measure. A measure that is minimally or unaffected by the total is descriptive. Descriptive measures are valuable for comparing results across data mining tasks. For example an unexplored area of research is data mining using previous data mining results from the same or disparate sources. The limitation of descriptive measures is that as they approach their maximum value they lose the ability to discriminate between the rules. In this application, statistical measures are preferential to descriptive.

Temporal association rule data mining discovers a large number of temporal relationships between events in the form of rules with an antecedent and consequent. To distinguish between rules that are interesting and those that are uninteresting, the rules are ranked according to a measure. Many different measures exist and are evaluated based on specific properties. For temporal association rule data mining the measure would have to meet the following criteria:

1. Asymmetrical
2. Decreases as the consequent increases in value
3. Constant value when the antecedent and consequent are independent
4. Concave approaching a constant maximal value as the number of counter examples approaches 0
5. A statistical measure

Currently in the literature, there is no single measure that meets all of these criteria. The closest is the measure by Zhang (Zhang 2000). It meets all the above criteria except that it is a descriptive measure. (See Table 4.4)

$$Zhang^{'}s\ measure = \frac{T(n_{AB}) - n_A * n_B}{max\{n_{AB}, n_{\bar{B}}, (n_{A\bar{B}} * n_B)\}} \qquad \textbf{Equation 4.1}$$

The denominator is the maximum value of the rule $A \rightarrow B$, the frequency of the rules without the consequent regardless of the antecedent, or the frequency of the rules with the antecedent when not followed by the consequent multiplied by the frequency of the consequent. The numerator is the frequency of the antecedent multiplied by the frequency of the consequent subtracted from the total number of rules times the frequency of $A \rightarrow B$. When the antecedent and consequent are independent Zhang's measure is 0. When there are few counter examples to the rule, Zhang's measure approaches 1. When the antecedent is rarely associated with the consequent, Zhang's measure is =-1. Understanding how the rule behaves is important however understanding what a measure says about a rule is also

important. If a rule has a Zhang's measure of 0.7 and another rule has a Zhang's measure of 0.8 the rule with the higher measure is considered more interesting because it has fewer counter examples and has a high frequency.

Zhang's measure is only one measure used to rank rules and by including other measures we can get a more complete picture of the temporal associations of interest. (See Table 4.3) The two most common metrics are support and confidence. Support is the proportion of rules that contain the antecedent, A, or the precipitating events. Confidence is the conditional probability of the consequent, B, given the antecedent, A. Confidence is asymmetric and support in symmetric. In this context the support is scaled by the total number of rules therefore the inverse rule has the same support. But the confidence of the inverse rule will most likely be different. Both support and confidence are bounded [0, 1] which eases interpretation. A support value close to 1 signals more complete coverage. A confidence value that is close to 1 is a strong relationship between the antecedent and the consequent. As the consequent frequency increases, neither the support nor the confidence decreases. When respect to the situation where there are few counter examples of the rule, both measures are linear as the number of counter examples approaches zero. The concern with the support/confidence framework is that it doesn't account for the number of times the consequent occurs and when the occurrence of the antecedent and the consequent are independent.

Table 4.3 Temporal Association Rule Interestingness Measures

Interestingness Measure	Definition
Bayes' Factor	$\frac{n_{AB}n_{\bar{B}}}{n_B n_{A\bar{B}}}$
Confidence	$\frac{n_{AB}}{n_A}$
Lift	$\frac{nn_{AB}}{n_A n_B}$
Likelihood	$\log\left(\frac{n_{AB}}{n_{A\bar{B}}}\right)$
Support	$\frac{n_{AB}}{n}$
Zhang's	$\frac{T(n_{AB}) - n_A * n_B}{max\{n_{AB}, n_{\bar{B}}, (n_{A\bar{B}} * n_B)\}}$

The lift is a measure of how independent the antecedent and consequent are in the data. (Brin, Motwani et al. 1997) Like support, lift is a symmetric measure and is linear when the number of counter examples approaches zero. Like support, lift is sensitive to the size of the rule set. Dissimilar to support and confidence, lift is constant when the antecedent and consequent are independent and its maximum value is dependent on the frequency of the rule, antecedent, consequent, and rules set size. Lift does have the advantage of being easily interpreted. If the lift value is 4, then the consequent following the antecedent is 4x what is expected by chance.

Bayes' Factor is a measure of the evidence for the relationship of the consequent given the antecedent versus the evidence for the independence of the antecedent and consequent. The likelihood is also a comparison or ratio of two conditions. In this case, the likelihood is the comparison of the rule and its counter-example. The counter-example is the consequent that is not preceded by the antecedent. The likelihood measure is fundamentally different from the other measures. Bayes factor, confidence, lift, support, and Zhang's are confirmatory measures. Confirmatory measures are measures whose value changes with respect to the rule of interest. The value of the likelihood varies with the evidence to support one condition versus another. If there is more evidence, i.e. a higher probability for the rule, than for the counter-example, the likelihood will be positive. If the converse is true and there is more evidence to support the counter-example, then the likelihood will be negative. If there is equal evidence, then the likelihood will be 0.

Table 4.4 Interestingness Measures and Confirmatory Measure Properties

Property	Bayes	Confidence	Lift	Support	Zhang's
Symmetrical/Asymmetrical	Asymmetric	Asymmetric	Symmetric	Symmetric	Asymmetric
Consequent frequency increases	Decreases	Non-decreasing	Decreases	Non-decreasing	decrease
Independence	1	n_B/n	1	$n_A n_B/n^2$	0
No counter examples	Convex	Linear	Linear	linear	concave
Sensitivity to rule set size	No	No	Yes	Yes	No
Minimum value	0	0	0	0	-1
Maximum value	∞	1	n/n_B	n_A/n	1

4.3 Subjective Measures and Incorporating Expert Knowledge

Subjective measures incorporate domain or expert knowledge. For example, an objective measure can express the number of sequences in the database that contain a specific rule, i.e. the support. A subjective measure of novelty captures whether a specific rule was previously included in the domain knowledge. In this analysis I will focus on discovering rules that are surprising. A surprising rule is one that contradicts

what was previously known or expected based on domain knowledge. A rule can be unexpected if the antecedent or consequent is unanticipated or the pairing of the antecedent or consequent is unexpected. It should be noted that whether a rule is surprising is conceptually district from the rule frequency. A common rule can be surprising if it contradicts previous knowledge and an infrequence rule is expected if it can be inferred from domain knowledge. For example, say that we know jam and jelly are related and often interchangeable. Therefore an infrequent rule $peanut\ butter + jam \rightarrow bread$ is not surprising.

Therefore modeling subjective measures must include a model of what is expected. Horvitz pioneered the use of surprise modeling. As part of a traffic prediction system, users were alerted to unexpected situations. Alerting users to unexpected situations requires modeling user's expectations. The model of user's expectations was created from a case library of unexpected traffic situations that were identified by low event probability. This is similar to outlier detection or rare event modeling. Horvitz does point out that more sophisticated or nuanced methods exist to model user expectations. In association rule mining, the expectation of the user isn't modeled but the knowledge of the user is represented. An interestingness measure to distinguish surprising rules from expected rules has three components, 1) the modeling of domain knowledge and 2) a comparison of domain knowledge with the discovered rule and 3) a measure of how similar or dissimilar the rule is compared to domain knowledge.

User knowledge is referred to as a knowledge base and it is derived directly from experts or from data. Silberschatz asked experts to represent their understanding of domain knowledge using "relates to" statements. The statements where used to construct a Bayesian network where the nodes where the entities and the relationship "relates to" was the edges. (Silberschatz and Tuzhilin 1995, Silberschatz and Tuzhilin 1996) A data driven model is described by Padmanabhan. (Padmanabhan and Tuzhilin 1998, Padmanabhan and Tuzhilin 2000) Rules with high support and confidence values are not considered interesting, instead they are common sense. These common sense rules become the knowledge base. This data driven process assumes that the rules in the database have a high support and high confidence and that these rules encode basic domain knowledge. The challenge to this approach is a statistical paradox. Simpson's paradox is when the trend, in this case high support and high confidence, is true for the entire sample but not true for all subgroups in the sample. Both the data driven and expert derived knowledge databases are represented as Bayesian networks. Jaroszewicz used the knowledge base to

compute the conditional probability of a rule. Then the probability of the same rule was calculated using the data. (Jaroszewicz, Scheffer et al. 2009) The difference between the knowledge base and the data is the measure of unexpectedness. The larger the difference is the more unexpected a rule.

The approach where experts create the knowledge base using "relates to" statements in not practical due to the extensive "size" of the domain knowledge. The data driven approach to knowledge base creation assumes that common sense rules exist. The number of possible events is large and the number of potential combinations is exponential. Therefore it is unreasonable to assume that there exists a set of common sense rules that is sufficient to be called a knowledge base. Using the conditional probability of a rule calculated from a Bayesian network to identify unexpected rules assumes that the knowledge base is complete. A rule derived from the data must exist in the knowledge base in order to compare the conditional probabilities.

Due to the limited coverage of the domain knowledge, a direct similarity measure is not appealing. Because when a rule is not similar to the domain knowledge, it may be unexpected or it may be dissimilar because of missing or incomplete domain knowledge. The similarity measure would then be more a commentary on the completeness of the knowledge base and not the discovery of unexpected rules. In clinical medicine, the domain knowledge is extensive and the knowledge base would also need to be extensive to use a direct similarity measure like comparing conditional probabilities. Therefore, instead of using the conditional probability of the rule, I will use the outcome of the rule. The ultimate patient outcome of a rule provides a more compelling and informative measure of surprise than conditional probability. A surprising rule is a rule with a more favorable patient outcome than the knowledge base.

The challenge of creating a knowledge base for the treatment of a complex condition, in this case STEMI, is non-trivial. Domain knowledge in clinical medicine is captured in text documents, and not in a machine readable form or rule representation. The challenge then is to translate or transform text-based domain knowledge into a rule-based representation. Fortunately, in clinical medicine knowledge is transmitted via clinical practice guidelines and corresponding measures of compliance. The compliance measures, referred to as performance measures, are presented as

fractions with explicit instructions on the composition of the numerator and denominator. (Krumholz, Anderson et al. 2006, Krumholz, Anderson et al. 2008) The rule representation of the performance measures for the treatment of STEMI will be the knowledge base. For this dissertation, I manually translated these performance measures into temporal rule representation. (See Table 4.6)

4.3.1 Performance Measures as a Knowledge Base

A rule representation of the performance measures must be compatible with the EMR data in the ACSPD, clearly delineate time dependencies, and only include unambiguous events from reliable sources. The rule representation is compatible with the EMR data if it uses only medication, laboratory tests, or procedures that are present in the ACSPD. The time dependencies are the temporal relationship(s) between the events in the rule. A clear time-based ordering of events is a basic assumption of temporal data mining. Therefore when using performance measures as a knowledge base, temporal order must be clear. Temporal association rules can express temporal relationships in two ways, 1) embedded in the antecedent-consequent structure of the rule, and 2) temporal constraints external to the rule. Finally, the events in the performance measures must be unambiguous and have a corollary in the ACSPD. These three criteria were used to create the rule representation of the performance measure.

The performance measures were read to identify and remove references, event, or constraints that were incompatible with the EMR data available in the ACSPD. (See Table 4.5) For example, performance measure 5 (PM5) reads "AMI patients with documentation in the hospital record that left ventricular (LV) systolic function was evaluated during hospitalization or is planned for after discharge". The ACSPD does not contain physician notes or other written documentation. The ACSPD also does not contain the finding of a procedure or examination. Therefore the documentation of left ventricular systolic function is not possible with the ACSPD. When the phrase "documentation in the hospital record that left ventricular (LV) systolic function" is removed for incompatibility, the measure becomes meaningless and was withdrawn from further consideration. Performance measures PM12, PM13, PMt7, and PMt8 were excluded for incompatibility because the ACSPD does not contain the specific data necessary. (See Table 4.5 and Table 4.6)

Performance measures could be incompatible with the ACSPD if the data required was outside the scope of the database. The ACSPD does contain some information on patient transfers; however this is usually transfers within the same facility such as emergency department to the intensive care unit. The facilities in the ACSPD had to use Cerner software, and agree to participate. Thus the information about patient care involving transfers to other facilities is limited. Performance measures 9, 10, and 11 were excluded due to database scope. Please note that AMI or acute myocardial infarction is a general term that includes STEMI, NSTEMI and UA. When a performance measure uses the term MI, the data used to create the temporal rules is only patients with a diagnosis of STEMI.

As with most databases, even large databases, the quality of the data can vary. There are certain fields that are more reliable than others, and a reliable field is referred to as valid. The rule representation of the performance measures should only use data from reliable sources within the ACSPD. Performance measures PMt2-PMt6 address excess dosage of medications used to treat STEMI. The calculation for the correct dose includes weight. In the ACSPD, weight contains a high number of missing values and those that remain are not valid. Because an important variable in the dosage calculation is unreliable, the performance measures PMt2-PMt6 were removed from consideration. (See Table 4.5)

The performance measures PM6, PM7, and PM8 contained clauses similar to "LBBB (left bundle branch block) on the ECG closest to the arrival time". (See Table 4.5) As mentioned previously, the ACSPD does not contain the findings of procedures. When the clause is removed from the performance measures, the remaining measure is compatible and within the scope of the ACSPD. Therefore these performance measures were retained for further consideration. (See Table 4.6)

The rule representation has events which comprise the antecedent and consequent and can express temporal relationships by structure or external constraint. The structure of the rule is the temporal order of events. The external constraint establishes a specific temporal boundary for events in the antecedent or consequent. A rule must have one temporal relationship and can have both structure and constraint. Performance measure PM1, "AMI patients who received aspirin within 24 hours

before or after hospital arrival", is an example of a rule with structural temporal relationships (before, after) and constraint (within 24 hours). (See Table 4.5)

To translate the remaining performance measures, PM1-PM4, PM6-PM8, PMt1 and PMt9, into rule representation, I first identified events and then temporal relationships. (See Table 4.5) In performance measure PM1, the events are "aspirin" and "hospital arrival". The events are combined with the temporal relationships, "before", "after", "within 24 hours". For example a text version of the rule would be "aspirin 24 hours after admission". The rule representation is $aspirin_{time \leq admission + 24\ hours}$. For the complete translation of the performance measures, see Table 4.6.

The translated performance measures form the limited knowledge base. A subjective measure of surprise is the outcome of the rule compared to the knowledge base. A surprising rule is a rule with a better outcome than the knowledge base. Performance measures can capture domain knowledge; admittedly the coverage of the breadth of knowledge is limited. However, the performance measure coverage is offset by the ability to use a common data representation and ease of outcome comparison.

4.4 Conclusion

Temporal association rule data mining discovers a large number of rules, far too many to be useful. Some rules may be more interesting to users than others. Interestingness measures can be used to distinguish important rules and unimportant rules based on user specified criteria. In this chapter I show the criteria used to select the interestingness measures and measure behavior under difference circumstances. Interestingness measures can be divided into subjective and objective. Objective measures are data driven and subjective measures incorporate expertise. For this dissertation I selected support, confidence, Bayes' factor, lift, likelihood and Zhang's as objective measures to identify clinically interesting results. For subjective measures, the challenge is expressing domain expertise in a manner sufficient for comparison with the rules. The AHA/ACC developed clinical practice performance measures for STEMI treatment to aid hospitals in assessing quality of care. These measures incorporate clinical knowledge and evidence-based medicine best practices for STEMI treatment. I translated these

performance measures into temporal association rule representation. In the next chapter I will compare clinical practice patterns derived from the TARM with the performance measures and analyze subsequent patient outcomes.

Table 4.5 AHA/ACC STEMI Performance Measures Published in 2006 and 2008

	Performance Measure	2006	2008
1	AMI patients who received aspirin within 24 hours before or after hospital arrival	X	X
2	AMI patients who are prescribed aspirin at hospital discharge	X	X
3	AMI patients who are prescribed beta blockers at arrival	X	
3	AMI patients who are prescribed a beta-blocker at hospital discharge	X	X
4	AMI patients who are prescribed a statin at discharge	X	X
5	AMI patients with documentation in the hospital record that left ventricular (LV) systolic function was evaluated during hospitalization or is planned for after discharge		X
6	AMI patients with left ventricular systolic dysfunction (LVSD) who are prescribed an ACEI or ARB at hospital discharge.	X	X
7	AMI patients with STEMI or LBBB on the ECG closest to arrival time receiving fibrinolytic therapy during the hospital stay and having a time from hospital arrival to fibrinolysis of 30min or less	X	X
8	AMI patients with STEMI or LBBB on the ECG closest to the arrival time receiving primary PCI during the hospital stay with a time from hospital arrival to PCI of 90 minute or less.	X	X
9	Percentage of AMI patients with STEMI or LBBB on the ECG performed closest to the arrival time receiving either fibrinolysis or primary PCI or who are transferred to another facility for PCI	X	X
10	Median time from emergency department arrival at STEMI referral facility to ED discharge from STEMI referral facility for AMI patients with STEMI or LBBB on the ECG performed closest to arrival time, who are transferred to a STEMI receiving facility for primary PCI.		X
11	Median time from patient arrival at a STEMI referral facility's ED to time of primary PCI at STEMI receiving facility for AMI patients presenting with STEMI or LBBB on the ECG performed closest to the arrival time who are transferred to a STEMI receiving facility for primary PCI		X
12	AMI patients with a history of smoking cigarettes who are given smoking cessation advice or counseling during hospital stay.	X	X
13	All patients hospitalized with a primary diagnosis of AMI referred to an early outpatient cardiac rehabilitation program.		X
T-1	AMI patients with documentation of low-density lipoprotein-cholesterol (LDL-C) level in the hospital record or documentation that LDC-C testing was done during the hospital stay or is planned for after discharge.	X	X
T-2	AMI patients who received excess dosing of unfractionated heparin (UFH) initially		X
T-3	AMI patients who received excessive dosing of subcutaneous enoxaparin initially.		X
T-4	AMI patients who received excess dosing of abviximab initially		X
T-5	AMI patients who received excess doing of epitifibatide initially.		X
T-6	AMI patients who received excess dosing of tirofiban initially		X
T-7	Presence of a protocol or other clinical aid in the hospital record of AMI patients that addresses dosing o anticoagulant therapy and intravenous antiplatelet therapy (i.e. unfractionated heparin, low molecular weight heparin, and glycoprotein IIb/IIIa inhibitors).		X
T-8	Evidence of a tracking system or identifying dosing errors in anticoagulation therapy in the hospital record of AMI patients.		X
T-9	Medically treated AMI patients who are prescribed clopidogrel or ticlopidine at hospital discharge		X

Table 4.6 Rule Representation of AHA STEMI Performance Measures

PM	Rule Representation
PM1	$aspirin_{time \leq admission\ +24\ hours} + event \rightarrow event$ $event + aspirin_{time \leq admission\ +24\ hours} \rightarrow event$ $event + event \rightarrow aspirin_{time \leq admission\ +24\ hours}$
PM2	$aspirin_{time \geq discharge\ -24\ hours} + event \rightarrow event$ $event + aspirin_{time \geq discharge\ -24\ hours} \rightarrow event$ $event + event \rightarrow aspirin_{time \geq discharge\ -24\ hours}$
PM3a	$beta-blocker_{time \leq admission\ +24\ hours} + event \rightarrow event$ $event + beta-blocker_{time \leq admission\ +24\ hours} \rightarrow event$ $event + event \rightarrow beta-blocker_{time \leq admission\ +24\ hours}$
PM3b	$beta-blocker_{time \geq discharge\ -24\ hours} + event \rightarrow event$ $event + beta-blocker_{time \geq discharge\ -24\ hours} \rightarrow event$ $event + event \rightarrow beta-blocker_{time \geq discharge\ -24\ hours}$
PM4	$stain_{time \geq discharge\ -24\ hours} + event \rightarrow event$ $event + statin_{time \geq discharge\ -24\ hours} \rightarrow event$ $event + event \rightarrow statin_{time \geq discharge\ -24\ hours}$
PM6	$ACEI_{time \geq discharge\ -24\ hours} + event \rightarrow event$ $event + ACEI_{time \geq discharge\ -24\ hours} \rightarrow event$ $event + event \rightarrow ACEI_{time \geq discharge\ -24\ hours}$ $ARB_{time \geq discharge\ -24\ hours} + event \rightarrow event$ $event + ARB_{time \geq discharge\ -24\ hours} \rightarrow event$ $event + event \rightarrow ARB_{time \geq discharge\ -24\ hours}$
PM7	$tenecteplase_{time \leq admission\ +30min} + event \rightarrow event$ $event + tenecteplase_{time \leq admission\ +30\ min} \rightarrow event$ $event + event \rightarrow tenecteplase_{time \leq admission\ +30\ min}$ $reteplase_{time \leq admission\ +30min} + event \rightarrow event$ $event + reteplase_{time \leq admission\ +30\ min} \rightarrow event$ $event + event \rightarrow reteplase_{time \leq admission\ +30\ min}$ $alteplase_{time \leq admission\ +30min} + event \rightarrow event$ $event + alteplase_{time \leq admission\ +30\ min} \rightarrow event$ $event + event \rightarrow alteplase_{time \leq admission\ +30\ min}$ $streptokinase_{time \leq admission\ +30min} + event \rightarrow event$ $event + streptokinase_{time \leq admission\ +30\ min} \rightarrow event$ $event + event \rightarrow streptokinase_{time \leq admission\ +30\ min}$
PM8	$PCI_{time \leq admission\ +90min} + event \rightarrow event$ $event + PCI_{time \leq admission\ +90\ min} \rightarrow event$ $event + event \rightarrow PCI_{time < admission\ +90\ min}$

PM	Rule Representation
PMt1	$LDL + event \rightarrow event$ $event + LDL \rightarrow event$ $event + event \rightarrow LDL$
PMt9	$clopidogrel_{time \geq discharge\ -24\ hours} + \begin{Bmatrix} tenecteplase \\ reteplase \\ alteplase \\ streptokinase \end{Bmatrix} \rightarrow event$ $\begin{Bmatrix} tenecteplase \\ reteplase \\ alteplase \\ streptokinase \end{Bmatrix} + clopidogrel_{time \geq discharge\ -24\ hours} \rightarrow event$ $event + \begin{Bmatrix} tenecteplase \\ reteplase \\ alteplase \\ streptokinase \end{Bmatrix} \rightarrow clopidogrel_{time \geq discharge\ -24\ hours}$ $\begin{Bmatrix} tenecteplase \\ reteplase \\ alteplase \\ streptokinase \end{Bmatrix} + event \rightarrow clopidogrel_{time \geq discharge\ -24\ hours}$ $ticlopidine_{time \geq discharge\ -24\ hours} + \begin{Bmatrix} tenecteplase \\ reteplase \\ alteplase \\ streptokinase \end{Bmatrix} \rightarrow event$ $\begin{Bmatrix} tenecteplase \\ reteplase \\ alteplase \\ streptokinase \end{Bmatrix} + ticlopidine_{time \geq discharge\ -24\ hours} \rightarrow event$ $event + \begin{Bmatrix} tenecteplase \\ reteplase \\ alteplase \\ streptokinase \end{Bmatrix} \rightarrow ticlopidine_{time \geq discharge\ -24\ hours}$ $\begin{Bmatrix} tenecteplase \\ reteplase \\ alteplase \\ streptokinase \end{Bmatrix} + event \rightarrow ticlopidine_{time \geq discharge\ -24\ hours}$

Chapter 5 SPM and TARM in ACSPD

In modern healthcare settings, diagnoses, course of treatment and outcomes are recorded in an electronic health record, including temporal (date and time) information. This time stamp for events is valuable because in the treatment of ACS what is done may not be as important as when it is done; variations in sequence and timing affect patient outcomes. (Eagle 2002, Armstrong, Collen et al. 2003) Although the quality of EHR data is not perfect, both in the accuracy of events and diagnoses, as well as the accuracy of the temporal information, in aggregate, such data can provide valuable insights about the practice of healthcare.

In this chapter, I introduce the acute coronary syndrome patient database, apply sequential and temporal association rule mining to discover order sets and clinical practice patterns, then I evaluate the clinical practice patterns with interestingness measures and investigate the influence of AHA performance measure compliance on subsequent outcomes.

5.1 Acute Coronary Syndrome Patient Database

HealthFacts® is a HIPAA compliant clinical data source from Cerner. The data are derived from EMR systems and contains information from pharmacy, laboratory, admission, and billing on patient encounters from healthcare institutions across the United States. I received permission to use HealthFacts® and extracted a portion for acute coronary syndrome research.

More precisely, I defined this extract as HealthFacts® data from all healthcare encounters with a medication, laboratory test, procedure, or diagnosis indicative of acute coronary syndrome, where the patient was discharged from January 2000 to December 2009. I also included anonymized patient demographic information and hospital or institution characteristics. Henceforth I will refer to this extracted data as "ACSPD", for the Acute Coronary Syndrome Patient Database. For all patients in the ACSPD, I also extracted all encounter information from HealthFacts® over the nine years, including non-ACS encounters, in order to gather information about chronic conditions, outcomes, and survival.

I believe that the ACSPD is unique in its size and national breadth of information about acute coronary syndrome. It includes 23 million encounters for 6 million patients in 128 healthcare institutions across the US. There are 88 million medication orders, 648 million laboratory results, 35 million diagnoses, and

almost 4 million procedures. Of the 6 million patients, 58% are female (n=3,694,642) and 73% are Caucasian (n=4,664,543). The patients averaged 4.1 encounters (range: 1-564, standard deviation: 7.7) over the nine year time period. Forty–five percent (n=58) of the healthcare entities have at least 200 beds, are mostly in urban locations (86%, n=110) and are split between non-teaching (56%, n=72) and teaching (35.1%, n=45; unknown 8.5%, n=11).

The US 2010 Census estimates that the resident population is 50.6% female and 79.5% self-identified as white. (Census 2010) In the ACSPD females and minorities are over represented. There is evidence that there is a gender disparity in ACS symptoms, treatment and outcomes and racial disparities in coronary artery disease, the underlying cause of ACS. The population in the ACSPD is representative of the ACS population in the US.

One of the challenges of electronic medical records is that they contain multiple time series and data are collected asynchronously. In addition, the granularity of the timestamps can vary. This challenge is further increased when using data from across multiple healthcare settings due to variations in what data are collected and the granularity of that data. Standardization of data collection, like that found in the ACSPD, can diminish the variation between healthcare settings and aid cross-site analysis.

The number of patients and encounters, and the various healthcare institutions make this database unique. In addition, the longitudinal nature of the database allows us to examine the changes in treatment over time. The changes in treatment over time may represent the adoption of the treatment guidelines, incorporation of novel procedures or drug regimens into standard practice (Hess, Wells et al. 2008), increased knowledge about the treatment of ACS from clinical trials, changes in resources for treatment or shifts in patient characteristics.

For this dissertation, I used only ACSPD encounters with an International Classification of Disease, 9th Revision (ICD9) code of 410.0-410.06 and 410.08, an admission date, and at least 3 events with a date or date and time stamp. I included inpatient, outpatient, and emergency room patients. Patient transfers and/or admission and discharge dates within 5 min were considered a continuation of the same encounter.

5.1.1 Events

In ACSPD data, there are 6 categories of events, but only 5 are of interest for this dissertation: admission, discharge, medication order, procedure, and laboratory testing. (Diagnosis codes are assigned as part of the billing process after discharge.) (See Table 5.1) Each category contains many specific events, e.g. aspirin or CABG, each type of event is identified by a unique numeric code and

Table 5.1 Event Type and Time Granularity in the ACSPD

Event type	Time granularity
Admission	Datetime
Diagnosis code*	Relative relationship
Discharge	Datetime
Laboratory test	Interval
Medication order	Interval
Procedure	Interval

*Diagnoses ICD 9 codes are assigned after discharge.

has time information of varying granularity. The admission and discharge of a patient during an encounter are a single time point with date and time granularity. The medication orders and laboratory testing are either a single event that occurred at a specific point during an interval or a continuous action over an interval. For example, a single dose of aspirin or an intravenous drug. Laboratory test have a date and time corresponding to when the test was administered, when it was processed and when the analysis was completed and results reported. A laboratory test can be represented as a single point in time, when it was administered or when the results were reported, or as a process from when it was administered to when the results were reported. I selected an interval representation, i.e. when the laboratory test was ordered and when the results were made available. Ordering a laboratory test is the result of a healthcare decision. When the results are available is an event that could affect subsequent decisions and patient care.

The data in the ACSPD is time stamped at different levels of granularity, date/time and date based on the type of event. (See Table 5.1) Procedures are date stamped. For this dissertation, the level of granularity of a date is interpreted to mean the procedure occurred during a 24 hour interval. The medication orders and laboratory tests have a starting date and time and an ending date and time. Admission and discharge date and time stamps indicate a specific point in time when the action occurred. Diagnoses are unique in that there is no time stamp associated with them in the ACSPD, because they are assigned after discharge. The various time granularities in the ACSPD mean that

the temporal relationships are point and intervals. In order for the data mining algorithms to process point and interval data, the intervals are treated as two separate events, a start and stop.

5.1.2 Encounters

An encounter is a combination of what happened (events) and when those events occurred (timing). In this dissertation, I represent each encounter using two vectors, one for the events and one containing the timing of each event. Medication orders and laboratory tests, being intervals are two events in the sequence, a start and an end. When two or more events occur at the same time, they are ordered descending by their unique identification number. In order to compare the timing of events across encounters, I transformed the time sequence from absolute date time stamps to the number of minutes from admission.

5.1.3 Outcomes

In healthcare, the treatment decisions are often judged by the ultimate outcome they produce for the patient. It is my assumption that the events in the encounter contribute to the effect of treatment. In this work, I focus on two outcomes of interest to cardiology and STEMI care, in-hospital mortality and bleeding. Bleeding is a risk associated with the medications used to restore normal cardiac function.

The outcomes of the treatment decisions are either collected as part of normal operations of the healthcare facility or derived from the ACSPD. Discharge disposition, including in-hospital mortality is recorded for the encounters in the data. A bleed is more complex to ascertain and there are several definitions of a bleed event. The differences in the definitions can impact the incidence of bleed events.(Steinhubl, Kastrati et al. 2007) I selected the definition used by the Can Rapid Risk Stratification of Unstable Angina Patients Suppress Adverse Outcomes with Early Implementation of the American College of Cardiology/American Heart Association Guidelines (CRUSADE) initiative and registry.(Hoekstra, Pollack et al. 2002, Subherwal, Bach et al. 2009) A bleed event was one of the following: 1) a drop in absolute hematocrit of 12%, 2) intracranial hemorrhage (ICD9 430-432), 3) gastrointestinal bleed (ICD9 456.0, 456.20, 459.0, 530.7, 530.82, 531-534.[0,2,4,6], 537.83, 562.[02,03,12,13], 569.3, 569.85, 578, 623.8, 626.8), and 4) red blood cell transfusion if baseline hematocrit was ≥ 28%. For patients undergoing surgical revascularization, the bleeding event

criteria had to occur before revascularization procedure. Due to the complexity of defining bleeding and the various levels of granularity, there are some encounters with a bleed and limited data as to the timing of the event.

5.2 Study Design

The syndromes of ACS are different both in severity and treatment. Angina, the mildest of the syndromes rarely requires surgical intervention. Non ST-elevated myocardial infarction is more severe and can cause significant morbidity. Because it causes the highest morbidity and mortality of the ACS syndromes, this research focuses on ST-elevated myocardial infarction (STEMI). A STEMI encounter had an ICD9 code of 410.0-410.06 and 410.08.

The events in a STEMI encounter are numerous and the timing is asynchronous. This means that the temporal relationships between events can be complex. In the ACSPD, there are specific points in time, time intervals, and relative temporal relationships. In order to mine temporal relationships I used sequential pattern mining and temporal association rule mining. (See Figure 5.1) The sequential pattern mining discovers order sets. These order sets are used to condense the STEMI encounter sequence of events prior to applying temporal association rule mining to discover clinical practice patterns. The clinical practice patterns are ranked by interestingness measures. The clinical practice patterns are also compared to AHA/ACC performance measures for compliance and analysis of subsequent patient outcomes.

The STEMI encounters were divided into 2 experimental groups using simple random sampling. (See Figure 5.2) The first experimental group, comprising 1/3 of the data is for discovering order sets, clinical practice patterns, interestingness measures and performance measure compliance analyses. The second group is to validate the findings from sequential pattern mining and temporal association rule data mining.

Figure 5.1 Study Design

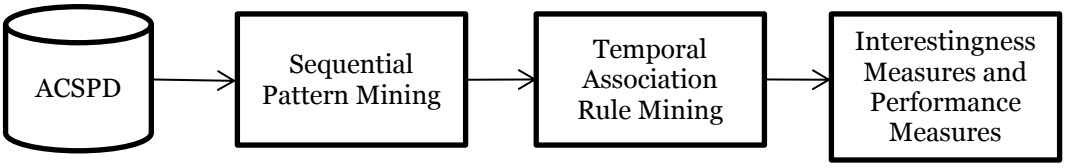
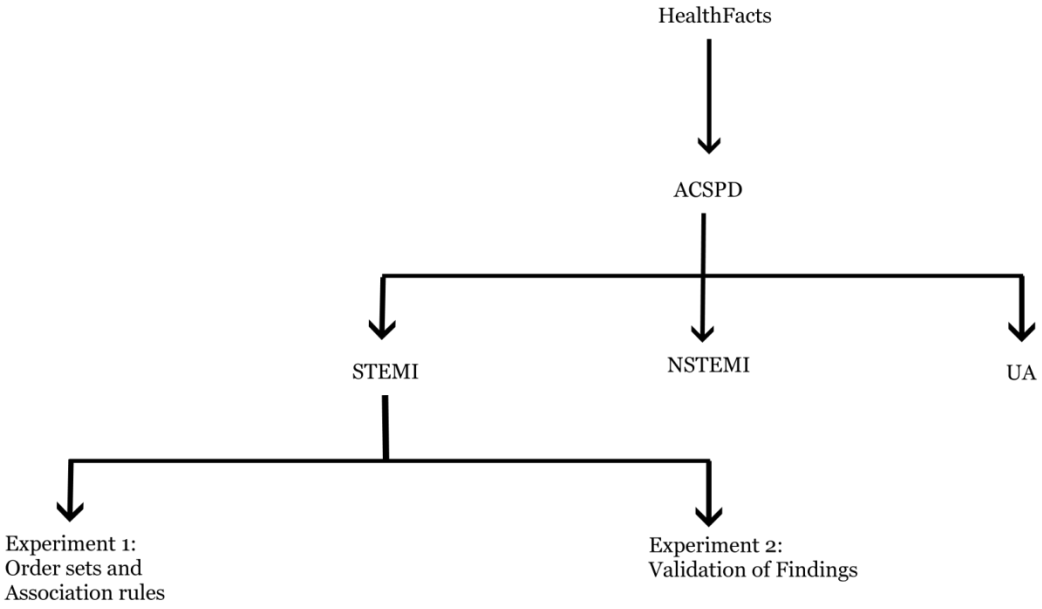


Figure 5.2 Study Data



5.3 Sequential Pattern Mining to Find Order Sets

In clinical medicine the way healthcare providers transmit information about how a patient is to be cared for is by an order. An order set is a group of orders that are commonly issued together. Order sets are included in many hospital electronic medical record systems as part of the CPOE. Order sets can vary from facility to facility and change over time. Due to the anonymization of the healthcare institutions in the ACSPD, it is not possible to ascertain directly from the facility the order sets used in patient care at a particular time. When creating and identifying interesting rules the co-occurrence of events is important and the basis for many of the measures of rule interestingness. Rules generated from order sets will demonstrate strong co-occurrence properties and could potentially be considered interesting. When, in

fact, the rules generated from the events in order sets are not interesting and are the result of institutional policy. But rules that contain order sets could be interesting when exploring healthcare practice.

In the ACSPD, order sets are not explicitly identified, they must be found. To find order sets in the ACSPD I make three assumptions. The first is that order sets contain frequent items and the second is that the constituents vary little and the third is that order sets do not contain gaps or "skip" events in the sequence. I used the SPM to identify order sets in the ACSPD. (See chapter 2) The SPM first identified events with a frequency of occurrence in the ACSPD that was above a threshold. An order set, in this data, is a sequence of events whose co-occurrence is unexpected given the background frequency of the events. The conditional probability of a sequence of events is the product of the conditional probability of all preceding events. Although using the conditional probability is ideal and potentially computationally intense, for this dissertation, using the event frequency was sufficient. To identify an order set I started with finding the events with a frequency above the threshold. To the best of my knowledge there is no set threshold for defining a frequent event. I experimented with various threshold values and selected a threshold that balances the need to find the most order sets without the analyses become intractable. Using the frequent events as a starting point, I searched for the longest sequences where the confidence threshold is above the cutoff. The confidence threshold is the frequency of the event in question given the previous events, i.e. the child sequence, divided by the frequency of the previous events, the parent sequence.

5.4 Discover Clinical Practice Patterns Using Temporal Association Rule Mining Algorithm

The temporal association rule data mining algorithm is based on the APRIORI or market basket analysis. The goal is to discover rules that describe the temporal relationships between events that occur frequently in the data. (Agrawal, Imieliński et al. 1993, Ale and Rossi 2000, Lee, Chen et al. 2003, Yan, Han et al. 2003, Wang and Han 2004, Pray and Ruiz 2005, Verma and Vyas 2005, Laxman and Sastry 2006) A temporal rule is a sequence of events ordered in time where the antecedent event or events precedes the consequent event. The antecedent can have an arbitrary number of events and the consequent is a single event. The approach to temporal association rule mining used in this dissertation differs from previous implementations in that it mines rules with non-sequential relationships, allows for interval events, and removes order sets. The temporal association rule mining algorithm has three phases, 1) condense rules

using order sets discovered by SPM, 2) rule generation, and 3) post processing using user selected metrics. The ACSPD transaction data is processed into sequences of events and event times relative to the encounter admission. The condensing of sequences phase, finds all order sets and replaces the given order sets in the encounter sequences with the new identifying code. The rule generation phase records all rules present in the encounter sequences. The third phase uses user selected post processing metrics to filter rules interesting. (See chapter 3) It is important to note that although I did not use any privacy preserving thresholds in this phase, the data from the ACSPD is HIPAA compliant and de-identified, and the rules and algorithm security were maintained during the entire process.

In this implementation of the temporal association rules I take the first event in the sequence and one of the subsequent events to constitute the antecedent. I then create a rule by combining the antecedent and a one of events that follows the antecedent. This is repeated for each of the subsequent events (consequents) in the event sequence. When all rules for the first event are created I use the next event in the sequence and continue the process. Here I report on the rules of 3 events, 2 events for the antecedent and 1 event for the consequent. (See chapter 3)

5.5 Clinical Performance Measures Represented as Temporal Association Rules

In the US, it is the American Heart Association (AHA) in conjunction with the American College of Cardiology (ACC) that publishes clinical practice guidelines for the treatment of acute coronary syndrome. The purpose of the guidelines is to provide clinicians with evidence-based treatments and expert opinion consensus for the optimal care of ACS. The AHA and ACC periodically update the clinical practice guidelines to incorporate new research findings. During the period of the ACSPD, 2000-2009, the AHA and ACC published 3 different updates in 1999, 2004, and 2007 for STEMI. (Ryan, Antman et al. 1999, Antman, Anbe et al. 2004, Anderson, Adams et al. 2007) In 2004 and 2008 the AHA published performance measures that are easily captured metrics of institutional adherence to the guidelines.(Krumholz, Anderson et al. , Krumholz, Anderson et al.) In addition, the AHA/ACC published guidelines pertaining to specific a population or procedure. This work will focus on the core STEMI guidelines (1999, 2004, and 2007) and the STEMI performance measures (2004 and 2008).

In 2001, the Institute of Medicine spotlighted the quality of medical care, specifically the difference between the care a patient received and the care the patient should have received according to evidence-based guidelines.(Richardson, Berwick et al. 2000) The AHA and ACC developed a set of measures to promote improvement in the quality of care for acute myocardial infarction patients.(Krumholz, Anderson et al. 2006, Krumholz, Anderson et al. 2008) The AHA/ACC strove to create measures that were easy to employ and could be used both as prospective and retrospective assessments of care quality. The measures were based on existing guideline Class I and Class III recommendations. Class I are recommendations for a treatment or procedure and Class III are recommendations prohibiting a treatment or procedure. Within each class are levels that correspond to the strength of the recommendation. Level A recommendation has the most unequivocal evidence from multiple studies or meta-analyses in diverse populations. Level B and C have decreasing weight of evidence. The quality measures use Level A recommendations. Not all recommendations meeting the class and level criteria were included in the performance measures. Instead only those measures with the greatest potential impact and lowest potential for unintended consequences were selected. The strength of evidence and ease of use criteria for the measures are so that they "may serve as vehicles to accelerate appropriate translation of scientific evidence into clinical practice" (Krumholz, Anderson et al. 2008).

The AHA/ACC measures are divided into performance measures and test measures. Strictly speaking, performance measures can be used for internal quality improvement and comparison between institutions. The AHA/ACC suggests that the test measures only be used for internal quality improvement and not for comparison purposes until further validation is completed. For the purposes of this dissertation, the distinction between test and performance measures is not made and both are referred to as performance measures. For consistency with the AHA/ACC naming convention is retained and the performance measures are designated PM and the test measures with a T.

Representation of performance measures in machine readable form is an active research area in artificial intelligence and decision support.(Bottrighi, Chesani et al. 2010) And there are several representation languages to choose from, Asbru (Miksch, Shahar et al. 1997), EON (Musen, Tu et al.), GLIF (Ohno-Machado, Gennari et al. 1998), GUIDE (Peleg, Boxwala et al. 2000), PRODIGY (Johnson, Tu et al. 2000), PROFORMA (Fox, Johns et al. 1998), GLARE (Terenziani, Montani et al. 2003), and PGROVE (Chesani,

Lamma et al. 2008), to facilitate the creation and use of clinical guidelines in patient data. Each representation language was developed for a specific purpose. Unfortunately these representation languages were not developed for use in data mining applications. Therefore the performance measures must be translated into a representation similar to the data mining format, i.e. the temporal association rule.

The performance measures focus on a wide array of treatment areas related to ACS, and not all of them are applicable or measurable in EMR data. The measures that are appropriate for the ACSPD I carefully translated into temporal association rule representation. With this common representation the performance measures and data mining results are comparable. A performance measure was translated into the temporal association rule format if it met the following criteria: 1) it is compatible with ACSPD data types, i.e. no physician notes or text and 2) it must contain unambiguous events. I compared the TARs discovered in the ACSPD with the performance measure derived rules to estimate compliance and corresponding patient outcomes.

5.6 Identifying Clinically Relevant Results using Interestingness Measures

Sequential pattern mining followed by temporal association rule data mining in clinical data will produce a substantial number of rules. Clearly the importance of those rules for clinical medicine practitioners is not uniform for rules. Having a domain expert, such as a clinician review all of the rules discovered via the data mining process for clinical relevance is impractical due to the volume of rules. Instead I apply a set of interestingness measures to cull the rules by ranking them according to a quantitative measure. (For a discussion of interestingness measures see chapter 4.) The idea is that rules which exhibit specific characteristics, such as a high frequency or strong correlation, are more clinically relevant. The challenge is to specify interestingness measures with characteristics that are congruent with the attributes a domain expert would deem essential for relevancy. In this dissertation I will use interestingness measures that focus on rule coverage, reliability, and rules that are unexpected based on domain expertise. I selected the traditional framework of support and confidence, as well as Bayes' factor, likelihood, lift, and Zhang's measure to assess rule interestingness.

5.7 Related Works

Electronic medical records contain information about the patient and treatment. The promises of EMR data are improvements in the areas of public health, pharmacovigilance, personalized care and more. The confluence of information technology, EMRs, natural language processing and data mining will forge the tools necessary to fulfill those promises.(Jensen, Jensen et al. 2012) And the integration of disparate data types will forge the techniques necessary to fulfill those promises. As EMR use increases in the US (especially with the passage of the HITECH Act), there is an opportunity to analyze EMR patient data to better understand the correlations between practice choices and outcomes.

STEMI treatment is time-dependent and exploring STEMI treatment in EMR data requires awareness of time and its effects on patient outcomes. EMR data are a record of the events during an encounter and often includes the date or date and time stamp and therefore lends itself to temporal ordering. The ordering of events in time, i.e. a sequence, carries important information that is lost if the same events are randomly arranged. In the clinical domain these sequences of events can represent the pathway of physician choices, patient preferences, and ultimately patient care. The goal of temporal data mining is to discover novel relationships between events ordered in time. Temporal data mining in EMR data is challenging because the data in EMRs can vary due to missing information or the system design. Furthermore, the data are collected asynchronously and the granularity can vary. Despite these challenges using EMR data, in the context of STEMI, the date and time information can be used to explore the temporal relationship between events and outcomes in the treatment of STEMI.

EMRs can incorporate medications, procedures, physician notes, laboratory tests, diagnoses, and vital sign monitoring, patient preferences and more. In addition to the various data contained in the EMR, researchers also use data that is external in conjunction with the EMR data to explore and understand healthcare. Hanauer et al used ICD9 codes and literature-based associations derived from abstracts to model temporal relationships in clinical diagnoses.(Hanauer and Ramakrishnan 2013) What makes Hanauer's and others' work interesting is that is used different types of data and data derived from disparate sources. In order to compare data from different sources, researchers employ a common representation, e.g. network, and/or analytic process, e.g. maximum entropy. Much like other inquiries into clinical medicine, the exploration of STEMI in EMRs requires use of data of different types and from

different sources. The choice of the analytic process is entirely driven by the importance of the temporal element in STEMI treatment. To gain new insights into STEMI treatment, we need to know not only what was done but when. Therefore, sequential pattern mining to discover order sets precedes temporal association rule mining to discover temporal relationships. Although a variation of sequential patterns and association rules has been used in computer science (Nkambou, Fournier-Viger et al. 2011) the combination is novel in clinical medicine. Sequential pattern mining and temporal association rule data mining are closely related. (See section 3.1) The difference is that sequential pattern mining explores frequent patterns in data and temporal association rule mining looks for a temporal relationship between events. To be clear, temporal rules are neither causal nor deterministic instead they imply a time-dependent relationship between elements.

To the best of my knowledge, using sequential pattern to discover *order sets* in EMR data is novel. On its own, sequential pattern mining in EMR data is neither novel nor widely implemented. For example, SPM discovered patterns of diagnoses codes across patient histories.(Patnaik, Butler et al. 2011) Where many SPM researchers focus on one aspect of the EMR what makes this implementation unique is the breath of data mined which included laboratory, medication, and procedure data.

Temporal rule association data mining methods, though abundantly applied in finance, economics, and marketing, are relatively new to health data and few studies use temporal association rule data mining.(Concaro, Sacchi et al. 2009, Somaraki, Broadbent et al. 2010) Recently, Sebastian et al. used association rule mining in medical texts and hospital derived data to discover novel insights and links in disparate medical hypothesis.(Sebastian and Then 2011) Sebastian's work asserts that the order of the items in an association rule is not important and therefore Sebastian does not consider the temporal dimension of treatment. In STEMI, the timing of treatment is central to optimal patient outcomes. Therefore the temporal relationship, which is often ignored by other researchers, is paramount.

5.8 Ethical Considerations

The ACSPD is derived from the HIPAA compliant HealthFacts® data warehouse. Patients are identified by a unique ID and the encounter dates have been randomly shifted within a narrow timeframe to protect patient identity. There is a risk, as with any data mining, that personal information can be inferred from

non-sensitive data. The sequential pattern mining and temporal association rule mining algorithms look for patterns and rules, and if a pattern or rule is rare there is a risk that person could be identified. In this dissertation, reported outcomes and models are based on aggregation of patient data and the risk of identification is minimal.

5.9 Findings

The ACSPD contains 25,186 STEMI encounters. From the STEMI encounters I created two groups, one group contained 1/3 of the encounters and the other 2/3, using simple random sampling. The first group had 8,397 encounters and 2,292,126 events and was used for SPM and TARM. The second group with 16,789 encounters and 4,526,200 events was used to validate the findings of TARM.

5.9.1 Order Set Discovery in ACSPD Using Sequential Pattern Mining

The SPM algorithm uses 3 user-defined thresholds, frequency, confidence and maximum pattern length. The SPM can mine patterns with a static or dynamic confidence threshold. (For a complete discussion of SPM, see chapter 2) To select the frequency and confidence thresholds, I used the SPM with several threshold values. The purpose is the find a frequency and confidence threshold that maximizes efficiency without sacrificing effectiveness. Performance evaluation of these two threshold types indicates that the static is more efficient at discovering known patterns in simulated data. (For a detailed explanation of SPM performance see chapter 2.) I used both the dynamic and static thresholds in the ACSPD data. The dynamic required a substantially higher frequency threshold and discovered considerably more order sets than the static. (See Figure 5.3 and Figure 5.4)

Initially, I used a dynamic confidence threshold. The frequency thresholds can be higher when using a dynamic threshold compared to a static confidence threshold. For the ACSPD, I found the execution time of the dynamic threshold was excessive when the frequency threshold was 6,000, meaning the algorithm efficiency was low. Therefore I only show results for minimum frequency of 6,000. The upper bound of algorithm effectiveness was 9,000. At a frequency threshold of 9,000 the SPM does not discover any order sets. I varied the dynamic confidence threshold from 0.70 to 1.0. The range that the SPM discovered order sets was 0.70 to 0.80. Below 0.70 the efficiency suffers and above 0.80 the effectiveness wanes. The sequential pattern mining algorithm using a

dynamic constraint discovered order sets when the frequency threshold was set between 6,000 and 9,000 and confidence threshold was 0.7-0.8. At a frequency threshold of 6000, the confidence thresholds of 0.7-0.8, yields approximately 5000 order sets. A frequency threshold of 7,000 yields approximately 6,000 order sets and a frequency threshold of 8,000 yields approximately 6,400 order sets. (See Figure 5.4) When the confidence threshold is adjusted to 0.85 for all frequencies, the discovered order sets decreases to less than 100. For each frequency threshold, the number of order sets decreases as the confidence threshold increases from 0.7 to 0.80. Therefore the dynamic constraint is effective when the frequency threshold is between 6,000 and 8,000 and the confidence threshold is adjusted to 0.7 to 0.8.

The static constraint is effective at lower frequency thresholds than the dynamic. (See section 2.10 To explore the static confidence threshold, I used a lower frequency threshold range than I did with the dynamic confidence threshold. With the static confidence threshold the frequency range was from 10 to 100. I varied the static confidence threshold from 0.70 to 1.0. In the ACSPD, the number of order sets discovered is greater for lower frequency thresholds. When the frequency threshold is 10, the number of order sets is 149 to 231 across the confidence thresholds. When the frequency threshold is 100, the number of order sets is 10 to 18 across the confidence thresholds. At lower frequency thresholds the differences in the number of discovered order sets by confidence threshold are more pronounced. When the frequency threshold is 10, the confidence thresholds of 0.75 and 0.8, 0.85 and 0.9, and 0.95 and 1.0 yielded the same number of order sets. As the frequency threshold increases, the number of discovered order sets decreases and the difference between the confidence thresholds diminishes. The SPM is more effective across a wider range of confidence values using the static confidence threshold than the dynamic. The SPM is more efficient using a static confidence threshold than the dynamic.

Based on the efficiency, the static constraint was used to discover order sets prior to temporal association rule mining. The frequency threshold of 40 and a confidence threshold of 0.9 discovered 39 order sets. (See Table 5.3) The discovered order sets appear in the ACSPD 46,450 times. (See Table 5.2) One of the advantages of using EMR data are that the events are time stamped. Therefore we know in what month and year each order set occurs. When we consider all the order sets over the

9 year span of the ACSPD, we can see that the use of the order sets changes over time. The use of order sets increases from 2000 to the zenith in late 2001. (See Figure 5.5 and Figure 5.6) Then the frequency of the order sets decreases with the nadir being in June 2004. The frequency of the order sets increases after 2004 but do not reach the previous highs seen early years. All 39 order sets are not present for the entire span of the database (2000 – 2009). In the year 2000, 67% (n = 26) of the order sets are present and the remaining 13 first appear after 2000. In 2005 there is a spike in the number of new order sets (See Figure 5.6). In late 2004 the AHA and ACC published a new set of STEMI treatment guidelines, which updated the previous one published in 1999. The spike in new order sets in 2005 could be a result of the introduction and implementation of the new guidelines.

Table 5.2 Order Sets by Time from Admission

Time from admission (hours)	Order sets (n)
< 1	5,608
1	6,902
2	2,138
3	1,291
4	867
5	711
6	612
>6	28,321
Total	46,450

Table 5.3 Order Set ID and Constituent Events

Order set ID	Events in order set
711199	Heparin; Heparin*
170199	Left heart cardiac catheterization*; Angiocardiography of left heart structures*; Coronary arteriography using two catheters*; Insertion of non-drug-eluting coronary artery stent(s)*; Insertion of one vascular stent*; Percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy*
665399	Nitroglycerin; Nitroglycerin*
283799	Left heart cardiac catheterization*; Angiocardiography of left heart structures*; Insertion of non-drug-eluting coronary artery stent(s)*; Procedure on single vessel*
437899	PTT*; PT*; BUN*; CKMB*; ALT/SGPT*; AST/SGOT*; Creatinine, serum*; PT INR*; Potassium, serum*; Magnesium*; Calcium, serum*; CO2*; HGB*; HCT*; WBC*; RBC*; Glucose, serum*; Phosphorus, serum*; Platelet count*; ALK PHOS, serum dup*; Bilirubin total*; CL, body fluid*
467499	Left heart cardiac catheterization*; Angiocardiography of left heart structures*; Coronary arteriography using a single catheter*
844799	Left heart cardiac catheterization*; Angiocardiography of left heart structures*; Coronary arteriography using two catheters*; Injection or infusion of platelet inhibitor*
51799	Left heart cardiac catheterization*; Angiocardiography of left heart structures*; Coronary arteriography using two catheters*; Multiple vessel percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy performed during the same operation, with or without mention of thrombolytic agent*
311799	Insertion of temporary transvenous pacemaker system*; Left heart cardiac catheterization*
852199	PTT*; PT*; BUN*; CK/CPK, total, serum*; ALT/SGPT*; AST/SGOT*; Creatinine, serum*; Potassium, serum*
565899	Heparin*; Heparin
404999	Morphine*; Morphine
872799	Left heart cardiac catheterization*; Single vessel percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy without mention of thrombolytic agent*; Insertion of non-drug-eluting coronary artery stent(s)*
636499	PTT*; PT*; BUN*; ALT/SGPT*; AST/SGOT*; Creatinine, serum*
100699	Left heart cardiac catheterization*; Insertion of non-drug-eluting coronary artery stent(s)*
760499	Left heart cardiac catheterization*; Angiocardiography of left heart structures*; Coronary arteriography using two catheters*; Insertion of drug-eluting coronary artery stent(s)*; Insertion of two vascular stents*; Percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy*
813999	Inject/INF thrombo agent*; PTT*; PT*; BUN*
297199	Implant of pulsation balloon*; Left heart cardiac catheterization*; Angiocardiography of left heart structures*; Coronary arteriography using two catheters*
317199	Morphine; Morphine*
451499	Left heart cardiac catheterization*; Angiocardiography of left heart structures*; Coronary arteriography using two catheters*; Single vessel percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy without mention of thrombolytic agent*; Insertion of non-drug-eluting coronary artery stent(s)*; Injection or infusion of platelet inhibitor*
653999	PTT*; PT*; BUN*; Troponin*; Creatinine, serum*; Potassium, serum*
252699	Left heart cardiac catheterization*; Angiocardiography of left heart structures*; Single vessel percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy without mention of thrombolytic agent*
709599	Nitroglycerin*; Nitroglycerin
751499	Aspirin; Aspirin*
949499	PTT*; PT*; BUN*; CK/CPK, total, serum*; Troponin*
850299	Left heart cardiac catheterization*; Coronary arteriography using two catheters*; Single vessel percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy without mention of thrombolytic agent*; Insertion of non-drug-eluting coronary artery stent(s)*; Injection or infusion of platelet inhibitor*
500499	Left heart cardiac catheterization*; Angiocardiography of left heart structures*; Coronary arteriography using two catheters*; Single vessel percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy without mention of thrombolytic agent*; Injection or infusion of platelet inhibitor*; Insertion of drug-eluting coronary artery stent(s)*
181699	Aspirin*; Aspirin
333399	PTT*; PT*; BUN*; CK/CPK, total, serum*; ALT/SGPT*; AST/SGOT*; Creatinine, serum*; PT INR*; Potassium, serum*; Albumin, serum*; ALK PHOS, serum*; Calcium, serum*; CL, serum*; HGB*; HCT*; WBC*; RBC*; Sodium, serum*; Protein total, serum*; Bilirubin total*; CKMB dup*; Troponin I*; Anion gap, blood*; Baso*; Baso abs*; BUN creat ratio*; CK calc*; CK rel index*
742599	BUN*; Creatinine, serum*; Potassium, serum*
215499	Insertion of endotracheal tube*; Continuous invasive mechanical ventilation for less than 96 consecutive hours*
63599	Left heart cardiac catheterization*; Injection or infusion of platelet inhibitor*; Insertion of drug-eluting coronary artery stent(s)*; Procedure on single vessel*
254899	Implant of pulsation balloon*; Left heart cardiac catheterization*; Coronary arteriography using two catheters*
412599	Aortography*; Left heart cardiac catheterization*; Angiocardiography of left heart structures*; Coronary arteriography using two catheters*; Single vessel percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy without mention of thrombolytic agent*
808999	Left heart cardiac catheterization*; Angiocardiography of left heart structures*; Coronary arteriography using two catheters*; Insertion of drug-eluting coronary artery stent(s)*; Insertion of one vascular stent*; Percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy*
255299	Left heart cardiac catheterization*; Angiocardiography of left heart structures*; Insertion of drug-eluting coronary artery stent(s)*; Procedure on single vessel*
781799	Left heart cardiac catheterization*; Angiocardiography of left heart structures*; Coronary arteriography using two catheters*; Insertion of drug-eluting coronary artery stent(s)*; Percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy*
390199	Left heart cardiac catheterization*; Coronary arteriography using two catheters*; Insertion of non-drug-eluting coronary artery stent(s)*; Procedure on single vessel*; Insertion of one vascular stent*; Percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy*
282899	Left heart cardiac catheterization*; Angiocardiography of left heart structures*; Coronary arteriography using two catheters*; Single vessel percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy without mention of thrombolytic agent*; Insertion of drug-eluting coronary artery stent(s)*

* Event start

Figure 5.4 Number of Order Sets by Frequency Using Static Confidence Threshold in the ACSPD

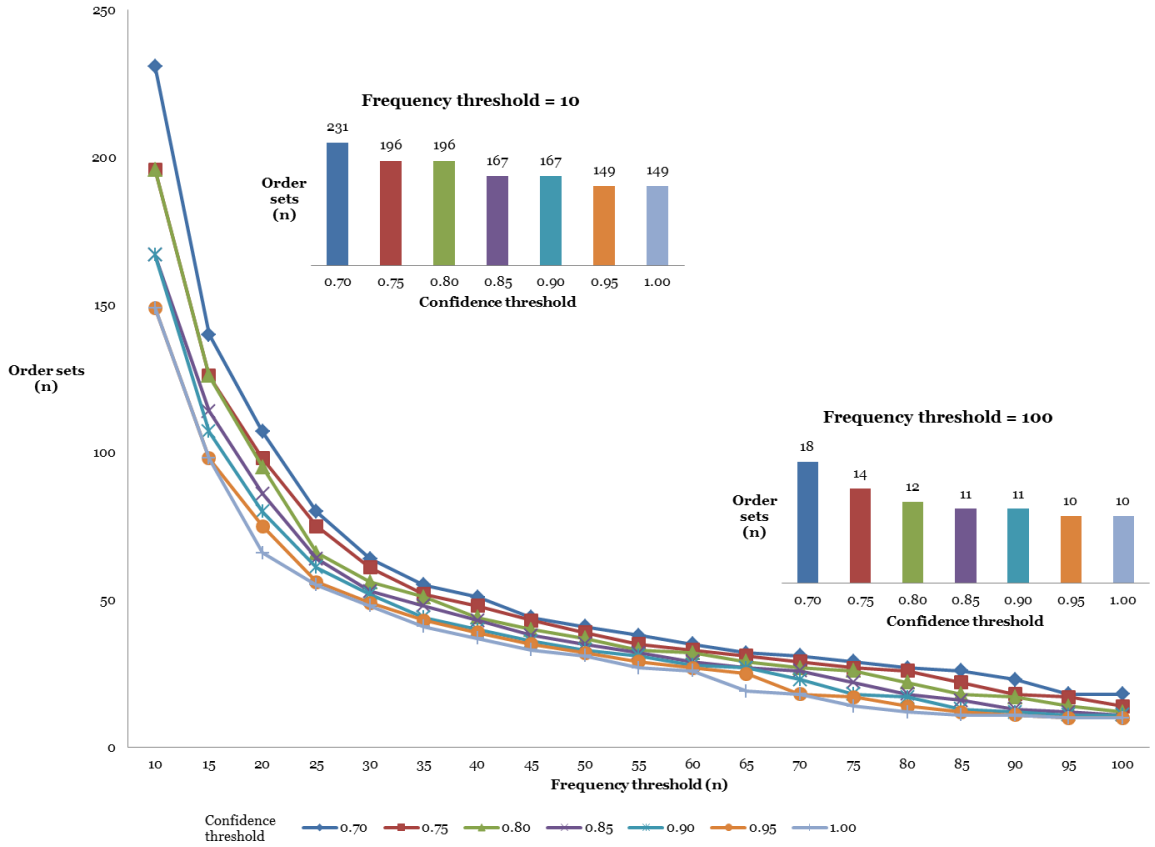


Figure 5.3 Number of Order Sets by Frequency Using Dynamic Confidence Thresholds in the ACSPD

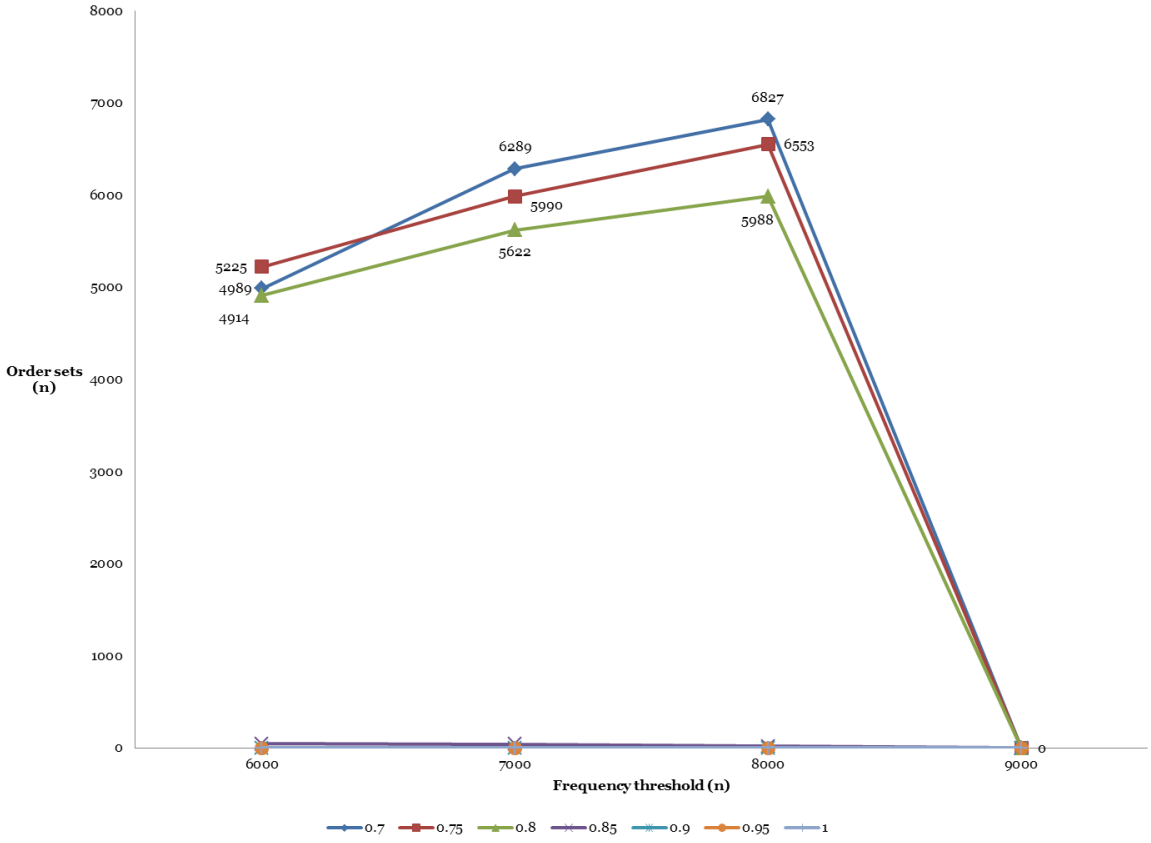


Figure 5.5 ACSPD Discovered Order Set Frequency by Year

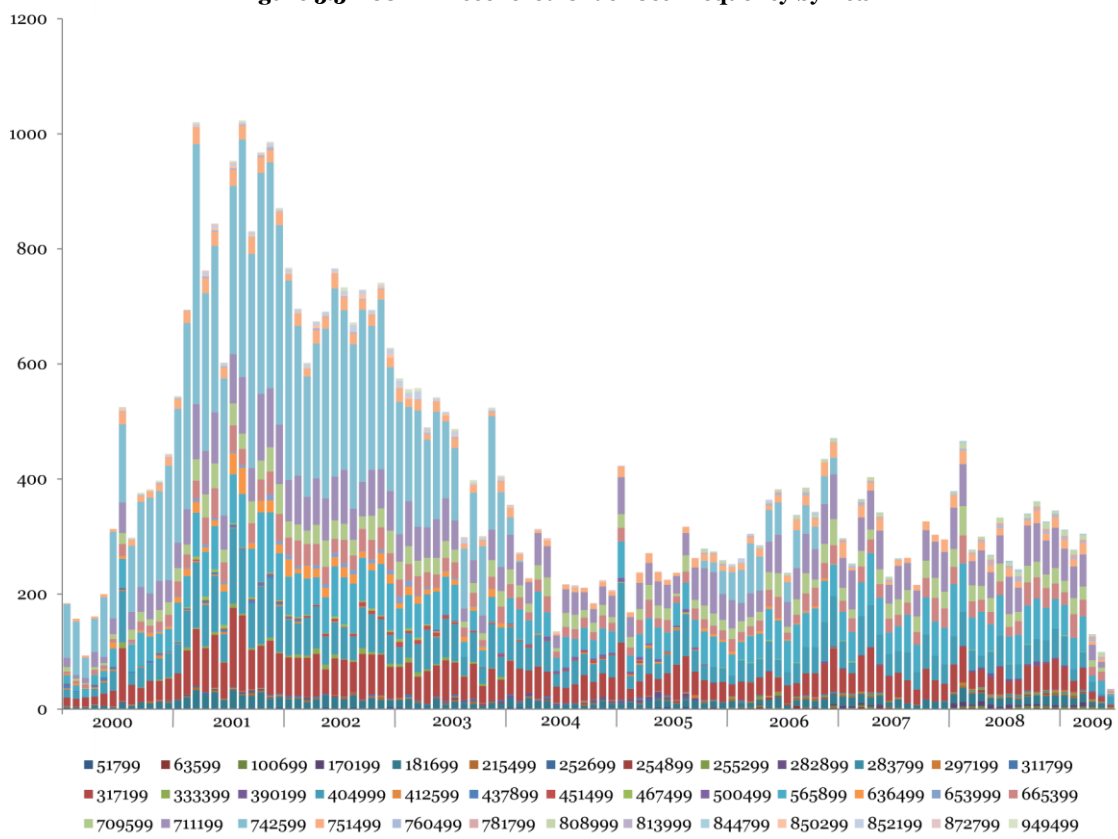
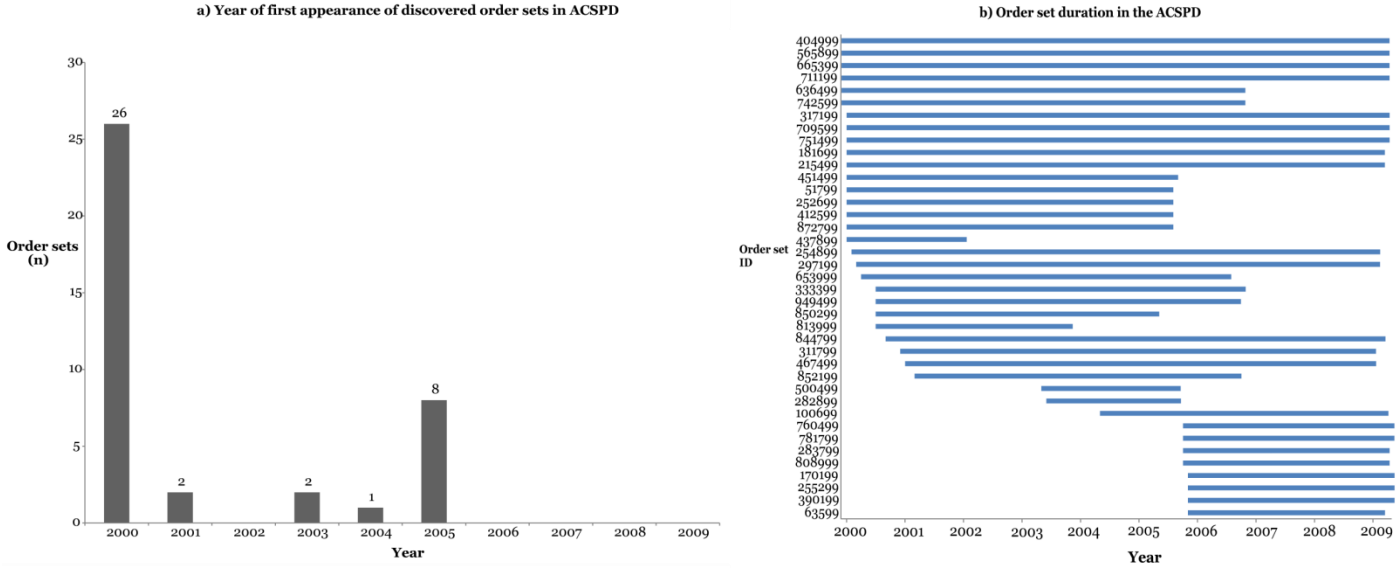


Figure 5.6 Discovered Order Sets by Appearance and Duration in the ACSPD



5.9.2 TARM in ACSPD

I used TARM to elucidate the temporal relationships between events that occur during the treatment of STEMI. Before applying the TARM algorithm to the ACSPD, I used the sequential pattern mining algorithm to identify and consolidate order sets. The ACSPD is an extremely large database of EMR encounters. For this exploration of practice patterns, due to resource limitations I used a third of the STEMI patient encounters in the ACSPD. Despite this reduction in data, the TARM algorithm produced in excess of 1 trillion rules. Therefore, I used a random sample (n = 1,000,000) of the discovered rules for these analyses. TARs represent clinical practice patterns and when discussing the findings of the TARM, I will refer to the TAR as clinical practice patterns and rules as the context dictates.

Table 5.4 Summary of Interestingness Measures for Practice Patterns Discovered Using TARM

Interestingness Measure	Minimum	Maximum	Mean	Standard Deviation	Count*
Bayes	0	999200	3157.194	22551	980069
Confidence	0.00521	1	0.474814	0.371447	980069
Lift	5.164	999200	6941.965	36882.06	980069
Likelihood	-7.949	0	-5.92142	1.493115	980069
Support	0	0.00002	0	1.22E-07	980069
Zhang	0.6413	1	0.994812	0.010725	980069

* Count is the number of unique practice patterns

Clinical practice patterns for STEMI treatment are highly variable. When we examine clinical practice pattern frequency, we can see that the majority of practice patterns are rare, with many occurring only once (sample size = 1,000,000 with 980,069 unique patterns). In this context of practice pattern scarcity, equating practice pattern frequency with importance renders all practice patterns irrelevant. For this reason, I used the interestingness measures of confidence, Bayes factor (BF), lift, likelihood and Zhang's which are metrics based on attributes other than frequency. (For a discussion of interestingness measure attributes see 4.2

The rules discovered by the TARM algorithm in the ACSPD are scarce and the reliability is weak. Rule coverage or support is the proportion of rules that fit a specific pattern. The vast majority of rules are rare (support mean= 0.0, standard deviation, SD = 0.0). Along with rarity of rules is the somewhat ambivalent relationship between the antecedent and consequent (confidence mean = 0.5, SD = 0.37). The conditional probability of the consequent following the antecedent is equivalent to the probability of the consequent not following the antecedent. Conversely if we consider the

consequents of the rules, the negative mean likelihood demonstrates that a consequent is more likely to *not* follow a given antecedent (likelihood mean = -5.92, SD = 1.49). Zhang's measure affirms and expands these findings. The mean Zhang's measure is 0.99 and the standard deviation is small (SD = 0.01). Zhang's measure will approach 1.0 when there are few counter examples (rules with neither the antecedent nor consequent) compared to the number of examples (rule support). However when rule support is extremely low and the number of counter examples is very large, as it is in this data, Zhang's measure will also approach 1. BF is a measure of the evidence provided by the data for the relationship of the consequent given the antecedent or antecedent/consequent independence. The data supports the existence of an antecedent-consequent relationship (BF mean = 3157, SD = 22551). What is interesting is that the relationship indicated by the BF most likely is the counter example relationship rather than the rule. Simply put, there is a significant rule diversity leading to low coverage (support) and weak reliability (confidence, lift, likelihood, and Zhang's). In terms of clinical treatment of STEMI, there is a wide variation in clinical practice patterns. There are temporal relationships between events but those relationships are weak.

Table 5.5 Top 5 Clinical Practice Patterns Ranked by Interestingness Measure

Interestingness Measure	Ranking	Clinical Practice Pattern	Value	Count
Bayes	1	LVP solution end, lepirudin end, lepirudin end	5612	5
	2	sodium chloride start, methylprednisolone start, sodium chloride end	2928	4
	3	LVP solution end, lepirudin end, lepirudin end	1871	6
	4	sodium chloride start, sodium chloride end, hydroxyzine end	1839	4
	5	Midazolam start, albuterol start, albuterol start	1774	4
Confidence	1	Granisetron end, albuterol end, albuterol end	1	5
	2	sodium chloride start, methylprednisolone start, sodium chloride end	0.6667	4
	3	Platelet Count end, Basophile % start, Mean platelet volume start	0.5714	4
	4	565899, potassium chloride start, LVP solution end	0.5714	4
	5	LVP solution end, lepirudin end, lepirudin end	0.5	5
Lift	1	LVP solution end , lepirudin end, lepirudin end	2807	5
	2	sodium chloride start, sodium chloride end, hydroxyzine end	1710	4
	3	Granisetron end, albuterol end, albuterol end	1674	5
	4	LVP solution end, lepirudin start, lepirudin end	1403	6
	5	insulin regular end, insulin regular end, insulin regular end	1269	4
Likelihood	1	LVP solution start, LVP solution start, lepirudin end	-3.356	6
	2	LVP solution start, LVP solution start, calcium gluconate end	-3.266	5
	3	LVP solution start, LVP solution end, calcium gluconate start	-3.239	6
	4	LVP solution start, LVP solution end, lepirudin start	-3.142	7
	5	sodium chloride start, sodium chloride end, hydroxyzine end	-2.225	4
Support	1	LVP solution start, LVP solution start, LVP solution start	0.00002	18
	2	LVP solution start, LVP solution start, LVP solution end	0.00001	13
	3	Metoprolol start, insulin regular end, insulin regular start	0.00001	7
	4	insulin regular start, insulin regular end, insulin regular start	0.00001	5
	5	insulin regular start, insulin regular end, insulin regular end	0.00001	6
Zhang's	1	LVP solution end, lepirudin end, lepirudin end	0.9982	5
	2	sodium chloride start, sodium chloride end, hydroxyzine end	0.9977	4
	3	Granisetron end, albuterol end, albuterol end	0.9970	5
	4	insulin regular end, insulin regular end, insulin regular end	0.9969	4
	5	LVP solution end , lepirudin start , lepirudin start	0.9959	4

Clearly there is a wide variation in the interestingness measure values. To find the more interesting rules, the rules are ranked by the measure value. In this data, the frequency of the rules is low; often there are only one or two occurrences. When the frequency is low, the value of the interestingness measure tends toward the extreme of the measure. For example, the confidence will be 1 and the support almost 0. Distinguishing between the rules becomes problematic. Using a lower bound frequency prior to ranking rules by interestingness measure can force a more intuitive ranking by eliminating the low frequency rules with extreme measure values. With a lower bound frequency of ≥ 4, the copious number of rules with extreme interestingness values is mitigated. The top 5 rules for BF, confidence, lift, likelihood, support and Zhang's are shown in Table 5.5.

When we look at the top clinical practice patterns based on potential user interest, we can see there are common events. The majority of the top ranked rules include large volume parenteral (LVP) solution, sodium chloride, potassium chloride, regular insulin, heparin, metoprolol, lepirudin and albuterol. LVP solutions are intravenous (IV) solutions used to deliver IV medications. Sodium chloride is commonly used to flush IV catheters. Potassium chloride is used to increase the potassium levels. Albuterol treats bronchial spasms. Albuterol is known to cause cardiovascular changes and aggravation of diabetes. Regular insulin, the treatment for diabetes, is a common element in the top discovered rules. The AHA/ACC STEMI guidelines contain specific recommendations for treatment of diabetic patients regarding disease management and secondary prevention via lifestyle changes, and as one deciding factor for the use of aldosterone blockade. (Antman, Anbe et al. 2004, Antman, Hand et al. 2008) Heparin is used to prevent clotting that can result when the body repairs damage to cardiac tissue. Lepirudin is used to treat coagulation as a result of an immune response to heparin. The order set discovered by sequential pattern mining, order set identification number 565899, encodes the administration of heparin. Finally, metoprolol is a treatment for high blood pressure and is shown to increase survival after a myocardial infarction.

The medications in the top ranked rules by interestingness measures that are not related to the treatment of STEMI or cardiac issues are methylprednisolone, granisetron, hydroxyzine and midazolam. Methylprednisolone, a corticosteroid, reduces inflammation. Granisetron is for nausea

from cancer treatment whereas hydroxyzine treats nausea caused by a number of conditions. Midazolam is for sedation prior to surgical procedure.

The laboratory tests included in the top 5 rules ranked by interestingness measure are platelet count, mean platelet volume and percentage of basophiles. Blood contains white cells, red cells, and platelets. Platelets are essential in clot formation. Basophiles are white blood cells that participate in the immune response. An increase in basophiles is an indicator of infection. The mean platelet volume is a measurement of the platelet size. It is included in the complete blood count, a standard set of laboratory tests which commonly incorporates white, red, and platelet measurements.

Now let's consider each interestingness measure, starting with BF. (See Table 5.5) BF is 5612 and the rule count is 5 for the highest ranked rule, *LVP solution end, lepirudin end, lepirudin end*. This implies that the numerator, $n_{AB}n_{\bar{B}}$, is greater than the denominator, $n_{A\bar{B}}n_{B}$, where A and B represent the antecedent (*LVP solution end, lepirudin end*) and consequent (*lepirudin end*), respectively. Since the count (n= 5) is small, the value of the numerator and therefore the BF is driven by $n_{\bar{B}}$, the number of consequents that are not lepirudin. The BF shows that there is more evidence for a relationship between the antecedent and consequent in the data than for independence. But this is more a statement about the counter example than the rule because of the low rule count compared to the number of antecedents and consequents. The same can be said for all rules in the top 5 for BF.

For confidence, the conditional probability of the consequent given the antecedent, the highest value is 1. When the antecedent *granisetron end, albuterol end* appears it is followed by only one consequent, *albuterol end*. The confidence decreases almost exponentially to 67% for the 2nd rule, and to 50% for the 5th rule. At 0.5, the confidence in the consequent given the antecedent is no better than a coin toss. The confidence ranking does highlight a rule with all three events are laboratory tests (3rd) and a rule with an order set discovered using the sequential pattern mining algorithm. In the ACSPD, the swift decline in the confidence value of the top ranked rules and the low rule frequency demonstrates that if we know the antecedent of the rule, it would be difficult to identify the consequent.

Lift is an easily interpreted measure of independence between the antecedent and the consequent. A large lift value means that the consequent follows the antecedent more often than is expected by chance alone. At first glance, the large lift values would seem to contradict the low confidence values of the rules. The 1st rule's lift is 2807. The consequent, *lepirudin end*, following the antecedent, *LVP solution end, lepirudin end*, is much larger than would be expected by chance but the confidence, the conditional probability of the consequent, is only 50% (confidence rank 5th). Lift is sensitive to rule set size and the consequent frequency. The rule set is very large, n= 980,069 unique rules in the 1,000,000 sampled. As the rule set size increases the lift will increase and as the frequency of the consequent decreases, the lift will increase. (For the lift equation, see Table 4.3) The frequency of the consequent *lepirudin end* is very small compared to the size of the rule set. Therefore the large lift value is more a function of the size and heterogeneity of the rules than the relationship between the antecedent and the consequent, which is weak.

The likelihood measures the amount of evidence (probability) to support the rule compared to the counter example (antecedent not followed by the consequent of interest). A negative likelihood means that there is more support for the counter example than the rule. Consider the 1st ranked rule, *LVP solution start, LVP solution start, lepirudin end*, it has a likelihood of -3.4. From the other interestingness measure ranked rules we can see that *calcium gluconate end* (n=5), *LVP solution start* (n=18), and *LVP solution end* (n=13) are consequents to the antecedent *LVP solution start, LVP solution start*. These consequences are only from the top 5 rules by interestingness measure; the entire sample could contain more consequences. The probability of the other the consequents that are not *lepirudin end* is greater than the probability of *lepirudin end*. Therefore the likelihood is negative. The 2nd – 5th top ranked rules similarly to the 1st have negative likelihood values.

In the ACSPD, Zhang's measure does not behave as anticipated. Theoretically Zhang's will approach its maximum when the number of counter examples approaches 0. It is -1 when the conditional probability of the consequent given the antecedent is low. The 1st rule by Zhang's is *LVP solution end, lepirudin end, lepirudin end*. The number of counter examples should be relatively low. The BF of 5612 and the lift of 2807 tell a different narrative, namely that there are more counter examples than examples. Therefore, the measure should not be approaching 1. The measure value is, in fact,

0.9982. Zhang's measure was designed to address the shortcomings of other interestingness measures. Unfortunately, in an environment with rare rules and significant heterogeneity, Zhang's measure does not perform as anticipated.

The interestingness measures for temporal association rules were created to differentiate between rules that would be interesting to users and rules that are not important to users. Temporal association rule data mining produces an enormous number of rules which necessitates using filtering methods like interestingness measures. The assumption is that the discovered rules will have substantial coverage in the rule set. The ACSPD specifically and EMR data in general has a substantial number of events and event types that can be used to generate temporal association rules. Furthermore, the structure of the temporal association rules, 2 antecedents and 1 consequent engenders low rule frequency by increasing the number of possible permutations. Clearly, from these results, the interestingness measures of BF, confidence, lift, support and Zhang's do not perform well under these conditions and can yield contradictory findings.

Clinical practice patterns are widely varied and unique. Clinical patterns share common elements (events) but rarely are the entire patterns repeated. The interestingness measures show more evidence for counter examples than for the clinical pattern. Also the internal relationship between the events in the clinical practice pattern is weak.

5.9.3 TAR Compliance with AHA/ACC STEMI Performance Measures and Outcomes

The AHA/ACC published performance measures based on the STEMI guidelines to aid healthcare institutions in guideline adherence. (See Table 5.6) The STEMI clinical practice guidelines are a compilation of evidence-based medicine and expert consensus on the best practices for optimal patient outcomes. I translated the AHA/ACC performance measures into rule representation similar to the TAR in the previous section. The AHA/ACC performance measures included unambiguous events and temporal constraints. The performance measures govern patient care from admission to discharge. At admission the performance measures focus on administration of aspirin (PM1), time to diagnosis (PM7) and time to treatment (PM8). One measure, the documentation of low-density lipoprotein (LDL) test is not time dependent (PMt1). The remaining performance measures are completed at discharge. The TARs are, in this context, clinical practice patterns. Since we know the

outcome for each patient whose treatment is described by a particular pattern, we can estimate the risk of mortality and bleeding events. The translated performance measures are the expert treatment pattern. The question then is "Does compliance with performance measures decrease the risk of suboptimal outcome?" To address this inquiry, each TAR is compared to each of the performance measure for compliance.

Compliance is binary; "compliant" versus "noncompliant". There is considerable research investigating the issues of compliance and it suggests compliance is more accurately assessed as a continuum rather than as an either/or proposition. For this dissertation, the all or nothing approach is adequate because the explicit design of the performance measures, i.e. "AMI patients who are prescribed beta blockers at arrival", do not easily lend themselves to assessment on a continuum.

In order to determine if the risk of suboptimal outcome is modulated by compliance, I used a comparative risk measure, the odds ratio. The odds are the number of clinical practice patterns associated with the outcome divided by the number of clinical practice patterns not associated with the outcome. There are 4,903 practice patterns compliant with an aspirin at admission (PM1), of those the patient subsequently died in 237 and 4,666 where the patient survived. The odds of mortality for compliance with PM1 is 237/4,666 = 0.05 or 5 in 100. The odds ratio, aptly named, is the ratio of odds. The odds ratio of mortality is the odds of compliance divided by the odds of noncompliance. The odds ratio (OR) is bounded by 0 to infinity. An OR of 1 means the odds (risk) is equivalent for compliance and noncompliance. An odds ratio greater than 1 is interpreted as an increased risk of mortality. An OR of 1.49 means that the odds of mortality for compliance are 49 times greater than the odds of mortality for noncompliance. An OR of less than 1 is not as easily interpreted. The odds of mortality are less for compliance than for noncompliance, i.e. the patient's risk of mortality decreases if practice behavior is compliant. But the amount is not readily understood from the OR. For ease of interpretation, the inverse of the OR converts it to an OR greater than one. However the inverse OR is the odds of mortality for noncompliance divided by compliance. The inverse of the OR of 0.23 for PMt1 becomes the odds of mortality for noncompliance versus the odds of mortality for compliance with an OR of 1/0.23 or 4.55. For each OR, the confidence limit is also given. All confidence limits are 95%. (See Table 5.7 and Table 5.8)

Table 5.6 AHA/ACC STEMI Performance Measures

PM	STEMI Performance Measure
1	AMI patients who received aspirin within 24 hours before or after hospital arrival
2	AMI patients who are prescribed aspirin at hospital discharge
3a	AMI patients who are prescribed beta blockers at arrival
3b	AMI patients who are prescribed a beta-blocker at hospital discharge
4	AMI patients who are prescribed a statin at discharge
6	AMI patients with left ventricular systolic dysfunction (LVSD) who are prescribed an ACEI or ARB at hospital discharge.
7	AMI patients with STEMI or left bundle branch block (LBBB) on the ECG closest to arrival time receiving fibrinolytic therapy during the hospital stay and having a time from hospital arrival to fibrinolysis of 30min or less
8	AMI patients with STEMI or LBBB on the ECG closest to the arrival time receiving primary PCI during the hospital stay with a time from hospital arrival to PCI of 90 minute or less.
T-1	AMI patients with documentation of low-density lipoprotein-cholesterol (LDL-C) level in the hospital record or documentation that LDC-C testing was done during the hospital stay or is planned for after discharge.
T-9	Medically treated AMI patients who are prescribed clopidogrel or ticlopidine at hospital discharge

So far, in the quest to address the query of "Does compliance with performance measures decrease the risk of suboptimal outcome?" we have practice patterns (TARs), expert knowledge (PM) and a comparison measure (OR). There is the issue of what it means to be compliant and noncompliant. Compliance of a rule, in this context, means that the rule has the same events and if necessary conforms to the time constraint of the performance measure. (For the rule representation of the PMs, see Table 4.6) For example, PM1 is administration of aspirin within 24 hours of admission. If a TAR has an aspirin order within the first 24 hours, then the rule is in compliance with the PM. Each rule in the sample was compared to each performance measure for compliance. It should be noted that a single practice pattern can be in compliance with more than one measure, 1 measure, or no performance measures. A noncompliant rule would be a rule that does not have the same events or conform to the time constraint of the performance measure.

There are two types of noncompliance, specific and general. When a practice pattern is noncompliant with the performance measure of interest, it is specific noncompliance. If a rule is not compliant with any performance measure, it is generally noncompliant. These two noncompliance types form two noncompliant groups; 1) all rules that were not compliant with the performance measure of interest (specific) and 2) only those rules that are noncompliant with all the performance measures (general). This allows for exploration of performance measure compliance risk with and without the effect of compliance with the remaining measures.

There are demographics and hospital characteristics in the ACSPD that affect outcomes. For each clinical practice pattern, the TARM also captured demographics, hospital characteristics, and

treatment as well as the outcomes of interest. The outcomes of interest are mortality and bleeding events. Results from previous outcomes analyses in the ACSPD (see Appendix), demonstrate that age, gender, the presence of at least one comorbid condition, primary reperfusion treatment (PCI, CAGB, or medication), the size of the hospital, presence of a full catheterization laboratory for the treatment of MI patients, and the census region of the hospital are associated with mortality and bleeding events. (See Appendix, all p-values < 0.5). However due to missing data and complete separation of data, age and primary reperfusion treatment were not included as covariates. The unadjusted p-value comes from the χ^2 and does not account for covariates known to be associated with the outcome. The significance level for the p-value is 0.05. The adjusted analyses, which account for demographics and hospital characteristics, are generalized estimating equations with a logit link function. The risk of outcome is reported as the odds ratio (OR) with a 95% confidence interval. (See Table 5.8 and Table 5.9) I would anticipate, (possibly naively) that clinical practice patterns which are compliant with performance measures will have fewer deaths and bleeding events.

When we consider compliance with performance measures, we can see that the frequency of compliant clinical practice patterns varies widely across the performance measures, from 7 to 484,836. (See Table 5.8) The documentation of low-density lipoprotein (LDL) test during hospitalization (PMt1) has the most compliant clinical practice patterns. PMt9, the prescription of clopidogrel or ticlopidine at discharge for medically treated patients, had the lowest number of compliant practice patterns. The mortality for compliant clinical practice patterns ranged from 0% for PMt9 to 21% for PM7.

First, we will consider if mortality is associated with clinical practice pattern compliance, without regard to any other features known to affect mortality or compliance with other measures. The comparison group is the specific noncompliant clinical practice patterns and the results are the unadjusted χ^2 in Table 5.8 for noncompliant with performance measure. Clinical patterns that are in compliance with PM1, PM2, PM3a, PM3b, PM4, PM6, PM7, or PMt1 are associated with mortality. The unadjusted χ^2 is symmetric; therefore it cannot identify the direction of the association. The risk of mortality for noncompliant clinical practice patterns is approximately 6% for all performance

measures except PMt1 which is 9%. The risk of mortality for compliant practice patterns is greater than the noncompliant practice patterns for aspirin discharge (PM2 compliant vs. noncompliant 8.4% vs. 5.8%, p-value <0.0001), beta-blocker at discharge (PM3b 10.4% vs. 5.8%, p-value <0.001), time from admission to fibrinolytic therapy (PM7 20.8% vs. 5.8%, p-value <0.0001). The risk of mortality for compliant practice patterns is less than the mortality risk for noncompliant clinical practice patterns for aspirin at admission (PM1 4.8% vs. 5.8%, p-value 0.004), beta-blocker at admission (PM3a 3.8% vs. 5.8%, p-value <0.001), statin at discharge (PM4 5.2% vs. 5.8%, p-value 0.03), ACEI or ARB at discharge (PM6 4.7% vs. 5.8%, p-value 0.001), and LDL test during the encounter (PMt1 1.9% vs. 9.4%, p-value <0.0001). It should be noted that the time from admission to PCI (PM8 6.9% vs. 5.8%, p-value 0.37) and clopidogrel or ticlopidine at discharge (PMt9 0% vs. 5.8%, p-value 0.51) are not significant. (See Table 5.8, unadjusted p-value) Without regard to factors known to influence mortality and compared to the specific noncompliance group, compliance with the performance measures can increase or decrease patient mortality risk depending on the measure.

When the patient and hospital characteristics that influence mortality are included in the mortality risk estimates for clinical practice patterns, the only associations that remain significant are compliance with beta-blocker at admission (PM3a), beta-blocker at discharge (PM3b), statin at discharge (PM4), ACEI or ARB at discharge (PM6), and time to fibrinolytic therapy within 30 minutes (PM7). (See Table 5.8, adjusted p-values ≤ 0.05) Compliance with beta-blocker at discharge (PM3b OR 1.49, 95%CI 1.39-1.59) and fibrinolytic therapy (PM7 OR 4.01, 95%CI 2.52-6.65) increases the risk of mortality. Practice patterns in compliance with administering beta-blocker at admission (PM3a OR 0.60, 95%CI 0.54-0.68), statin at discharge (PM4 OR 0.71, 95%CI 0.62-0.81), and ACEI or ARB at discharge (PM6 OR 0.57, 95%CI 0.49-0.66), and LDL testing (PMt1 OR 0.23, 95%CI 0.22-0.23), even when adjusted for other factors, decreases the risk of mortality compared to noncompliance practice patterns. (See Table 5.8, odds ratio)

Patient mortality is associated with clinical practice pattern compliance with some of the STEMI performance measures, when the comparison group is specific noncompliant. Compliance with beta-blocker at discharge and fibrinolytic therapy increase the risk of mortality, and compliance with beta-

blocker at admission, statin at discharge, ACEI or ARB at discharge, and LDL monitoring decrease the risk of mortality.

There are 481,486 practice patterns discovered by TARM that are not compliant with any performance measure, i.e. general noncompliance. In 9.5% of these noncompliant practice patterns, the patient subsequently died. (See Table 5.8, Noncompliant with All Performance Measures) When we don't consider factors known to influence mortality, practice patterns that are compliant with aspirin in at admission (PM1 4.8% vs. 9.5%, p-value <0.0001), beta-blocker at admission (PM3a 3.8% vs. 9.5%, p-value <0.0001), statin at discharge (PM4 5.2% vs. 9.5%, p-value <0.0001), ACEI or ARB at discharge (PM 4.7% vs. 9.5%, p-value < 0.0001), and LDL monitoring (PMt1 1.9% vs. 9.5%, p-value <0.0001) are associated with a decrease in mortality, whereas beta-blocker at discharge (PM3b 10.3% vs. 9.5%, p-value=0.02) and fibrinolytic treatment within 30 minutes (20.8% vs. 9.5%, p-value <0.0001) are associated with an increase in mortality. (See Table 5.8, Noncompliant with All Performance Measures, unadjusted p-value). When we do consider other mortality risk factors, compliance with aspirin at admission (PM1 OR 0.54, 95%CI 0.46-0.62), aspirin at discharge (PM2 OR 0.68, 95%CI 0.60-0.76) beta-blocker at admission (PM3a OR 0.36, 95%CI 0.32-0.41), statin at discharge (PM4 OR 0.43, 95%CI 0.38-0.49) and fibrinolytic therapy (PM7 OR 0.36, 95%CI 0.31-0.41) decrease mortality. (See Table 5.8, Noncompliant with All Performance Measures, adjusted p-value and odds ratio) Compliance with beta-blocker at discharge (PM3b p-value = 0.14) or time from admission to PCI (PM8 p-value = 0.07) has no statistically significant effect on patient mortality when adjusted for other mortality risk factors.

In order to better understand the impact of compliance on patient mortality, consider the inverse comparison that is the impact of noncompliance on patient mortality. (See Table 5.7) It should be noted that the noncompliant group is comprised of practice patterns that are noncompliant with all performance measures (general noncompliance) and the ORs are adjusted for mortality risk factors. As expected, based on the previous results, beta-blocker at discharge (PM3b OR 1.05, 95%CI 0.98-1.12) time to PCI (PM8 OR 1.47, 95%CI 0.97-2.17) and clopidogrel at discharge (PMt9 OR could not be calculated) are not associated with mortality risk. Noncompliance with beta-blocker at admission (PM3a OR 2.78, 95%CI 2.44-3.13), statin at discharge (PM4 OR 2.33, 95%CI 2.04-2.63), ACEI or

ARB at discharge (PM6 OR 2.78, 95%CI 2.44-3.23) and LDL monitoring (PMt1 OR 4.55, 95%CI 4.55-4.76) have a substantial impact on patient mortality. Noncompliance with aspirin at admission (PM1 OR 1.85, 95%CI 1.61-2.17) and aspirin at discharge (PM2 OR 1.47, 95%CI 1.32-1.67) have a significant effect on patient mortality, although not as striking as other measures. Using the general noncompliance comparison group and adjusting for mortality modulating factors, demonstrates that compliance with performance measures tends to decrease patient mortality risk. Using the inverse OR demonstrates the non-trivial influence of clinical practice pattern compliance on that risk.

The choice of comparison group matters. The risk of mortality differs when compliant practice patterns are compared to noncompliant practice patterns for the specific measure of interest (specific noncompliance) or practice patterns generally noncompliant with all performance measures (general noncompliance). This could be due to additive effects or masking of risk. This occurs when compliance with other measures modifies the risk relationship between compliance and noncompliance with the measure of interest because it increases/decreases the mortality in the noncompliant group. For compliance with beta-blocker at discharge (PM3a) the overall effect changes from increasing the risk of mortality to having no effect on patient mortality (OR 1.49 95% CI 1.39-1.56 vs. 0.95, 0.89-1.02). (See Table 5.8) Subsequent mortality in practice patterns that are compliant with aspirin at admission (PM1 OR 0.93, 95%CI 0.80-1.08 vs. 0.54, 0.46-0.62) is not significantly different than the risk of mortality in specific noncompliant practice patterns but decreases the risk when compared to practice patterns that are generally noncompliant. Likewise with receiving an aspirin at discharge (PM2 OR 1.09, 95%CI 0.97-1.22 vs. 0.68, 0.60-0.76), compliance is less risk when compared to generally noncompliant patterns, but it is equivalent to the mortality risk of practice patterns that are specifically noncompliant. With these two performance measures, PM1 and PM2, practice patterns that comply with other measures (specific noncompliance) modifies the effect of patient mortality, specifically it makes noncompliance with PM1 and PM2 less risky and shifts the OR from less than 1 to 1. Compliance with other performance measures masks the positive impact on patient survival that practice which includes the administration of a beta-blocker at admission has on patient mortality (PM3a OR 0.60, 95%CI 0.54-0.58 vs. 0.32, 0.32-0.41). Practice patterns that conform to administration of statins (PM4 OR 0.71,

95%CI 0.62-0.81 vs. 0.43, 0.38-0.49) or ACEI or ARB at discharge (PM6 OR 0.57, 95%CI 0.49-0.66 vs. 0.36, 0.31-0.41) also mask the effect. (See Table 5.8) Monitoring patient LDL levels decreases mortality regardless of the noncompliant comparison group (PMt1 OR 0.23, 95%CI 0.22-0.23 vs. 0.22, 0.21-0.22). But including compliance with other performance measures augments the risk from 2.58 to 4.01. In contrast, compliance with fibrinolytic therapy increases the risk of mortality regardless of noncompliant comparison group (PM7 OR 4.01, 95%CI 2.02-6.65 vs. 2.58, 1.58-4.21) Overall, removing the additive or masking effect of compliance with other measures, leads to the conclusion that compliance does decrease the risk of patient mortality.

Therefore, the answer the inquiry "Does compliance decrease the risk of patient mortality?" is a guarded affirmative. Generally compliance with the performance measure decreases the risk of mortality compared to noncompliance. For fibrinolytic therapy, compliance increases the risk of mortality. For some of the measures, there is evidence of an interaction between compliance with other measures that can modify the risk relationship. To understand how these measures interact with each other and influence patient mortality is an area of further research. The underlying volatility of the risk relationship in the presence of compliance with other measures could indicate that there are sub-groups or situations where compliance with the performance measures is actually detrimental to the patient.

Table 5.7 Risk of Mortality for Noncompliance with the AHA/ACC Performance Measures

Performance Measure	Mortality		
	OR[1]	LCL	ULC
PM1	1.85	1.61	2.17
PM2	1.47	1.32	1.67
PM3a	2.78	2.44	3.13
PM3b	1.05	0.98	1.12
PM4	2.33	2.04	2.63
PM6	2.78	2.44	3.23
PM7	0.39	0.24	0.63
PM8	1.47	0.97	2.17
PMt1	4.55	4.55	4.76
PMt9	--	--	--

Research demonstrates that bleeding events in STEMI patients are associated with other complications, including mortality. Although not specifically stated in the performance measures, many of the STEMI treatment guidelines are aimed at preventing bleeding. The percentage of compliant practice patterns that result in a bleeding event varies from 3.3% for time to fibrinolytic

therapy (PM7) to 14% for clopidogrel at discharge (PMt9). This range is smaller than patient mortality which was 0-21%. Also the narrative of practice patterns and bleeding events is much simpler than mortality. Aspirin at admission (PM1), beta-blocker at admission (PM3a), and LDL monitoring (PMt1) reduce bleeding events and aspirin at discharge (PM2), beta-blocker at discharge (PM3b), ACEI or ARB at discharge (PM6), and clopidogrel at discharge (PMt9) increase bleeding events. (See Table 5.9 unadjusted p-values ≤ 0.05) When we account for factors known to influence bleeding, aspirin at discharge (PM2), beta-blocker at discharge (PM3b), stain at discharge (PM4), ACEI or ARB at discharge (PM6) and LDL monitoring (PMt1) remain associated with bleeding events. (See Table 5.9 Noncompliance with Performance Measure, adjusted p-values ≤0.05) It is interesting that when adjusting for other bleeding risk factors, statins at discharge becomes statistically significant (PM4 unadjusted p-value = 0.77 vs. adjusted p-value = 0.0003). (See Table 5.9 Noncompliant with Performance Measure) When we restrict the comparison group to generally noncompliant practice patterns, there are five performance measures that influence patient bleeding events after accounting for other covariates. Three performance measures that decrease the risk of bleeding are aspirin at admission (PM1 OR 0.73, 95%CI 0.61-0.88), beta-blocker at admission (PM3a OR 0.87, 95%CI 0.78-0.98) and LDL monitoring (PMt1 OR 0.81, 95%CI 0.78-0.83). Compliance with the administration of aspirin at discharge (PM1 OR 1.25, 95%CI 0.10-1.41) or beta-blocker at discharge (PM3b OR 1.22, 95%CI 1.13-1.32) increases the risk of bleeding events. (See Table 5.9) Therefore the answer to the query "Does compliance with the performance measures decrease the risk of bleeding?" is yes, except for aspirin and beta-blocker at discharge.

The performance measures were created to aid healthcare institutions in adherence to clinical practice guidelines. I anticipated that compliance with the performance measures would decrease the risk of non-optimal outcomes of mortality and bleeding events. I found that for mortality, generally compliance decreases the risk of mortality except for fibrinolytic therapy. Furthermore, the data indicates a complex interaction between compliance with the performance measures and patient mortality. For bleeding events, I found an unanticipated answer; compliance with aspirin at discharge or beta-blocker at discharge can increase the risk.

Table 5.8 TAR Compliance with AHA/ACC STEMI Performance Measures and Mortality

Performance Measure	Compliant		Noncompliant with Performance Measure							Noncompliant with All Performance Measures						
			Noncompliant		p-value		Odds Ratio			Noncompliant		p-value		Odds Ratio (OR)		
n=1000000	Mortality	Survival	Mortality	Survival	Unadjusted	Adjusted†	OR†	LCL	UCL	Mortality	Survival	Unadjusted	Adjusted†	OR†	LCL	UCL
PM1	237	4666	57814	937283	0.0035	0.3751	0.93	0.80	1.08	45867	435619	<0.0001	<0.0001	0.54	0.46	0.62
PM2	433	4719	57618	937230	<0.0001	0.1409	1.09	0.97	1.22	45867	435619	0.0063	<0.0001	0.68	0.60	0.76
PM3a	413	10497	57638	931452	<0.001	<0.0001	0.60	0.54	0.68	45867	435619	<0.0001	<0.0001	0.36	0.32	0.41
PM3b	1277	11036	56774	930313	<0.001	<0.0001	1.49	1.39	1.59	45867	435619	0.0016	0.1406	0.95	0.89	1.02
PM4	400	7266	57651	934683	0.0273	<0.0001	0.71	0.62	0.81	45867	435619	<0.0001	<0.0001	0.43	0.38	0.49
PM6	244	4925	57807	937024	0.0008	<0.0001	0.57	0.49	0.66	45867	435619	<0.0001	<0.0001	0.36	0.31	0.41
PM7	25	95	58026	941854	<0.0001	<0.0001	4.01	2.52	6.65	45867	435619	<0.0001	0.0001	2.58	1.58	4.21
PM8	26	352	58025	941597	0.3721	0.5224	1.14	0.76	1.72	45867	435619	0.0796	0.0677	0.68	0.46	1.03
PMt1	9400	475436	48651	466513	<0.0001	<0.0001	0.23	0.22	0.23	45867	435619	<0.0001	<0.0001	0.22	0.21	0.22
PMt9	0	7	58051	941942	0.5113	0.9938	--	--	--	45867	435619	0.3906	0.9949	--	--	--

†Adjusted for gender, comorbid condition present, hospital beds, full catheterization laboratory, and census region
LCL is the lower confidence limit and UCL is the upper confidence limit

Table 5.9 TAR Compliance with AHA/ACC STEMI Performance Measures and Bleed Event

Performance Measure	Compliant		Non-Compliant with Performance Measure							Noncompliant with All Performance Measures						
			Noncompliant		p-value		Odds Ratio			Noncompliant		p-value		Odds Ratio (OR)		
n=1000000	Bleed	No Bleed	Bleed	No Bleed	Unadjusted	Adjusted†	OR†	LCL	UCL	Bleed	No Bleed	Unadjusted	Adjusted†	OR†	LCL	UCL
PM1	166	4737	42343	952754	0.0026	0.5026	0.94	0.78	1.13	22027	459459	<0.0001	0.0007	0.73	0.61	0.88
PM2	310	4842	42199	952649	<0.0001	<0.0001	1.66	1.46	1.88	22027	459459	<0.0001	0.0008	1.25	1.10	1.41
PM3a	417	10493	42092	946998	0.0256	0.8415	0.98	0.88	1.11	22027	459459	0.0002	0.0186	0.87	0.78	0.98
PM3b	835	11478	41674	946013	<0.0001	<0.0001	1.75	1.61	1.89	22027	459459	<0.0001	<0.0001	1.22	1.13	1.32
PM4	331	7335	42178	950156	0.7708	0.0003	1.27	1.11	1.44	22027	459459	0.2850	0.4473	0.85	0.84	1.08
PM6	372	4797	42137	952694	<0.0001	<0.0001	1.66	1.46	1.89	22027	459459	<0.0001	0.2100	1.09	0.95	1.24
PM7	4	116	42505	957375	0.6183	0.8411	0.90	0.33	2.47	22027	459459	0.5152	0.8454	0.90	0.33	2.49
PM8	15	363	42494	957128	0.7853	0.9487	0.98	0.58	1.66	22027	459459	0.5726	0.3816	0.79	0.47	1.34
PMt1	18633	466203	23876	491288	<0.0001	<0.0001	0.78	0.76	0.80	22027	459459	<0.0001	<0.0001	0.81	0.78	0.83
PMt9	1	6	42508	957485	0.1882	0.1842	4.33	0.50	37.7	22027	459459	0.2188	<0.0001	2.87	0.33	25.2

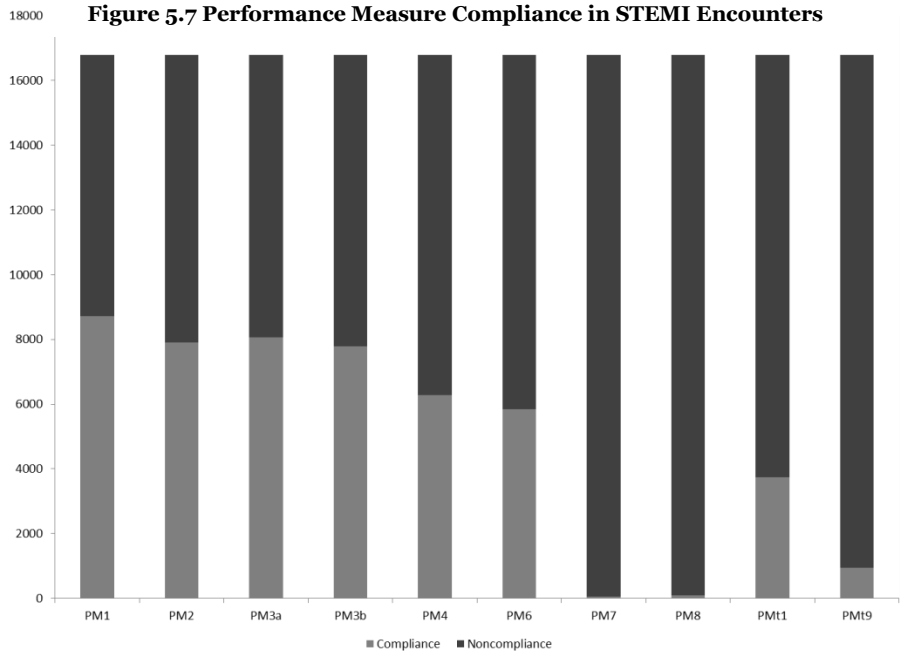
†Adjusted for gender, comorbid condition present, hospital beds, full catheterization laboratory, and census region
LCL is the lower confidence limit and UCL is the upper confidence limit

5.10 Verification of Findings

TARM is an exploratory method that will uncover previously unknown temporal relationships. In the previous section, TARs were filtered by interestingness measures and when in comparison with performance measures were linked to patient outcomes. Based on those findings, performance measures either singularly or in concert affect patient outcomes. Recall that the STEMI patients in the ACSPD were divided into two groups, the first provided data for SPM and TARM and the second is for verifying the TARM findings. This second group is comprised of 16,789 encounters and 4,526,200 events.

Compliance during the encounter varies across the performance measures. In slightly over half the encounters, the patient received an aspirin within 24 hours of admission (PM1 52%). (See Figure 5.7) In less than half of the encounters, the patient received an aspirin at discharge (PM2 47%), a beta-blocker at admission (PM3a 48%) or a beta-blocker at discharge (PM3b 47%). In fewer encounters, the patient received a statin at discharge (PM4 37%) or an ACEI or ARB at discharge (PM6 35%). Patients were tested for LDL levels in less than a quarter of the encounters (PMt1 22%) and clopidogrel was administered at discharge in 6% (PMt9). Fibrinolytic therapy within 30 minutes (PM7 0.4%) or a PCI within 90 minutes of admission (PM8 0.5%) were extremely rare events. The low compliance with the performance measures is due to incomplete data in the ACSPD. Incomplete data makes assessing the quality of care based on the performance measures challenging.

Figure 5.7 Performance Measure Compliance in STEMI Encounters



In order to assess if compliance with the performance measures affects mortality and bleeding events, I used the odds ratio. The odds ratios are the exponential parameter from a generalized estimating equation with logit link. Gender, the presence of at least one comorbid condition, primary reperfusion treatment (PCI, CAGB, or medication), the size of the hospital, presence of a full catheterization laboratory for the treatment of MI patients, and the census region of the hospital are included in the models. Since compliance with fibrinolytic therapy within 30 minutes or PCI within 90 minutes is rare, then the outcomes of mortality and bleeding events in encounters that are compliant are rarer. (See Figure 5.8 and Figure 5.9) Because of this, fibrinolytic therapy and PCI are not included in the models of mortality and bleeding events. In terms of mortality, compliance with the performance measures generally decreases the risk, except for aspirin at admission. Administration of aspirin at admission increases the risk for mortality (OR 1.86, 95%CI 1.55-2.23). In bleeding events, compliance generally decreases risk, except for LDL monitoring (OR 3.97, 3.38-4.67). (See Table 5.10)

The findings from the TARM, demonstrate that clinical practice patterns that are compliant with aspirin at admission decrease the risk of mortality compared to practice patterns that are generally noncompliant with all performance measures and is not significantly different from practice patterns which are noncompliant with this performance measure specifically. When the focus of the analysis is shifted to the encounter, aspirin at admission is associated with an increased risk of mortality. Temporal data mining finds that aspirin at discharge and beta-blocker at discharge increase the risk of bleeding events whereas the encounter-level analysis finds that LDL monitoring increases risk. The finding that LDL monitoring increases risk is most likely a proxy for another behavior or patient characteristic. The sickest patients, for example, are more likely to get LDL testing and more likely to die. But mortality is not due to LDL testing but to the severity of the illness. In summary, using sequential pattern mining and temporal association rule mining, uncovers a different narrative than statistical analysis.

Table 5.10 Encounter Compliance with Performance Measures and Outcomes

Performance Measures	Mortality			Bleed		
	OR	LCL	UCL	OR	LCL	UCL
PM1	1.86	1.55	2.23	0.72	0.59	0.88
PM2	0.66	0.52	0.83	0.91	0.73	1.14
PM3a	0.73	0.62	0.88	0.91	0.76	1.12
PM3b	0.49	0.39	0.62	1.22	0.98	1.53
PM4	0.33	0.25	0.43	0.54	0.44	0.66
PM6	0.46	0.36	0.59	0.57	0.47	0.70
PMt1	0.56	0.47	0.68	3.97	3.38	4.67
PMt9	1.22	0.88	1.71	1.29	0.95	1.75

Figure 5.8 Mortality and Performance Measure Compliance in STEMI Encounters

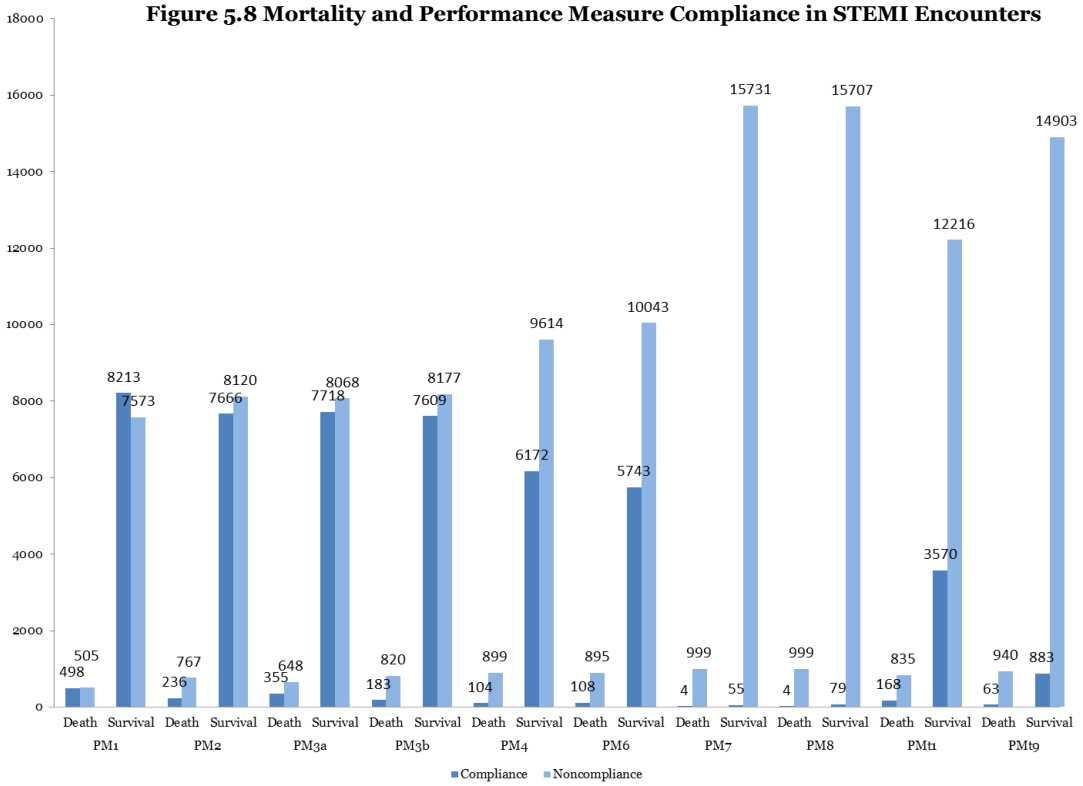
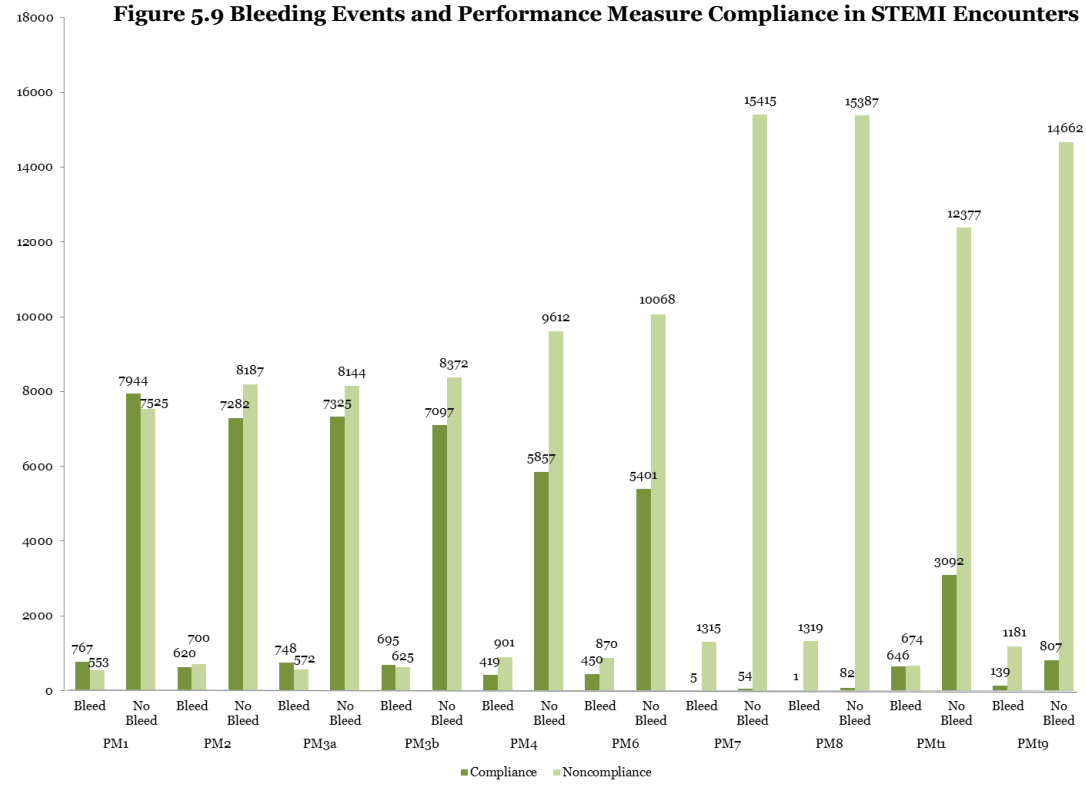


Figure 5.9 Bleeding Events and Performance Measure Compliance in STEMI Encounters



5.11 Conclusion

STEMI is the cause of significant morbidity and mortality in the US each year. I used the ACSPD, a large longitudinal EMR database to investigate the temporal relationship between events in STEMI treatment. In clinical medicine, orders represent the record of how a patient was treated. An order set is a group of orders that are commonly issued together. Finding order sets in EMR data is only one aspect of understanding STEMI treatment. The other is the temporal relationship between events, the clinical practice patterns. To find order sets and clinical practice patterns I used SPM and TARM.

Using the SPM in the ACSPD, I discovered 39 order sets. Not all order sets are present for the 9 year span and overall order set use drops precipitously in 2004. I postulate that this denotes a shift in medical practice. The cause is unknown, but in late 2004, the AHA published new STEMI treatment guidelines. I condensed the ACSPD sequences using the order sets then applied the TARM algorithm. Using support, confidence, Bayes' factor, lift, likelihood, and Zhang's measures, I found substantial variation, rarity and weak antecedent-consequent pairing in the clinical practice patterns. To explore the interaction between clinical decisions and patient outcomes, I compared the clinical practice patterns with AHA STEMI performance measures for compliance and analyzed the risk of bleeding and mortality. I found that for mortality, generally compliance with the AHA performance measures decreases the risk of mortality except for fibrinolytic therapy. Furthermore, the data indicates a complex interaction between compliance with the performance measures and patient mortality. For bleeding events, I found an unanticipated answer; compliance with aspirin at discharge or beta-blocker at discharge can increase the risk. It would require additional analyzes to uncover possible causes of this result.

The STEMI encounters were divided into two groups, one third for SPM and TARM, and two-thirds to validate the findings. The validation analysis focused on performance measure compliance for the encounter rather than the clinical practice pattern. Compliance across the measures is low due to incomplete data more than quality of care. I found that aspirin at admission increased the risk of mortality and LDL monitoring increased the risk of bleeding events. The encounter level validation did not have the same findings as the clinical practice patterns. At the encounter level, fibrinolytic therapy within 30 minutes is a rare event; too rare to analyze its effect on morality or bleeding. The clinical practice patterns show that fibrinolytic therapy does influence patient mortality. LDL monitoring during

an encounter decreases the risk of bleeding when we examine clinical practice patterns and outcomes. At the encounter level, LDL monitoring is most likely a proxy for another behavior or patient characteristic that increases the risk of bleeding. Using both the temporal data mining approach and the statistical outcomes analysis provides a richer insight into STEMI treatment.

Chapter 6 Future Work

STEMI is the cause of significant morbidity and mortality in the US each year. The AHA/ACC clinical practice guidelines and evidence-based best practices indicate that timing of treatment is important for optimal patient outcomes. There is substantial research in the treatment of STEMI that focuses on a single event and the risk/benefits thereof. However there is a dearth of research into the relationship between events. The interplay between events during encounter is especially important in ST-elevated myocardial infarction (STEMI), where the timing of treatments is crucial for maximizing positive patient outcomes.

In this chapter, I summarize my findings, discuss the strengths, limitations, assumptions and implications of my work for clinical medicine, and finally highlight directions for future work.

6.1 Dissertation Summary

I employed a two-step data mining process using (1) sequential pattern mining and (2) temporal association rule data mining to generate temporal association rules (TAR). The TARs were evaluated for potential interest to a user and risk of outcomes in the patient. To accomplish this, I created a sequential pattern mining (SPM) algorithm and a temporal association rule data mining (TARM) algorithm. Sequential pattern mining is an exploratory mining process to discover temporally ordered patterns of events. The SPM algorithm incorporates user-defined thresholds in order to maximize efficiency and effectiveness. The SPM also includes novel refinements to discover the longest, non-redundant patterns. The user-defined thresholds and the novel refinements combine to create an algorithm that is able to find patterns in noisy data. The TARM algorithm is designed to utilize large data repositories to discover rules. Unlike previous temporal association rule algorithms, this one does not generate candidate rules by combining events and searching for the candidates in the data, and then pruning those candidates with a frequency below a predefined threshold. Instead, rule creation is governed by the sequence of events in the data. This allows for rules with gaps and intervals. TARM, as with any temporal association rule algorithm, can discover a large number of rules. Some of the rules are more interesting to the user than others. To differentiate these rules, a set of interestingness measures are used. I used the support, confidence, Bayes' factor, lift, likelihood and Zhang's measure. Another method of analyzing rules is similarity to current knowledge, which in this context is compliance with clinical practice guideline

recommendations. The interestingness measures, compliance with guidelines and the SPM and TARM algorithms were used in the ACSPD as a unique approach to exploring STEMI treatment and patient outcomes.

When I applied SPM to the ACSPD, it uncovered 39 order sets. Not all 39 order sets were present for the 9 year span of the EMR data, with new order sets emerging while others fade. Overall the use of order sets drops precipitously in 2004. I postulate that this denotes a shift in the practice of medicine related to STEMI. The direct cause is unknown, but in late 2004, the AHA/ACC published a new set of STEMI treatment guidelines, the first in nearly 5 years.

I used the 39 order sets to condense the encounter sequences in the ACSPD. I then used the condensed sequences in TARM to discover temporal association rules. In this context, the TARs are clinical practice patterns. I found that substantial variation in clinical practice patterns exist and many practice patterns are unique and rare. Also, I found that generally, the antecedent-consequent pairing is weak. Furthermore, the low frequency was challenging for the interestingness measures to differentiate between rules of interest by ranking.

The American Heart Association (AHA) in conjunction with the American College of Cardiology (ACC) created a set of performance measures for the treatment of ACS. The performance measures are quantifiable metrics of the quality of care a patient receives and are derived from the ACS clinical practice guidelines. I used the TARs and the AHA/ACC performance measures to analyze the risk of bleeding and mortality. The mortality and bleed risk varies by performance measure. Clinical practice pattern compliance with some measures increases the risk, some have no statistically significant effect and others decrease the risk. The direction of the effect is strongly influenced by the comparison group. If the risk of compliant TARs is compared to a group that includes rules compliant with other measures, the risk profile can change direction, i.e. become more or less risky, when compared to rules that are not compliant with any of the measures. The presence of multiple performance measures in the comparison group could be augmenting or masking the effect of compliance with the measure of interest.

The STEMI encounters were divided into two groups, one third for SPM and TARM, and two-thirds to validate the findings. The validation analysis focused on performance measure compliance for the

encounter rather than the clinical practice pattern. Compliance across the measures is low due to incomplete data rather than the quality of care. I found the encounter level validation did not have the same findings as the clinical practice patterns. Using both the temporal data mining approach and the statistical outcomes analysis provides a richer insight into STEMI treatment.

6.2 Strengths, Limitations, and Assumptions

The strengths and limitations of this research stem from the same sources, the data, the methods, and the resources. EMR data can be a rich source for data mining and knowledge discovery. Large electronic medical record repositories are an opportunity to investigate the complex interaction between treatment decisions and patient outcomes. In this dissertation, the strengths of the ACSPD are that it is representative of the ACS population and its size, which allows for investigating the treatment of STEMI. EMR data are also challenging because it is often incomplete and the data are uncertain; the ACSPD is no different. The ACSPD data include medications, laboratory tests, and procedures. But it is not the case that every encounter has medication, laboratory test and procedure information. When data are missing, such as medications, it does not mean that the medication was not administered. Rather it means that the information was not recorded. The uncertainty arises from measurement error, e.g. laboratory timestamp values. Consider, the temporal record of an event, despite the precision of the date and time stamped value, there still mains some uncertainty about when the event occurred versus when it was recorded. Incomplete and uncertain data means that the estimates derived from the data are the lower bound. For example, when measuring compliance with aspirin in the first 24 hours, the missing information about medications or the uncertainty in timestamp could mean that there are more encounters that comply than the data indicates. Another limitation of using the ACSPD is that the sources of the EMR data are healthcare institutions that use Cerner products. Using Cerner products may self-select healthcare providers that treat STEMI differently or have better/worse outcomes than the general population. This could weaken the generalizability of my findings. The data, while providing a unique opportunity to investigate the temporal relationships between events in the treatment of STEMI, are also a primary source of the limitations in this research.

The methods, like the data, are a source of strengths and limitations. Data mining, metrics, and statistical analyses are powerful methods to uncover insights in data. The SPM and TARM discover novel insights

into STEMI treatment that we could not see if using statistical methods only. Each method has assumptions and limitations which influence the interpretation of the results. Therefore care must be taken when using multiple methods, because the assumptions of one method could contradict the assumptions of another. This is especially true when, as in this work, the results of one algorithm are used as the input for another. In this work, the SPM and TARM both assume timed-based ordering of data. Therefore the basic assumption of one algorithm (SPM) is in concert with the basic assumption of the other (TARM). The design and implementation of methods can also provide limitations. The SPM uses 3 user-defined thresholds, but pattern growth is initiated by frequency. The assumption is that frequency is an important attribute. But if frequency is not an important attribute to the user, the effectiveness of the SPM diminished greatly. The TARM relies on 3 event rules. Shorter rules with only 2 events would, theoretically increase rule support and improve the identification of important rules by the interestingness measures. Another method-derived source of strengths and limitation is simulation. In this dissertation, I also used simulation to evaluate algorithm performance. Simulation is a method that allows researchers to better understand algorithm performance by controlling the data. It is particularly useful when labeled data are not available. By varying the simulation parameters I was able to demonstrate that SPM performance is governed more by the user's threshold selections than the attributes of the data. Simulation is able to mimic aspects of real world EMR data albeit a simplified version. The simulation I used is unable to capture complex patterns that exist in EMR data. It is unable, for example, to inject specific medications or treatments associated with a specific comorbid condition. The methods, while providing valuable insights into STEMI treatment, are also a primary source of limitations.

The strengths and limitations of the data and methods are influenced by the strengths and limitations of the resources available. The cost of computing is decreasing at the same time that the size of the data and the complexity of the methods are increasing. For this work, I did not have access to cluster computing resources. The algorithms were designed to take advantage of parallel processing and all available memory. However if I had access to more robust resources, the temporal investigation could have been more detailed and nuanced.

Currently there are no methods for directly measuring clinician intent and motivations for treatment choices on a large scale. Electronic medical records capture the effects of those decisions in the form of orders for medications, procedures, and laboratory tests. The results of the orders, the condition of the patient and environmental factors form a feedback loop, influencing future decisions by the clinician. In this dissertation I make strong assumptions about the relationship between healthcare decisions and the date and time stamped events in the database. I implicitly hold that each action is the observable and measurable result of a decision. This means that the events are not random and independent, but are conditioned on an unknown decision and are dependent on other events. I also acknowledge that electronic medical record data are uncertain and incomplete. I accept these assumptions and their limitations as necessary to the computational evaluation of healthcare practice in the context of treatment for STEMI and where feasible I have adjusted my analyses to take into account these limitations.

6.3 Implications for Clinical Practice

There is substantial research on the use of aspirin, beta-blockers, ARB's etc. for treating STEMI.(Armstrong, Collen et al. 2003, Eagle, Lim et al. 2004, Spencer, Allegrone et al. 2004, Fox, Eagle et al. 2010, Yeh, Sidney et al. 2010, Jernberg, Johanson et al. 2011) The evidence suggests that using these medications decreases the risk of detrimental patient outcomes. Therefore it is not surprising that the AHA/ACC used this evidence as the basis for the performance measures. Research into compliance with evidence-based best practices from the Global Registry of Acute Coronary Events (GRACE) demonstrates that adherence decreases mortality. (Fox, Steg et al. 2007) The GRACE analysis spans from 1999 to 2005 and examined each medication separately. In this longitudinal analysis, the GRACE researchers did not use a noncompliant with any performance measure cohort for comparison. Granted, due to the high compliance with aspirin use, >95% of encounters, building such a cohort would be difficult.

The Worchester Heart Attack Study was a 30 year community surveillance study based in Worchester, MA.(Floyd, Yarzebski et al. 2009) In the period from 1997-2005, researchers found a decrease in the morbidity and mortality associated with STEMI and an increase in the use of aspirin, beta-blockers, lipid-lowering therapy, ACE/ARB's, thrombolytics, cardiac catheterization and PCI. They did find that the rates of CABG declined during the study period.(McManus, Gore et al. 2011) In the Myocardial Ischeamia

National Audit Project (MINAP), a multicenter prospective registry study from 2008-2009 in England and Wales, hospital performance of evidence-based therapies (aspirin, beta-blocker, ACE/ARB and statin) was associated with lower 30 day and 6 month patient mortality.(Simms, Baxter et al. 2013) In Germany, analysis from the Secondary Prevention after Acute Myocardial Infarction registry found that for STEMI patients receiving optimal medical therapy at discharge reduced 1-year mortality.(Bramlage, Messer et al. 2010) A study of acute myocardial infarction patients treated at the University of Michigan Medical Center from 1999-2002 found that patients who receive more evidence-based medications (antiplatelet agents, beta-blockers, ACEI, and lipid lowering therapy) have a lower risk of mortality vs. patients receiving none of the evidence-based medications.(Mukherjee, Fang et al. 2004)

The arc of research bends toward following the guideline recommendations for the treatment of STEMI. Compliance is associated with increased survivorship in-hospital and after. But when we consider the work as a whole, researchers use composite scores of compliance. Composite scores quantify compliance, higher scores mean better care. But composite scores cannot transmit information about which guidelines are followed and the effect of individual guidelines on the outcomes. This research shows that compliance with other performance measures can have a masking/augmenting effect. Furthermore, previous research is conducted at the encounter level only. This research is one of the first to examine compliance with each guideline from the perspective of association rules and the patient encounter. I found that using both approaches provides richer insights into STEMI treatment.

Comparison across studies for bleeding events is more difficult. Mortality is a relatively easy outcome to measure compared to bleeding events. The definition of bleeding events varies between studies depending on the type of information available. Studies that include physicians' notes may require a specific clinical notation. Studies with transfusion information may require a time stamped transfusion. Studies with laboratory results may require a change in laboratory readings. Although a standardized definition would ease comparisons, it may not be appropriate for all disease states. (Steinhubl, Kastrati et al. 2007)

6.4 Future Work

Exploratory analysis to discover meaningful patterns in data is challenging. Future work will push the current boundaries of temporal data mining, increasing effectiveness and efficiency. As the size of available data, and the complexity of systems (clinical treatment) increases, it is imperative to create algorithms to handle these challenges. In addition, when using more than one algorithm or data from multiple sources, researchers must pay special attention to the limitations and assumptions of algorithms and select those that are complimentary to address the new and interesting research questions of the future. An integral part of understanding the strengths, limitations and assumptions of novel methods in future work will be simulation. Simulations will not only mimic overall attributes of data but also model the complex relationships within real world data.

In the current work, I have found that compliance with the performance measures does affect patient outcomes. But this work does not address *why* a clinician would not comply with the clinical guidelines. There are many reasons for a clinician to decide not to adhere to the guidelines. Compliance with clinical practice guidelines improves patient outcomes, as this dissertation demonstrates, but is difficult to implement and maintain compliant practice behavior in a healthcare setting. One of the barriers to guideline compliance is the guidelines themselves. Guidelines are often vague and are frequently updated to reflect changes in the understanding of disease and treatments. This makes it challenging to include compliance in healthcare setting practice patterns. Another barrier is the complexity of patient health issues. Patients in a healthcare setting may have multiple concurrent health issues and the best practices for one issue may conflict with those of another. Therefore, the healthcare practitioners, for various reasons, may make care decisions that are not in compliance with the guidelines for a specific disease.

This research work does not address the appropriateness of care. Appropriateness and compliance are closely related issues in clinical care. A course of treatment may be appropriate for the patient's specific circumstances but noncompliant with the guidelines. The inverse could also occur, where a course of treatment is not appropriate for the patient, but in compliance with the guidelines. Ideally, the treatment is appropriate and in compliance. It is outside the scope of this research to determine the appropriateness of care. Future endeavors would be to understand how appropriateness and compliance affect patient outcomes in the treatment of ACS.

6.5 Contributions

The primary contribution of this work is in exploring clinical practice patterns and their effect on patient outcomes. The focus of this dissertation is an exploratory analysis of the temporal relationships between events during STEMI treatment. Clinical practice patterns are the representation of the temporal relationships, i.e. the temporal association rules that describe STEMI treatment. This work is the first to mine clinical practice patterns and assess those practice patterns for compliance with treatment guidelines. Furthermore, to the best of my knowledge, this is the first work to analyze clinical practice pattern compliance and patient outcomes.

Novel methods for temporal data mining in big data are the second area of contribution. Gaining insights into clinical practice using EMR big data is challenging because of the large amount of data and often the data are incomplete. Existing methods are not designed to work with large incomplete data. Furthermore, existing methods are not designed for exploratory analysis of temporal relationships. Therefore I created two algorithms for temporal data mining, the sequential pattern mining algorithm (SPM) and the temporal association rule mining algorithm (TARM). To the best of my knowledge, no research exists on the combined used of sequential pattern mining and temporal association rule data mining in EMR data.

The third area of contribution is in using simulation. Simulation of data is a powerful technique that can inform researchers about the properties of an algorithm or system. In this dissertation, I use simulated data to evaluate the performance of the SPM algorithm. The purpose of the SPM is to find sequential patterns in data. However large amounts of labeled data for algorithm performance evaluation are not available. Therefore simulated data are generated to mimic real world encounter data. In EMR data, events do not have the same probability of occurrence, some events happen more often than others. Likewise sequential patterns, i.e. order sets, do not have the same probability. By varying a set of data simulation parameters, I was able to evaluate the performance of the SPM algorithm in different conditions. It is my hope that future work will unlock more insights into STEMI treatment and reduce the morbidity and mortality it causes.

Appendix

Table A.1 STEMI Encounter Outcomes by Demographics, Hospital Characteristics, and Treatment

Characteristic	Death (n = 2205)	Survival (n = 29516)	p-value	Bleed (n =2269)	No bleeding (n = 29452)	p-value
Demographics						
Age (yr)						
18-25	3	51	<0.0001	1	53	<0.0001
26-30	3	89		2	90	
31-35	4	272		11	265	
36-40	19	766		31	754	
41-45	25	1706		83	1648	
46-50	60	2743		130	2673	
51-55	92	3545		188	3449	
56-60	146	3847		239	3754	
61-65	172	3436		250	3358	
66-70	223	3117		276	3081	
71-75	288	3117		333	3072	
75+	1170	6810		725	7255	
mean ± SD	73.96±12.50	63.57±13.86	<0.0001			
Gender						
Female	1135	10208	<0.0001	925	10418	<0.0001
Male	1069	19306		1344	19031	
Race/Ethnicity						
African-Am	172	1979	0.069	176	1975	0.080
Asian	12	235		26	221	
Biracial	1	7		0	8	
Caucasian	1792	24157		1894	24055	
Hispanic	23	350		31	342	
Native Am	3	103		5	101	
Pacific Islander	0	0		0	0	
Other	23	430		29	424	
Unknown	65	733		74	724	
Comorbid conditions	1797	15711	<0.0001	1474	1077	<0.001
Hospital						
Acute care facility	1760	21165	0.209	1612	21215	0.751
Bed size, range						
< 5	4	19	<0.0001	0	22	<0.0001
6-99	93	1262		68	1281	
100-199	100	1545		70	1572	
200-299	476	4422		302	4587	
300-499	510	5561		578	5485	
500+	585	8355		598	8330	
Catheterization lab						
Full	1641	19748	0.001	1544	19782	<0.0001
Diagnostic	1268	14745		1189	14796	0.001
Census region						
Midwest	539	6172	0.008	643	6061	<0.0001
Northeast	671	8897		491	9056	
South	500	5453		431	5512	
West	58	642		51	648	
Teaching	1178	14439	0.167	1182	14414	<0.0001
Non-teaching	590	6725		434	6863	
Urban	1765	21060	0.059	1614	21175	0.040
Rural	3	104		2	102	
Procedure						
PCI	511	13058	<0.0001	765	12792	<0.0001
CABG	118	2401	<0.0001	750	1766	<0.0001
Order of therapies						
Fibrinolytic only	66	888	<0.0001	62	892	<0.0001
PCI only	476	23360		608	12108	
CABG only	93	1964		610	1447	
Fibrinolytic before PCI (no CABG)	10	436		50	396	
PCI before Fibrinolytic (no CABG)	1	14		3	12	
PCI before CABG	22	352		98	276	
CABG before PCI (no fibrinolytic)	1	8		1	8	
Fibrinolytic, then PCI then CAGB	0	6		4	2	
PCI then Fibrinolytic, then CABG	0	1		1	0	
PCI, CABG, Fibrinolytic	1	0		0	1	

References

Agrawal, R., T. Imieliński and A. Swami (1993). "Mining association rules between sets of items in large databases." ACM SIGMOD Record **22**(2): 207-216.

Agrawal, R. and R. Srikant (1994). Fast algorithms for mining association rules in large databases, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*. Morgan Kaufmann Publishers Inc.

Ale, J. and G. Rossi (2000). An approach to discovering temporal association rules, *Proceedings of the 2000 ACM Symposium on Applied Computing*. ACM (pp.294-300).

Anderson, J. L., C. D. Adams, E. M. Antman, C. R. Bridges, R. M. Califf, D. E. Casey Jr, W. E. Chavey, F. M. Fesmire, J. S. Hochman and T. N. Levin (2007). "ACC/AHA 2007 guidelines for the management of patients with unstable angina/non-ST-elevation myocardial infarction: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Revise the 2002 Guidelines for the Management of Patients With Unstable Angina/Non-ST-Elevation Myocardial Infarction) developed in collaboration with the American College of Emergency Physicians, the Society for Cardiovascular Angiography and Interventions, and the Society of Thoracic Surgeons endorsed by the American Association of Cardiovascular and Pulmonary Rehabilitation and the Society for Academic Emergency Medicine." Journal of the American College of Cardiology **50**(7): e1-e1.

Antman, E. M., D. T. Anbe, P. W. Armstrong, E. R. Bates, L. A. Green, M. Hand, J. S. Hochman, H. M. Krumholz, F. G. Kushner and G. A. Lamas (2004). "ACC/AHA guidelines for the management of patients with ST-elevation myocardial infarction." Circulation **110**(9): e82-e82.

Antman, E. M., M. Hand, P. W. Armstrong, E. R. Bates, L. A. Green, L. K. Halasyamani, J. S. Hochman, H. M. Krumholz, G. A. Lamas, C. J. Mullany, D. L. Pearle, M. A. Sloan and S. C. Smith Jr (2008). "2007 Focused Update of the ACC/AHA 2004 Guidelines for the Management of Patients With ST-Elevation Myocardial Infarction." Journal of the American College of Cardiology **51**(2): 210-247.

Armstrong, P. W., D. Collen and E. Antman (2003). "Fibrinolysis for acute myocardial infarction the future is here and now." Circulation **107**(20): 2533-2537.

Bottrighi, A., F. Chesani, P. Mello, M. Montali, S. Montani, S. Storari and P. Terenziani (2010). "Analysis of the GLARE and GPROVE Approaches to Clinical Guidelines." Knowledge Representation for Health-Care. Data, Processes and Guidelines: 76-87.

Bramlage, P., C. Messer, N. Bitterlich, C. Pohlmann, A. Cuneo, E. Stammwitz, J. Tebbenjohanns, H. Gohlke, J. Senges and U. Tebbe (2010). "The effect of optimal medical therapy on 1-year mortality after acute myocardial infarction." Heart **96**(8): 604-609.

Braunwald, E., E. Antman, J. Beasley, R. Califf, M. Cheitlin, J. Hochman, R. Jones, D. Kereiakes, J. Kupersmith and T. Levin (2002). "ACC/AHA 2002 guideline update for the management of patients with unstable angina and non–ST-segment elevation myocardial infarction—summary article A report of the American College of Cardiology/American Heart Association task force on practice guidelines (Committee on the Management of Patients With Unstable Angina)." Journal of the American College of Cardiology **40**(7): 1366-1374.

Brin, S., R. Motwani and C. Silverstein (1997). Beyond market baskets: generalizing association rules to correlations. ACM SIGMOD Record, ACM.

Ceglar, A. and J. F. Roddick (2006). "Association mining." ACM Computing Surveys (CSUR) **38**(2): 5.

Census. (2011, January 06, 2011). "The 2011 Statistical Abstract: Estimates and Projections by Age, Sex, Race/Ethnicity." Retrieved January 20, 2011, from

http://www.census.gov/compendia/statab/cats/population/estimates_and_projections_by_age_sex_raceethnicity.html.

Centers for Disease Control and Prevention. "Behavioral Risk Factor Surveillance System: Prevalence and Trends Data, Chronic Health Indicators - 2012." Retrieved Nov 1, 2013, from http://apps.nccd.cdc.gov/brfss/list.asp?cat=CH&yr=2012&qkey=8091&state=All.

Centers for Disease Control and Prevention. a. "Vital Statistics; Underlying Cause of Death, 1999-2010 Results." Retrieved Nov 1, 2013, from http://wonder.cdc.gov/controller/datarequest/D76.

Cheitlin, M., W. Armstrong, G. Aurigemma, G. Beller, F. Bierman, J. Davis, P. Douglas, D. Faxon, L. Gillam and T. Kimball (2003). "ACC/AHA/ASE 2003 Guideline update for the clinical application of echocardiography." J Am Coll Cardiol **42**: 954-970.

Chen, X. and I. Petrounias (1998). A framework for temporal data mining. Database and Expert Systems Applications, Springer.

Chen, Y., G. Dong, J. Han, B. W. Wah and J. Wang (2002). Multi-dimensional regression analysis of time-series data streams, *Proceedings of the 28th international conference on Very Large Data Bases*.VLDB Endowment.

Chesani, F., E. Lamma, P. Mello, M. Montali, S. Storari, P. Baldazzi and M. Manfredi (2008). "Compliance checking of cancer-screening Careflows: an approach based on Computational Logic." Studies in health technology and informatics **139**: 183.

Chu, C., V. Tseng and T. Liang (2008). "An efficient algorithm for mining temporal high utility itemsets from data streams." Journal of Systems and Software **81**(7): 1105-1117.

Concaro, S., L. Sacchi, C. Cerra, M. Stefanelli, P. Fratino and R. Bellazzi (2009). Temporal Data Mining for the Assessment of the Costs Related to Diabetes Mellitus Pharmacological Treatment, *AMIA Annual Symposium Proceedings*. American Medical Informatics Association.

Dahlbäck, B. (2000). "Blood coagulation." The Lancet **355**(9215): 1627-1632.

Davies, M. J. (2000). "The pathophysiology of acute coronary syndromes." Heart **83**(3): 361-366.

Denny, J. C., J. F. Peterson, N. N. Choma, H. Xu, R. A. Miller, L. Bastarache and N. B. Peterson (2010). "Extracting timing and status descriptors for colonoscopy testing from electronic medical records." Journal of the American Medical Informatics Association **17**(4): 383-388.

Eagle, K. (2002). "Lessons from GRACE: The global registry of acute coronary events." European Heart Journal Supplements **4**(suppl E): E24-E31.

Eagle, K. A., R. A. Guyton, R. Davidoff, F. H. Edwards, G. A. Ewy, T. J. Gardner, J. C. Hart, H. C. Herrmann, L. D. Hillis and A. M. Hutter Jr (2004). "ACC/AHA 2004 guideline update for coronary artery bypass graft surgery: summary article: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to Update the 1999 Guidelines for Coronary Artery Bypass Graft Surgery)." Circulation **110**(9): 1168-1168.

Eagle, K. A., M. J. Lim, O. H. Dabbous, K. S. Pieper, R. J. Goldberg, F. Van de Werf, S. G. Goodman, C. B. Granger, P. G. Steg, J. M. Gore, A. Budaj, A. Avezum, M. D. Flather, K. A. Fox and G. Investigators (2004). "A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry." JAMA **291**(22): 2727-2733.

Epstein, F. H., R. D. Rosenberg and W. C. Aird (1999). "Vascular-bed–specific hemostasis and hypercoagulable states." New England Journal of Medicine **340**(20): 1555-1564.

Ewens, W. W. J. and G. R. Grant (2001). Statistical methods in bioinformatics: an introduction, Springer.

Floyd, K. C., J. Yarzebski, F. A. Spencer, D. Lessard, J. E. Dalen, J. S. Alpert, J. M. Gore and R. J. Goldberg (2009). "A 30-Year Perspective (1975–2005) Into the Changing Landscape of Patients Hospitalized With Initial Acute Myocardial Infarction: Worcester Heart Attack Study." Circulation: Cardiovascular Quality and Outcomes **2**(2): 88-95.

Fox, J., N. Johns, A. Rahmanzadeh and R. Thomson (1998). "Disseminating medical knowledge: the PROforma approach." Artificial Intelligence in Medicine **14**(1-2): 157-182.

Fox, K., K. Eagle, J. Gore, P. Steg and F. Anderson (2010). "The Global Registry of Acute Coronary Events, 1999 to 2009–GRACE." Heart **96**(14): 1095.

Fox, K. A. A., P. G. Steg, K. A. Eagle, S. G. Goodman, F. A. Anderson Jr, C. B. Granger, M. D. Flather, A. Budaj, A. Quill and J. M. Gore (2007). "Decline in rates of death and heart failure in acute coronary syndromes, 1999-2006." Jama **297**(17): 1892-1892.

Garofalakis, M. N., R. Rastogi and K. Shim (1999). SPIRIT: Sequential pattern mining with regular expression constraints. VLDB **99**: 7-10.

Geng, L. and H. J. Hamilton (2006). "Interestingness measures for data mining: A survey." ACM Computing Surveys (CSUR) **38**(3): 9.

Giannella, C., J. Han, J. Pei, X. Yan and P. S. Yu (2003). "Mining frequent patterns in data streams at multiple time granularities." Next generation data mining **212**: 191-212.

Hamm, C. W., M. Bertrand and E. Braunwald (2001). "Acute coronary syndrome without ST elevation: implementation of new guidelines." The Lancet **358**(9292): 1533-1538.

Han, J., J. Pei and Y. Yin (2000). Mining frequent patterns without candidate generation. ACM SIGMOD Record, ACM **29**(2): 1-12.

Hanauer, D. A. and N. Ramakrishnan (2013). "Modeling temporal relationships in large scale clinical associations." Journal of the American Medical Informatics Association **20**(2): 332-341.

Hess, E. P., G. A. Wells, A. Jaffe and I. G. Stiell (2008). "A study to derive a clinical decision rule for triage of emergency department patients with chest pain: design and methodology." BMC Emerg Med **8**: 3.

Hoekstra, J. W., C. V. Pollack, M. T. Roe, E. D. Peterson, R. Brindis, R. A. Harrington, R. H. Christenson, S. C. Smith, E. M. Ohman and W. B. Gibler (2002). "Improving the Care of Patients with Non-ST-elevation Acute Coronary Syndromes in the Emergency Department: The CRUSADE Initiative." Academic emergency medicine **9**(11): 1146-1155.

Jaroszewicz, S., T. Scheffer and D. A. Simovici (2009). "Scalable pattern mining with Bayesian networks as background knowledge." Data Mining and Knowledge Discovery **18**(1): 56-100.

Jensen, P. B., L. J. Jensen and S. Brunak (2012). "Mining electronic health records: towards better research applications and clinical care." Nature Reviews Genetics **13**(6): 395-405.

Jernberg, T., P. Johanson, C. Held, B. Svennblad, J. Lindbäck, L. Wallentin and SWEDEHEART/RIKS-HIA (2011). "Association between adoption of evidence-based treatment and survival for patients with ST-elevation myocardial infarction." JAMA **305**(16): 1677-1684.

Johnson, P., S. Tu, N. Booth, B. Sugden and I. Purves (2000). Using scenarios in chronic disease management guidelines for primary care, *Proceedings of the AMIA Symposium*. American Medical Informatics Association.

King III, S., S. Smith Jr, J. Hirshfeld Jr, A. Jacobs, D. Morrison and D. Williams (2008). "2007 Focused Update of the ACC/AHA/SCAI 2005 Guideline Update for Percutaneous Coronary Intervention: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines: 2007 Writing Group to Review New Evidence and Update the ACC/AHA/SCAI 2005 Guideline Update for Percutaneous Coronary Intervention, Writing on Behalf of the 2005 Writing Committee." Circulation **117**(2): 261.

Krumholz, H. M., J. L. Anderson, B. L. Bachelder, F. M. Fesmire, S. D. Fihn, J. M. Foody, P. M. Ho, M. N. Kosiborod, F. A. Masoudi and B. K. Nallamothu (2008). "ACC/AHA 2008 Performance Measures for Adults With ST-Elevation and Non–ST-Elevation Myocardial InfarctionA Report of the American College of Cardiology/American Heart Association Task Force on Performance Measures (Writing Committee to Develop Performance Measures for ST-Elevation and Non–ST-Elevation Myocardial Infarction) Developed in Collaboration With the American Academy of Family Physicians and American College of Emergency Physicians Endorsed by the American Association of Cardiovascular and Pulmonary Rehabilitation, Society for Cardiovascular Angiography and Interventions, and Society of Hospital Medicine." Journal of the American College of Cardiology **52**(24): 2046-2099.

Krumholz, H. M., J. L. Anderson, N. H. Brooks, F. M. Fesmire, C. T. Lambrew, M. B. Landrum, W. D. Weaver, J. Whyte, R. O. Bonow and S. J. Bennett (2006). "ACC/AHA clinical performance measures for adults with ST-elevation and non–ST-elevation myocardial infarctionA report of the American College of Cardiology/American Heart Association Task Force on Performance Measures (Writing Committee to Develop Performance Measures on ST-Elevation and Non–ST-Elevation Myocardial Infarction)." Journal of the American College of Cardiology **47**(1): 236-265.

Krumholz, H. M., A. R. Merrill, E. M. Schone, G. C. Schreiner, J. Chen, E. H. Bradley, Y. Wang, Y. Wang, Z. Lin and B. M. Straube (2009). "Patterns of hospital performance in acute myocardial infarction and heart failure 30-day mortality and readmission." Circulation: Cardiovascular Quality and Outcomes **2**(5): 407-413.

Kushner, F., M. Hand, S. Smith Jr, S. King III, J. Anderson, E. Antman, S. Bailey, E. Bates, J. Blankenship and D. Casey Jr (2009). "2009 Focused Updates: ACC/AHA Guidelines for the Management of Patients With ST-Elevation Myocardial Infarction (Updating the 2004 Guideline and 2007 Focused Update) and ACC/AHA/SCAI Guidelines on Percutaneous Coronary Intervention (Updating the 2005 Guideline and 2007 Focused Update): A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines." Journal of the American College of Cardiology.

Laxman, S. and P. Sastry (2006). "A survey of temporal data mining." Sadhana **31**(2): 173-198.

Lee, C. H., M. S. Chen and C. R. Lin (2003). "Progressive partition miner: An efficient algorithm for mining general temporal association rules." IEEE Transactions on Knowledge and Data Engineering: 1004-1017.

Lee, C. H., C. R. Lin and M. S. Chen (2001). Sliding-window filtering: an efficient algorithm for incremental mining, ACM.

Levenshtein, V. (1965). Binary codes with correction of deletions, insertions and substitution of symbols. Dokl. Akad. Nank. SSSR.

Li, J., H. Shen and R. Topor (2002). "Mining the optimal class association rule set." Knowledge-Based Systems **15**(7): 399-405.

Libby, P. (2001). "Current concepts of the pathogenesis of the acute coronary syndromes." Circulation **104**(3): 365-372.

Lin, W., M. A. Orgun and G. J. Williams (2002). An overview of temporal data mining. *Proceedings of the 1st Australian Data Mining Workshop* (pp 83-90).

McCormick, T., C. Rudin and D. Madigan (2011). "A hierarchical model for association rule mining of sequential events: An approach to automated medical symptom prediction." MIT Sloan Research Paper

McManus, D. D., J. Gore, J. Yarzebski, F. Spencer, D. Lessard and R. J. Goldberg (2011). "Recent Trends in the Incidence, Treatment, and Outcomes of Patients with STEMI and NSTEMI." The American Journal of Medicine **124**(1): 40-47.

Mehta, R. H., L. Parsons, S. V. Rao and E. D. Peterson (2012). "Association of Bleeding and In-Hospital Mortality in Black and White Patients With ST-Segment–Elevation Myocardial Infarction Receiving ReperfusionClinical Perspective." Circulation **125**(14): 1727-1734.

Miksch, S., Y. Shahar and P. Johnson (1997). Asbru: A task-specific, intention-based, and time-oriented language for representing skeletal plans.

Mooney, C. H. and J. F. Roddick (2013). "Sequential pattern mining--approaches and algorithms." ACM Computing Surveys (CSUR) **45**(2): 19.

Mukherjee, D., J. Fang, S. Chetcuti, M. Moscucci, E. Kline-Rogers and K. A. Eagle (2004). "Impact of combination evidence-based medical therapy on mortality in patients with acute coronary syndromes." Circulation **109**(6): 745-749.

Musen, M., S. Tu, A. Das and Y. Shahar (1996). "EON: A component-based approach to automation of protocol-directed therapy." Journal of the American Medical Informatics Association **3**(6): 367.

Nkambou, R., P. Fournier-Viger and E. M. Nguifo (2011). "Learning task models in ill-defined domain using an hybrid knowledge discovery framework." Knowledge-Based Systems **24**(1): 176-185.

Ohno-Machado, L., J. Gennari, S. Murphy, N. Jain, S. Tu, D. Oliver, E. Pattison-Gordon, R. Greenes, E. Shortliffe and G. Barnett (1998). "The guideline interchange format." Journal of the American Medical Informatics Association **5**(4): 357.

Ozden, B., S. Ramaswamy and A. Silberschatz (1998). Cyclic association rules. 14th International Conference on Data Engineering. IEEE.

Padmanabhan, B. and A. Tuzhilin (1998). A Belief-Driven Method for Discovering Unexpected Patterns. KDD **1998**: 94-100.

Padmanabhan, B. and A. Tuzhilin (2000). Small is beautiful: discovering the minimal set of unexpected patterns. Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

Patnaik, D., P. Butler, N. Ramakrishnan, L. Parida, B. J. Keller and D. A. Hanauer (2011). Experiences with mining temporal event sequences from electronic medical records: initial successes and some challenges. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

Pei, J., J. Han and W. Wang (2007). "Constraint-based sequential pattern mining: the pattern-growth methods." Journal of Intelligent Information Systems **28**(2): 133-160.

Peleg, M., A. Boxwala, O. Ogunyemi, Q. Zeng, S. Tu, R. Lacson, E. Bernstam, N. Ash, P. Mork and L. Ohno-Machado (2000). GLIF3: the evolution of a guideline representation format, American Medical Informatics Association.

Pray, K. and C. Ruiz (2005). "Mining expressive temporal associations from complex data." Machine Learning and Data Mining in Pattern Recognition: 384-394.

Raissi, C. and P. Poncelet (2007). Sampling for sequential pattern mining: From static databases to data streams. Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, IEEE.

Ramaswamy, S., S. Mahajan and A. Silberschatz (1998). On the discovery of interesting patterns in association rules. VLDB, Citeseer.

Reis, B. Y., I. S. Kohane and K. D. Mandl (2009). "Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study." BMJ: British Medical Journal **339**.

Richardson, W. C., D. M. Berwick, J. C. Bisgard, C. Newhall, R. Melker and T. Brennan (2000). "The Institute of Medicine report on medical errors." N Engl J Med **343**(9): 633.

Roddick, J. F. and M. Spiliopoulou (2002). "A survey of temporal knowledge discovery paradigms and methods." IEEE Transactions on Knowledge and data engineering: 750-767.

Roger, V. L., A. S. Go, D. M. Lloyd-Jones, E. J. Benjamin, J. D. Berry, W. B. Borden, D. M. Bravata, S. Dai, E. S. Ford and C. S. Fox (2012). "Heart Disease and Stroke Statistics—2012 Update A Report From the American Heart Association." Circulation **125**(1): e2-e220.

Ryan, T. J., E. M. Antman, N. H. Brooks, R. M. Califf, L. D. Hillis, L. F. Hiratzka, E. Rapaport, B. Riegel, R. O. Russell and E. E. Smith (1999). "1999 Update: ACC/AHA Guidelines for the Management of Patients With Acute Myocardial Infarction: Executive Summary and Recommendations A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee on Management of Acute Myocardial Infarction)." Circulation **100**(9): 1016-1030.

Schluter, T. and S. Conrad (2010). Mining Several kinds of Temporal association rules enhanced by Tree structures. Second International Conference on Information, Process, and Knowledge Management eKNOW'10, IEEE.

Sebastian, Y. and P. H. H. Then (2011). "Domain-driven KDD for mining functionally novel rules and linking disjoint medical hypotheses." Knowledge-Based Systems **24**(5): 609-620.

Silberschatz, A. and A. Tuzhilin (1995). On subjective measures of interestingness in knowledge discovery. KDD. KDD **95**: 275-281.

Silberschatz, A. and A. Tuzhilin (1996). "What makes patterns interesting in knowledge discovery systems." Knowledge and Data Engineering, IEEE Transactions on **8**(6): 970-974.

Simms, A. D., P. D. Baxter, B. A. Cattle, P. D. Batin, J. I. Wilson, R. M. West, A. S. Hall, C. F. Weston, J. E. Deanfield and K. A. Fox (2013). "An assessment of composite measures of hospital performance and associated mortality for patients with acute myocardial infarction. Analysis of individual hospital performance and outcome for the National Institute for Cardiovascular Outcomes Research (NICOR)." European Heart Journal: Acute Cardiovascular Care **2**(1): 9-18.

Smith Jr, S. C., T. E. Feldman, J. W. Hirshfeld Jr, A. K. Jacobs, M. J. Kern, S. B. King Iii, D. A. Morrison, W. W. O'Neill, H. V. Schaff and P. L. Whitlow (2006). "ACC/AHA/SCAI 2005 guideline update for percutaneous coronary intervention: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (ACC/AHA/SCAI Writing Committee to update the 2001 Guidelines for Percutaneous Coronary Intervention)." Journal of the American College of Cardiology **47**(1): e1-e1.

Somaraki, V., D. Broadbent, F. Coenen and S. Harding (2010). "Finding Temporal Patterns in Noisy Longitudinal Data: A Study in Diabetic Retinopathy." Advances in Data Mining. Applications and Theoretical Aspects: 418-431.

Spencer, F. A., J. Allegrone, R. J. Goldberg, J. M. Gore, K. A. Fox, C. B. Granger, R. H. Mehta, D. Brieger and G. Investigators (2004). "Association of statin therapy with outcomes of acute coronary syndromes: the GRACE study." Ann Intern Med **140**(11): 857-866.

Srikant, R. and R. Agrawal (1996). "Mining sequential patterns: Generalizations and performance improvements." Advances in Database Technology—EDBT'96: 1-17.

Steinhubl, S. R., A. Kastrati and P. B. Berger (2007). "Variation in the definitions of bleeding in clinical trials of patients with acute coronary syndromes and undergoing percutaneous coronary interventions and its impact on the apparent safety of antithrombotic drugs." American heart journal **154**(1): 3-11.

Subherwal, S., R. G. Bach, A. Y. Chen, B. F. Gage, S. V. Rao, L. K. Newby, T. Y. Wang, W. B. Gibler, E. M. Ohman and M. T. Roe (2009). "Baseline Risk of Major Bleeding in Non–ST-Segment–Elevation Myocardial Infarction The CRUSADE (Can Rapid risk stratification of Unstable angina patients Suppress ADverse outcomes with Early implementation of the ACC/AHA guidelines) Bleeding Score." Circulation **119**(14): 1873-1882.

Terenziani, P., S. Montani, A. Bottrighi, M. Torchio, G. Molino, L. Anselma and G. Correndo (2003). "Applying Artificial Intelligence to clinical guidelines: the GLARE approach." AI* IA 2003: Advances in Artificial Intelligence: 536-547.

Torpy, J. M., A. E. Burke and R. M. Glass (2010). "Acute coronary syndromes." JAMA **303**(1): 90-90.

Verma, K. and O. Vyas (2005). "Efficient calendar based temporal association rule." ACM SIGMOD Record **34**(3): 63-70.

Wang, J. and J. Han (2004). BIDE: Efficient mining of frequent closed sequences, *Proceedings. 20th International Conference on Data Engineering*. International Conference on Data Engineering IEEE.

Writing Group Members, W. Rosamond, K. Flegal, K. Furie, A. Go, K. Greenlund, N. Haase, S. M. Hailpern, M. Ho, V. Howard, B. Kissela, S. Kittner, D. Lloyd-Jones, M. McDermott, J. Meigs, C. Moy, G. Nichol, C. O'Donnell, V. Roger, P. Sorlie, J. Steinberger, T. Thom, M. Wilson, Y. Hong, f. t. A. H. A. S. Committee and Stroke Statistics Subcommittee (2008). "Heart Disease and Stroke Statistics--2008 Update: A Report From the American Heart Association Statistics Committee and Stroke Statistics Subcommittee." Circulation **117**(4): e25-146.

Yan, X., J. Han and R. Afshar (2003). CloSpan: Mining closed sequential patterns in large datasets.

Yeh, R. W., S. Sidney, M. Chandra, M. Sorel, J. V. Selby and A. S. Go (2010). "Population trends in the incidence and outcomes of acute myocardial infarction." N Engl J Med **362**(23): 2155-2165.

Zaki, M. J. (2001). "SPADE: An efficient algorithm for mining frequent sequences." Machine learning **42**(1-2): 31-60.

Zhang, S. and X. Wu (2011). "Fundamentals of association rules in data mining and knowledge discovery." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **1**(2): 97-116.

Zhang, T. (2000). Association rules. Knowledge Discovery and Data Mining. Current Issues and New Applications, Springer**:** 245-256.

Zimbrão, G., J. M. de Souza, V. T. de Almeida and W. A. da Silva (2002). An algorithm to discover calendar-based temporal association rules with item's lifespan restriction. The Second Workshop on Temporal Data Mining.