IDENTIFYING MYOCARDIAL INFARCTION PATIENTS USING AUTOMATED TEXT MINING IN FAMILY PRACTICE ELECTRONIC MEDICAL RECORDS: A VALIDATION STUDY OF EMERGING METHODS

by

Tezeta Fasika Mitiku

A thesis submitted to the Department of Community Health and Epidemiology
In conformity with the requirements for
the degree of Master of Science

Queen’s University
Kingston, Ontario, Canada
(September, 2008)

Copyright © Tezeta Fasika Mitiku, 2008
Abstract

**Background:** As Electronic Medical Records (EMRs) are being utilized increasingly in primary care physician’s offices, the potential exists to collect a vast amount of clinical information for research purposes. However, variation in clinician documentation of diagnoses makes it challenging to accurately identify diseases within the EMR. We set out to develop and validate a text-mining tool to identify MI diagnoses within the EMR.

**Methods:** We selected a random 5% sample from the 19,376 active adult patients in our EMR database. This sample of 1293 charts was reviewed by trained abstractors and used as the gold standard for the evaluation of the validity and reliability of the automated text-mining tool in identifying myocardial infarction patients. We also compared the results of the manual EMR abstraction with the hospitalization records of each patient to evaluate the validity and reliability of myocardial infarction diagnosis in the EMR. We manually reviewed all discordant records to investigate and categorize the reasons for discordance. Lastly, we compared the results of using administrative data versus EMR data for the measurement of selected MI quality indicators.

**Results:** When compared with the gold standard of manual chart review, the text-mining tool had a sensitivity of 97.4% (95% confidence interval [CI] 94.8%–99.2%), a specificity of 96.2% (CI 94.9%–97.4%), a positive predictive value (PPV) of 88.9% (CI 85.5%–92.7%) and a negative predictive value (NPV) of 99.1% (CI 98.6%–99.3%). When compared with the current standard of hospital discharge abstracts the EMR manual chart review had a sensitivity of 94.9% (CI 92.0%–97.7%), a specificity of 91.7% (CI 90.0%–93.5%) and PPV of 71.6% (CI 66.6%–79.2%) and a NPV of 98.8% (CI 98.1%–99.0%). The assessment of MI quality indicators were the same whether measured using the EMR or administrative data, with the exception of the proportion of patients on ASA (p<0.001).

**Conclusion:** The text-mining tool identified myocardial infarction diagnosis in the EMR with a high level of accuracy. In addition, EMRs may represent an important data source for a comprehensive identification of MI patients and the evaluation of quality of care.
Co-Authorship

This thesis presents research conducted by Tezeta Mitiku, in collaboration with Drs. Karen Tu, Linda Levesque, Douglas Lee and Jack Tu.

Manuscript: The research question regarding the validation of an automated text mining tool in capturing myocardial infarction diagnosis in the family practice electronic medical record was provided by Dr. Karen Tu, with collaboration from Tezeta Mitiku and Dr. Linda Levesque and Dr. Douglas Lee.

The primary investigator and author of the manuscripts was Tezeta Mitiku, who conducted statistical analyses with guidance from Mrs. Helen Guo and Dr. Linda Levesque, Dr. Karen Tu and Dr. Michelle Mattern. This study was the first validation conducted for this project and established the methodology for all future work utilizing data from this database.

Methodology, statistical design, and interpretation were a collaboration between Tezeta Mitiku and supervisors Dr. Linda Levesque and Dr. Karen Tu.

The Manuscript was written by Tezeta Mitiku, with editing assistance in consultation with Dr. Karen Tu, Dr. Linda Levesque and Dr. Michelle Mattern.
Acknowledgements

This work was made possible by the rigorous guidance of my thesis supervisors Drs. Linda Levesque and Karen Tu, whose unlimited patience and wisdom I am grateful for. I would also like to thank Drs. Jack Tu and Doug Lee for their critical input during the development of this project. The National Research Council of Canada provided technical expertise regarding the development of the automated text mining tool.

This work was supported by a Canadian Institutes of Health Research Team Grant in Cardiovascular Outcomes Research to the Canadian Cardiovascular Outcomes Research Team (CCORT) and a Queen’s Graduate Award.

I dedicate this thesis to my parents Aster Kebede and Fasika Mitiku for their unrelenting support and a life-time of sacrifices to facilitate my dreams.
# Table of Contents

Abstract ............................................................................................................................................ ii  
Co-Authorship ................................................................................................................................ iii 
Acknowledgements ......................................................................................................................... iv  
Table of Contents ........................................................................................................................... v  
List of Figures ............................................................................................................................... viii  
List of Tables .................................................................................................................................. ix  

Chapter 1 Introduction ..................................................................................................................... 1  
1.1 Thesis Outline ........................................................................................................................ 1  
1.2 Purpose ................................................................................................................................... 1  
1.3 Background and Rational ..................................................................................................... 2  
1.3.1 EMR Usage ..................................................................................................................... 2  
1.3.2 Measuring Quality of Care Using EMR Data ................................................................. 2  
1.4 Objectives ............................................................................................................................. 5  
1.4.1 Primary Objective ........................................................................................................... 5  
1.4.2 Secondary Objectives ..................................................................................................... 5  
1.5 References .............................................................................................................................. 6  

Chapter 2 Literature Review ......................................................................................................... 7  
2.1 Epidemiology of Myocardial Infarction ................................................................................ 7  
2.1.1 Defining Myocardial Infarction ...................................................................................... 7  
2.1.2 Mortality and Morbidity ................................................................................................. 9  
2.2 Quality of EMR Data ............................................................................................................. 9  
2.3 Impact of EMR Use ............................................................................................................... 9  
2.4 Perceptions of the EMR ....................................................................................................... 10  
2.5 Automated Data Extraction from EMRs ............................................................................ 10  
2.6 References ............................................................................................................................ 11  

Chapter 3 Methods Overview ..................................................................................................... 13  
3.1 Overview .............................................................................................................................. 13  
3.2 Privacy and Security ............................................................................................................ 13  
3.3 Data Transfer ....................................................................................................................... 14  
3.4 Consideration of Data Quality ........................................................................................... 14  
3.5 EMR Database ..................................................................................................................... 14  
3.6 Abstracting Diagnoses ......................................................................................................... 15
3.7 Sample Size and Power Calculations ................................................................. 15
   3.7.1 Method for calculating the estimated precision around the point estimates .... 16
3.8 Ethical Consideration .......................................................................................... 17
3.9 References ......................................................................................................... 18

Chapter 4 Using Data from Electronic Medical Records: Theory versus Practice ... 19
4.1 The Issue .................................................................................................................. 20
4.2 The Findings ............................................................................................................. 21
4.3 Data Extraction ......................................................................................................... 21
4.4 Data De-identification ............................................................................................ 21
4.5 Data Transfer .......................................................................................................... 23
4.6 Disease Identification ............................................................................................. 23
4.7 Philosophical Approach to EMR Data ................................................................... 24
4.8 Conclusions .......................................................................................................... 24
4.9 Implications .......................................................................................................... 24
4.10 Recommendations ............................................................................................... 25
4.11 References ............................................................................................................ 26

Chapter 5 Identifying Myocardial Infarction in Primary Care Electronic Medical Records ... 27
5.1 Abstract .................................................................................................................. 28
5.2 Introduction .......................................................................................................... 30
5.3 Methods ................................................................................................................. 32
   5.3.1 Physician and Patient Sampling ................................................................. 32
   5.3.2 Data Sources ................................................................................................. 32
       5.3.2.1 Electronic Medical Records ................................................................. 32
       5.3.2.2 CIHI Discharge Abstract Database .................................................... 36
       5.3.2.3 Ontario Health Insurance Plan (OHIP) Billing Records ...................... 37
       5.3.2.4 Ontario Drug Benefits (ODB) Claims ................................................ 37
   5.3.3 Identifying Myocardial Infarctions and Quality Indicators ......................... 37
       5.3.3.1 Manual Abstraction – Gold Standard ............................................... 37
       5.3.3.2 EMR Text-Mining Strategy ............................................................... 40
       5.3.3.3 CIHI-DAD – Linking Abstraction Set ................................................. 41
   5.3.4 Assessing Discordance .................................................................................. 41
   5.3.5 MI Quality Indicators ................................................................................... 41
   5.3.6 Analysis ......................................................................................................... 41
   5.3.7 Ethical Consideration .................................................................................... 42
5.4 Results.................................................................................................................... 42
  5.4.1 Intra-Observer and Inter-observer Reliability ....................................................... 46
  5.4.2 Text-Miner vs. Abstractor (Validation Set, N=977) .............................................. 46
    5.4.2.1 Discordance........................................................................................................ 49
  5.4.3 Manual Abstraction vs. CIHI (Abstraction Set, N=1293) ..................................... 49
    5.4.3.1 Discordance........................................................................................................ 49
  5.4.4 MI Quality Indicators: EMR vs. Administrative Data ........................................... 51
5.5 Discussion .................................................................................................................. 54
5.6 Conclusion .................................................................................................................. 56
5.7 References .................................................................................................................. 57

Chapter 6 Summary and Conclusions ............................................................................. 60
  6.1 Main Findings ........................................................................................................... 60
    6.1.1 General Summary ............................................................................................... 60
    6.1.2 Secondary Analyses ............................................................................................ 61
  6.2 Consistency of the Evidence .................................................................................... 61
  6.3 Strengths and Limitations ....................................................................................... 61
  6.4 Clinical Practice Implications ................................................................................ 63
  6.5 Future Research Directions ................................................................................... 63
  6.6 Conclusions ............................................................................................................. 64
  6.7 References ............................................................................................................. 65

Appendix A Ethics Approvals ........................................................................................... 66
List of Figures

Figure 1.1 Conceptual Model For MI Diagnosis Setting ................................................................. 4
Figure 2.1 Manifestations of AMI Presentation ............................................................................... 8
Figure 4.1 Overview of EMR Data ................................................................................................ 22
Figure 5.1 Overview of Physician Inclusion .................................................................................. 34
Figure 5.2 Overview of Active Adult Cohort Selection ................................................................. 35
Figure 5.3 Creation of Training and Validation Sets ................................................................. 39
Figure 5.4 EMR Population Compared to 2001 Standard Ontario Population ......................... 45
List of Tables

Table 5.1 Description of Measures of Validity and Reliability ..................................................... 43
Table 5.2 Physician Characteristics ............................................................................................... 44
Table 5.3 Patient Characteristics for Active Patients ..................................................................... 47
Table 5.4 Text Miner vs. Abstractor* ........................................................................................... 48
Table 5.5 EMR (Abstractor) vs. CIHI* .......................................................................................... 50
Table 5.6 Selected MI Quality Indicators Using Prescriptions ...................................................... 52
Table 5.7 Selected MI Quality Indicators Using Laboratory Values and Blood Pressure Measurements ........................................................................................................... 53
Chapter 1

Introduction

1.1 Thesis Outline

This thesis is structured as a series of papers reporting the results of a landmark Canadian study evaluating the feasibility of using information from primary care electronic medical records (EMRs) to measure cardiovascular related quality of care. The first manuscript outlines the major barriers to accessing and utilizing EMR data in Ontario. The second manuscript reports the results of a validation study of an automated text mining tool for the extraction of myocardial infarction (MI) diagnoses from EMRs. It also provides data on the validity and reliability of MI diagnoses obtained from the EMR when compared with the current standard of administrative hospitalization records. Lastly, it summarizes the results of a comparison between the use of EMR data and administrative health databases to evaluate selected MI care quality indicators. In the final chapter of this thesis, the major findings, implications, limitations and future directions of this work are discussed.

1.2 Purpose

The primary purpose of this thesis is to evaluate the validity and reliability of an automated text mining tool for identifying patients that have had an MI, using primary care electronic medical records (EMRs). This thesis will also evaluate the validity and reliability of the EMR as a source of information for the occurrence of MI. In addition, it will compare the use of information from the EMR and administrative databases for the measurement of selected MI quality indicators.
1.3 Background and Rational

1.3.1 EMR Usage

Electronic recording of an individual’s medical information using specialized software known as an electronic medical record (EMR), was first introduced to primary care practices in Ontario in the mid 1980s.\(^1\) At first, EMRs were utilized primarily for the storage of limited patient information such as laboratory tests and immunizations. However, the utilization of EMRs has evolved rapidly and has replaced the paper medical record in some clinical settings. In 2006, approximately 22 percent of primary care physicians in Canada (24 percent in Ontario) were using EMRs.\(^2\)

Electronic medical records in primary care offices contain comprehensive information on an individual’s medical history, health risks, and the team of individuals involved in providing care which are important for evaluating the processes and structure of care.\(^3\) They also include longitudinal information on laboratory and clinical measurements such as lipid profile and blood pressure. As this type of information is not currently available in administrative health databases, it would facilitate a comprehensive evaluation of patients’ quality of care.

1.3.2 Measuring Quality of Care Using EMR Data

Interest in the quality of health care has continued to grow in the past two decades. The catalyst for this is the need for greater accountability of health expenditures given increasingly scarce healthcare resources. There is no conclusive definition for measuring the quality of care. It has been described by some as “the measure of certain properties or attributes of the processes of care” and the “objective or goals of these processes” by others.\(^4\) It is in essence a “reflection of values and goals current in the medical care system”.\(^5\)

Although the measurement of quality of care has been expanding in Canada, it has mainly focused on specialized hospital-based care. As a result, little is known about the quality of primary and preventative health care in Canada.
An aspect of primary care that could have a significant impact on the health of Canadians is the prevention and management of coronary heart disease (CHD), its associated risk factors, and health events such as MI, as it is a leading cause of death in Canada. Moreover, primary care physicians have a prominent role in the identification, prevention and treatment of CHD. Yet, little is known about whether Canadian patients have access to high-quality primary care for the prevention and management of CHD. With the increased utilization of computerized patient records in primary care, the potential exists for the efficient extraction of information from this rich data source for the purposes of surveillance, chronic disease prevention, management, and health services research. However, important information about diagnoses, symptoms and some aspects of treatment are only available in the narrative free text portions of the EMR which limits their accessibility since this information must be processed and coded before it can be analyzed.

Another limitation of research using the EMR is the lack of validation studies of the diagnoses captured in these records. MI diagnoses captured in the family practice EMR may differ from those captured in the administrative databases. This is because administrative hospitalization databases only capture patients admitted to hospital as a result of an MI (i.e., acute MI). It does not capture an MI that is diagnosed in an ambulatory setting during routine screening (i.e., silent MI). This highlights the importance of validating not only the methods used to extract clinical information but also the EMR’s ability to capture certain diagnoses. Figure 1.1 outlines the conceptual model for the various settings where MI diagnoses can be captured.

Given the quantity of patient records and encounters that need to be reviewed, and the variety of ways information is recorded, an automated text mining tool was developed by researchers at the Institute for Clinical Evaluative Sciences (ICES) in partnership with researchers at the National Research Council of Canada (NRC) to facilitate the extraction of such diagnoses from the EMR. However, this text mining tool requires validation.
Figure 1.1 Conceptual Model For MI Diagnosis Setting

- Circle: EMR CAPTURED
- X: CIHI CAPTURED
1.4 Objectives

1.4.1 Primary Objective
To assess the validity and reliability of an automated text mining tool for identifying patients with an MI compared with a manual review of the primary care EMR as the gold standard.

1.4.2 Secondary Objectives
(i) To assess the validity and reliability of the primary care EMR for identifying patients with an MI compared with the CIHI hospital separation diagnosis as the gold standard.
(ii) To explore the reasons for discordance in identifying patients with MI between the text mining tool, the EMR and the hospital separations database (CIHI).
(iii) To explore the generalizability of the text mining tool for identifying patients with an MI.
(iv) To compare the use of administrative and EMR data for the measurement of selected MI quality indicators.
1.5 References


2.1 Epidemiology of Myocardial Infarction

2.1.1 Defining Myocardial Infarction

An acute coronary syndrome (ACS) is a reduction in blood flow to the tissues of the heart. However, if this blood flow is interrupted due to a blockage of coronary vessels, this results in permanent death of the tissues of the heart, which is an acute MI (AMI).\(^1\)

An individual suffering from an AMI can present, and be diagnosed, in a variety of manners depending upon the severity and type of symptoms experienced. Individuals suffering from an AMI will most often present to hospital with serious symptoms such as chest pain and shortness of breath. These individuals then undergo a series of diagnostic tests in order to distinguish between two types of AMI: an ST segment elevation myocardial infarction (STEMI), and a Non ST-segment elevation myocardial infarction (Non-STEMI) or unstable angina (UA).\(^2\) However, in certain cases individuals may have an AMI with less severe symptoms that resolve on their own. In these instances the person may not seek medical attention and this event is classified as a silent MI. The latter can potentially be identified in the family practice EMR record during routine screening or the investigation of patient symptoms. However, they are less likely to be captured by hospitalization records. The Framingham study estimated the rate of silent or unrecognized MI to be between 5-10 %.\(^3,4\) Figure 2.1 depicts the various manifestations of MI.

Currently, the identification of patients with MI for research purposes is accomplished by query of administrative health databases, in particular hospital discharge abstract databases.
Figure 2.1 Manifestations of AMI Presentation
2.1.2 Mortality and Morbidity

CHD is one of the leading causes of mortality and morbidity in Canada. The financial impact of this disease is also tremendous. A 1998 report estimated that heart disease costs the Canadian economy approximately 18.4 billion dollars, representing the largest cost share of any disease. Although the rate of mortality for heart disease has declined over the past three decades as a result of improved prevention and treatments, recent data suggests that some of the major risk factors for CHD including obesity, hypertension and diabetes, are increasing. This coupled with an aging population is disconcerting as it has the potential to increase the incidence of adverse health events related to CHD such as MI.

2.2 Quality of EMR Data

Studies evaluating the quality of information documented within EMRs have focused on various aspects of these systems. A systematic review evaluating the quality and scope of EMR data in primary care defined data quality as a measure of the reliability of the data. Studies have found that because of the tremendous variation in the methods used to measure the quality of data within EMRs it is difficult to compare the results from different studies. However, other studies suggest that EMR use can in fact improve data accuracy.

2.3 Impact of EMR Use

EMRs have been hailed for their potential to increase the efficiency and integration of the health care system. A systematic review found that information technology can improve the delivery of care by promoting adherence to clinical guidelines and reducing medical errors. However, a recent study evaluating the impact of computer based patient records found that the simple implementation of the technology does not in and of itself lead to an improvement in the quality of patient care. This demonstrates the importance of research to evaluate efficient methods for the automated extraction of information from patient records in order to facilitate
continuity of care, population health research, and surveillance systems which are well established methods for the improvement of patient care. This will ensure that the enormous investment in new technologies translates into an improvement in the structure and delivery of care.17

2.4 Perceptions of the EMR

A recent survey, reports that 88 percent of Canadians believe that EMRs will lead to more timely diagnosis of medical conditions.18 Surveys have also been conducted on Canadian physicians evaluating their perception at three months and one year following the adoption of EMRs into their practice. The majority of physicians did not perceive any benefit after three months. Yet at one year, the majority attributed improvements to quality of care and reduction of wait times to the implementation of the EMR. This would suggest there is an initial learning and adjustment period that physicians’ must overcome before they perceive the intended benefits of the EMR. Federal Minister of Health Tony Clement echoed these sentiments when he stated “Once fully implemented, private and secure electronic health records will increase efficiencies, reduce wait times and result in significant savings in our health care system.” 19

2.5 Automated Data Extraction from EMRs

Studies have been conducted internationally evaluating the validity and reliability of automated text mining tools for the accurate diagnosis of patients with various conditions such as hypertension, pharyngitis, smoking status and wrist fractures.20-21 The results of these studies have been variable. Early text mining tools had high rates of misclassification.22-24 In recent years, many studies of contemporary tools have reported high levels of accuracy and precision in identifying disease conditions from narrative free text 25 26. However, commercially produced text mining tools are costly and most of the academic systems are not yet freely available for use.
2.6 References


(15) McDonald CJ. The barriers to electronic medical record systems and how to overcome them. J Am Med Inform Assoc 1997; 4(3):213-221.


(25) Mamlin BW, Heinze DT, McDonald CJ. Automated extraction and normalization of findings from cancer-related free-text radiology reports. AMIA Annu Symp Proc 2003;420-424.
Chapter 3
Methods Overview

3.1 Overview

We retrospectively reviewed the records of physicians using the Practice Solutions® EMR platform to document all clinical encounters, as well as laboratory results and hospitalizations using discharge summaries. Practice Solutions® EMR is a leading software company and is currently being used by over 400 primary care physicians in Ontario, encompassing more than 400,000 patients. This EMR system has been in existence for over 15 years and currently at least 200 of the physician users have at least 3 years of patient information. The EMRs were both manually reviewed and electronically abstracted using an automated text mining tool, to classify the records as to the absence or presence of 20 different risk factors, disease conditions and procedures related to cardiovascular disease. The first disease condition selected to validate the text mining tool was myocardial infarction (MI). MI was selected because of its tremendous impact on the morbidity and mortality of Canadians. In addition, it has the lowest prevalence of all of the selected disease conditions and thus required us to devise a sampling strategy that would ensure adequate power of the validation study for this outcome.

3.2 Privacy and Security

Anonymized patient level data was sent electronically to the Institute for Clinical Evaluative Sciences (ICES) through high level encryption software and a secure and continually monitored internet connection. The data was only accessible at ICES, from prespecified stand alone computer terminals connected to stand alone servers housed in specially developed monitored and secure rooms with restricted access. This is permitted because ICES is named as a prescribed entity under Section 45 of Ontario’s Personal Health Information Protection Act.
(PHIPA), 2004, which provides physicians the legal authority to share individualized health data with a prescribed entity without patient consent.

### 3.3 Data Transfer

Data from the EMR record was extracted directly from the physician’s EMR system using a Practice Solutions® designed “plug-in”. The plug-in is programmed to access the records database, identify all the patients from participating physicians, and extract patient demographic data and all cardiac-relevant fields (i.e., progress notes, consult letters, and lab results). During the extraction phase, the plugin automatically anonymizes these records by comparing the contents of the patient and physician names field and replacing any identifying instances in the patient record with a string of capital letters. The plugin then completes the process by compressing the four extracted data files into a zip archive which is then encrypted and uploaded to ICES via a secure, password-protected registry portal. These data were used to obtain information on 20 disease conditions, risk factors and procedures related to cardiovascular disease for the validation of the automated text mining software and the validation of the EMR diagnoses. MI Cases were identified using both manual and automated electronic methods.

### 3.4 Consideration of Data Quality

Quality assurance procedures including random manual reviews were employed in order to ensure that the EMR data were not compromised during processing at ICES.

### 3.5 EMR Database

The EMR database contained 170 variables, 36 of which were free text fields that were manually inputted by the physician and clinic staff or scanned into the EMR using optical character recognition (OCR) software. Optical character recognition is a process whereby textual information that is found within documents that are in non-textual file formats such as JPEG and TIFF, are converted to narrative text. Free text variables included the progress notes, consultation
letters, and hospital discharge summaries. The free text variables were unstructured and required processing in order to obtain coded clinical information that could be analyzed using traditional methods. The remaining variables were from coded portions of the EMR and arrived in the database ready for analysis. These variables include laboratory values, anthropomorphic measures and vital statistics.

3.6 Abstracting Diagnoses

To assess the validity of the text mining tool, diagnoses were first manually abstracted from the EMRs using a standardized data extraction form. This form classified the occurrence of disease conditions using the same linguistic rules as those developed for the text mining tool. This form was pre-tested by the three independent and blinded physician reviewers using a random sample of 15 EMRs. Reviewers then discussed the interpretation of different questions on the form, and adjustments were made as required. The manually abstracted MI diagnoses were used as the “gold standard” against which MI diagnoses, and other disease conditions obtained using the automated text mining tool were compared. To assess inter-observer reliability, a random 5% sample of EMRs was selected and the reviewers independently extracted the necessary information from the same records. A random sample of the adult patients in the database (N=1293) were then divided among the reviewers. Data abstracters were blinded to the study hypothesis and the purpose of the data abstraction. For the automated data extraction, the diagnosis of MI was identified through a customized text mining tool developed by researchers at the Institute for Clinical Evaluative Sciences (ICES) in collaboration with the National Research Council of Canada (NRC).

3.7 Sample Size and Power Calculations

Most diagnostic accuracy studies do not publish their methods for evaluating sample size and there are competing perspectives on how best to address the issue of sample size for these
types of studies. We used the method most often presented in the literature after a review of the literature in the area of diagnostic accuracy.¹

We estimated that the incidence rate of MI for the EMR population was approximately 1% per year as determined by using 2004 CIHI data obtained from the Chronic Disease Infobase surveillance system maintained by the Public Health Agency of Canada². This was expected to yield approximately 200 cases of MI in any given year for our sample of 19,376 active adult patients.

The results of a sample size calculation using a method based upon the desired precision of the estimates of sensitivity and specificity using the width of the confidence interval are as follow ³⁵. For our sample size of 19,376 patients we expected our estimate of sensitivity to be accurate to within +/- 7.0% and our specificity to within +/- 0.7%.

3.7.1 Method for calculating the estimated precision around the point estimates

Width of Confidence Intervals: Of the 40,228 patients in our source population approximately 20,000 will be adults who meet the inclusion and exclusion criteria. If we use an estimated incidence rate of MI of approximately 1% per year, we expect 200 patients per year with a heart attack as confirmed by clinical review. Sensitivity will be based on these 200 heart attack cases, while specificity will be based on patients without heart attack.

Suppose ps is the sensitivity (proportion of MI cases correctly identified by the data extraction tool) and pe is the specificity (proportion of non-MI cases correctly identified by the data extraction tool), then the 95% CI for ps is ps +/-1.96*SEs, and the 95% CI for pe is pe +/- 1.96*SEe. The standard error (SE) of a proportion is the sqrt(p*q/n). SE will be the largest when p=q=0.5. This implies that the sensitivity can be estimated within +/- 1.96*sqrt(0.5*0.5/200) = 0.069 (ie, +/- 7%), while specificity can be estimated within 0.007 (ie, +/- 0.7%) at a confidence level of at least 95%. 
This is a conservative estimate as we expect greater than 200 cases since we have many years of follow up in our database. In addition, we expect the sensitivity and specificity of our tool to be greater than 0.5.

3.8 Ethical Consideration

This study obtained ethics approval from the Queen’s University Health Sciences and the Sunnybrook Health Sciences Centre Research Ethics Boards prior to commencing (Appendix A). All data were encrypted and anonymized and no patients were contacted for this study. In addition, all study results are reported in aggregate form only.
3.9 References


Chapter 4

Using Data from Electronic Medical Records: Theory versus Practice

This chapter is the first of two manuscripts comprising this thesis. It provides an overview of the challenges inherent in the utilization of data from electronic medical records for research and evaluation purposes. It outlines the major barriers to accessing and utilizing this data source and provides the preface to the second manuscript which presents the results of a study utilizing this data source. The chapter was published as Mitiku, T.F. and K. Tu. 2008. Using data from electronic medical records: Theory versus practice. *Health Care Quarterly* 11(4): 23-25.

**Word Count:** 1289  
**Tables:** 0  
**Figures:** 1

**Keywords:** Electronic Medical Records, Primary Care, Quality of Care
4.1 The Issue

In Canada, the measurement of quality of healthcare has historically focused on specialized hospital-based care. Considerably less is known about the quality of care provided in the offices of primary care physicians. Primary care research has relied on data collected manually from physicians’ offices or from administrative databases. Manual data collection from paper-based patient charts in primary care physicians’ offices is costly and time consuming, and often only a small portion of the information in the charts is useable due to the lack of uniform documentation. Although data from administrative databases are more readily accessible and encompass the entire population, they are limited in their depth of clinical information.

The increased use of electronic medical records (EMRs) by primary care physicians presents an opportunity for the efficient extraction and use of large quantities of clinical information. EMRs capture comprehensive longitudinal information on individual patients not available from other sources, including important risk factors for health outcomes such as smoking status, family history and clinical and laboratory measurements (e.g., blood pressure, body mass index and cholesterol levels).

EMR use in Canada has expanded rapidly, with many provincial governments providing funding for primary care physicians to adopt EMRs into their practices. In 2006, approximately 22% of primary care physicians in Canada (24% in Ontario) were using EMRs.1 Canada Health Infoway, an organization established to spearhead the movement toward a national electronic health records system, has estimated that $10–12 billion dollars are needed to establish basic EMR infrastructure in Canada by 2015.2

Researchers in the United States and the United Kingdom have demonstrated the utility of EMR data for chronic disease surveillance, management and prevention and for health services research.3-5 However, research organizations in these countries obtain the data from a common EMR format. In Canada, there are multiple provincially accredited EMR formats in use (currently
11 in Ontario and 10 in Alberta, and there is no single EMR software program accredited for use in those provinces that have an EMR funding and accreditation program in place. Although Canada Infoway and the provincial agencies are setting standards for interoperability between EMRs, the multitude of software vendors represents a particular challenge for research. In theory, gathering data that are already in an electronic format should be uncomplicated; but in practice, extracting comprehensive information from even a single EMR poses many challenges.

4.2 The Findings

Over the past two years, we have been working with a leading EMR software vendor in Ontario to conduct a pilot study to evaluate the feasibility of using data from an EMR to measure cardiovascular-related primary care quality indicators. In so doing, we have gained a better appreciation of the challenges inherent in this undertaking.

4.3 Data Extraction

To extract data from an EMR, it is necessary to engage an EMR software vendor to develop methods for getting information from the dynamic database environment of the EMR into a format that contains specific clinical information under exact variable headings. Data-extraction programs developed by vendors are designed at a specific point in time on a particular version of the EMR software. However, these software programs are constantly being modified and upgraded, thus requiring ongoing engagement between users and vendor software technicians. Further, physicians’ offices have different versions of the EMR software depending on when it was installed and whether they have chosen to upgrade their software. This presents a challenge to establishing automated data-collection procedures.

4.4 Data De-identification

Data contained within an EMR comes in both a structured and an unstructured format (Figure 4.1). Structured data not containing identifying information can be entered numerically
Figure 4.1 Overview of EMR Data

**Structured Data**

**Identifying Information**
- Name
- Address
- Telephone number
- Postal code
- Date of birth
- Health card number

**Numerical Data**
- Laboratory results
- Blood pressures
- Height
- Weight

**Predefined List**
- Prescriptions

**Unstructured Data**

**Non-nominal**
- Cumulative patient profiles
- Progress notes

**Nominal**
- Investigations/diagnostic tests
- Radiology/pathology reports
- Consultation letters
- Hospital discharge summaries

---

- Data elements that are easily partitioned
- Data elements requiring free text searching and anonymization

---

22
(e.g., blood pressure readings) or as single words or word combinations from a finite list (e.g.,
prescriptions). Thus, these structured data are readily extracted and can be linked to date, age,
gender and coded patient identification, but they do not contain any other identifying information.
This exceeds what is currently available in administrative data. EMR data contained in an
unstructured or free text format can further enhance research capabilities, but they carry the risk
of containing personal identifying information. Data that are unstructured or free text can be
classified as unlikely to contain identifying information (e.g., progress notes) or highly likely to
contain identifying information (e.g., consultation letters). This can present a further challenge in
that some EMRs contain segments of free text (e.g., a consultation letter) that are captured in a
file format such as PDF or TIFF. The text in these image files requires further processing in order
to be converted into a searchable format whereby identifying information can be automatically
removed or anonymized.

4.5 Data Transfer

Once data have been extracted and anonymized, they must be transported to a central
location where they can be analyzed. This is a costly but necessary step that requires sophisticated
technological security and expertise along with regular risk assessment.

4.6 Disease Identification

Identifying patients with particular disease conditions in the EMR database is necessary
to measure the presence of disease-specific quality indicators and to examine patterns of practice;
this requires the development of automated techniques. While some disease conditions can be
deduced from prescription profiles or laboratory values, many require examination of the free
text. Simply searching for the occurrence of a particular disease condition is not sufficient
because the inclusion of the condition in the free text does not necessarily mean the patient has
that condition. For instance, the free text might note “no evidence of a myocardial infarction,”
“mother had an MI,” “rule out MI” or “patient had a heart attack.” All these phrases relate to the
same disease condition – myocardial infarction – but each conveys a different message. Furthermore, physicians often use abbreviations and acronyms that are unique. We are currently developing methods to address variations in physician documentation to better identify patients with specific disease conditions.

4.7 Philosophical Approach to EMR Data

A philosophical approach to analyzing EMR data was adopted in order to make use of existing data generated by physicians already using EMR software. This approach was chosen to lessen the disruption to physicians’ established clinical flow, making it less likely to meet with resistance. In addition, this approach is not subject to the selection bias that is present in methods that require special input or coding by physicians who are likely to have varying levels of participation in research studies that use a prospective approach requiring them to perform additional data-entry steps.

4.8 Conclusions

In theory, accessing and analyzing data contained within an EMR appears to be a straightforward process. In reality, there are many barriers and challenges that need to be overcome to set up an EMR database that preserves the richness of information contained within the EMR.

4.9 Implications

Although we anticipate that the initial investment for developing an EMR database will be substantial, ongoing maintenance and upkeep with software changes and dealing with multiple software vendors are also likely to incur substantial costs. In addition, the development and validation of automated ways of processing and extracting information from EMR records is costly and time consuming. The large estimated expenditure for EMR implementation do not take into account costs for the extraction and use of data for research and evaluation purposes. This
initial investment of time and finances is necessary to use this rich data source in Canada to its fullest. To our knowledge, this retrospective approach to capturing and evaluating EMR data is a leading initiative in Canada.

4.10 Recommendations

If seamless portability of data between accredited EMR vendors cannot be attained, then the number of such vendors within provinces and across the country should be considerably reduced so that the substantial investment in EMRs maximize their research potential in addition to achieving enhanced patient care goals.
4.11 References


Chapter 5
Identifying Myocardial Infarction in
Primary Care Electronic Medical Records

This chapter is the second and final manuscript comprising this thesis. It presents the findings of the primary and secondary objectives of this project designed to evaluate the validity and reliability of emerging methods for the extraction and use of data from primary care EMRs.

Word Count: 5,689  Tables: 7  Figures: 4

Keywords: Myocardial Infarction, Electronic Medical Records, Quality Indicators
5.1 Abstract

**Background:** As Electronic Medical Records (EMRs) are being utilized increasingly in primary care physician’s offices, the potential exists to collect a vast amount of clinical information for research purposes. However, variation in clinician documentation of diagnoses makes it challenging to accurately identify diseases with this data source. Methods for the automated identification of diseases from the narrative free-text portions of EMRs have been developed, however, none have been validated for myocardial infarction specifically, and little work has been done to evaluate them in a Canadian context.1-6 We set out to develop and validate a text-mining tool to identify MI diagnoses within the EMR.

**Methods:** We selected a random 5% sample from the 19,376 active adult patients in our EMR database drawn from a convenience sample of 17 physicians practicing in Ontario, Canada, all of whom were using the same EMR software. To ensure that an adequate number of MIs were in our sample, we added patients with MI terminology in their EMR as identified through a manual search strategy. This final sample of 1293 charts was reviewed by trained abstractors and used as the gold standard for the evaluation of the validity and reliability of the automated text-mining tool in identifying patients that had suffered from a myocardial infarction. We also compared the results of the manual EMR abstraction with the hospitalization records of each patient to evaluate the validity and reliability of myocardial infarction diagnoses in the EMR. We manually reviewed all discordant records to investigate and categorize the reasons for discordance. Lastly, we compared the results of using administrative data versus EMR data for the measurement of selected MI quality indicators.

**Results:** When compared with the gold standard of manual chart review, the text-mining tool had a sensitivity of 97.4% (95% confidence interval [CI] 94.8%–99.2%), a specificity of 96.2% (CI 94.9%–97.4%), a positive predictive value (PPV) of 88.9% (CI 85.5%–92.7%) and a negative...
predictive value (NPV) of 99.1% (CI 98.6%–99.3%). When compared with the current standard of hospital discharge abstracts for identifying myocardial infarction patients, the EMR manual chart review had a sensitivity of 94.9% (CI 92.0%–97.7%), a specificity of 91.7% (CI 90.0%–93.5%) and PPV of 71.6% (CI 66.6%–79.2%) and a NPV of 98.8% (CI 98.1%–99.0%). Of the 88 charts where the abstractors found an MI diagnosis in the EMR with no information in the hospitalization data, 15.5 % were likely silent MIs. The results of the assessment of MI quality indicators were the same whether measured using the EMR or administrative data, with the exception of the proportion of patients on ASA (p<0.001).

**Conclusion:** The text-mining tool identified myocardial infarction diagnoses in the EMR with a high level of accuracy. In addition, EMRs may represent an important data source for a comprehensive identification of MI patients and the evaluation of quality of care.
5.2 Introduction

The increased utilization of computerized patient records in primary care presents an opportunity to access large amounts of clinical information not previously available. As uptake of this technology expands it is important to develop methods to efficiently and accurately extract diagnosis from these records. It has been established in previous studies “that the accurate identification of patient cohorts with a specific disease is important for the measurement and surveillance of quality of care, the management of chronic disease conditions and the creation of prospective research cohorts”. 2,7,8

However, the challenge in using the electronic medical record (EMR) for research purposes is that diagnoses such as myocardial infarction (MI) are entered in a free-text format. Furthermore, physicians record information in a variety of ways using different acronyms, short forms or language such as “heart attack” “hrt atk” “MI” or “AMI”, all denoting myocardial infarction. Given the quantity of patient records and encounters that need to be reviewed and the variety of ways information is recorded, automated text mining tools are needed to efficiently and accurately extract information from the EMR.

The use of computer technology to identify concepts within the narrative free-text of medical records has been evolving since the time health records began to be electronically documented. Commercially available text-mining programs to identify specific disease conditions within health records are either costly or are not designed with the flexibility necessary to easily classify new disease conditions.9

The current standard for identifying acute MI patients in Ontario and Canada is hospital discharge abstract databases. In Ontario, this information is obtained from the Canadian Institute for Health Information (CIHI) Discharge Abstract Database (DAD). However, in certain cases, individuals may have an MI with less severe symptoms or atypical symptoms that resolve on their own. In these instances, the person may not seek medical attention and this is defined as a silent
MI. The Framingham Study estimated the prevalence of silent or unrecognized MI to be between 5 and 10%. However, silent MIs, those diagnosed out of the hospital setting or MIs that occur out of country, out of province or before 1988 (CIHI data for Ontario is available to us beginning in 1988) are not captured in CIHI. Furthermore, administrative databases do not capture comprehensive information on patients, including important risk factors for health outcomes such as smoking status, family history and clinical and laboratory measurements (e.g., blood pressure measurements, body mass indices, cholesterol and glucose levels). As such, their ability to assess indicators of quality of care may be limited. It has been suggested that EMRs represent a source of clinical data that complements or could even replace administrative sources for the purposes of identifying patient populations and measuring quality of care. Previous quality indicator measurements for MI care in Ontario have been based on data from administrative databases and chart abstraction. However, prescription administrative data in Ontario is limited by the unavailability of complete prescription information for patients less than 65 years of age and chart abstraction is a cost prohibitive process.

In this study, we evaluated the performance of an automated text mining tool in identifying patients that had an MI, using data from one of the largest EMR software vendors in Canada. The validity and reliability of using the EMR as a data source to identify MI patients compared with the current gold standard of CIHI administrative data was also assessed. We manually reviewed all discordant records for each comparison in order to determine the reasons for discordance. In addition, to evaluate the comprehensiveness of EMRs, we compared the results of using administrative data with those of EMR data for the measurement of selected MI quality indicators.
5.3 Methods

5.3.1 Physician and Patient Sampling

A convenience sample of 17 physicians using Practice Solutions® EMR contributed their data to the EMR database. These physicians were recruited: (1) at an annual general meeting held by Practice Solutions® (the company that provides the software for the EMRs) and (2) through referral by participating physicians. Of the 30 physicians that were registered at the event, 21 agreed to participate in the study (figure 5.1). Participating physicians had to have a minimum of two years of data on their EMR system before being included in this study. We calculated physician time on the EMR from the date the data was obtained. This criterion was established in order to ensure that a reasonable amount of health information would be recorded in the physician EMR. Eight of these physicians were not included in the current study as they did not meet this criterion. An additional four physicians meeting the inclusion criteria were recruited by referral from participating physicians, bringing the total number of participating physicians to 17. We asked physicians to report their duration of EMR usage when recruiting for study participation. Duration varied from a maximum of 18 years to a minimum of two.

The 17 physician practices provided data for 40,228 patients to the EMR database. (figure 5.2) This database was cleaned by removing patients with invalid health card numbers and date of birth, resulting in 32,008 patients. The cohort was then further restricted to active adult patients only, defined as having at least two physician visits within the three years prior to the data download date. Adults were defined as all patients greater than or equal to 20 years on or before December 31, 2007. These exclusion criteria reduced our sample size to 19,376 patients.

5.3.2 Data Sources

5.3.2.1 Electronic Medical Records

Data from the EMR was extracted directly from the physicians’ EMR systems, between June and December of 2007, using a Practice Solutions®-designed ‘plug-in’ that automatically
pulled the relevant patient information and anonymized, partitioned, encrypted and transmitted it through a monitored and secured data transmission line.
Figure 5.1 Overview of Physician Inclusion
Figure 5.2 Overview of Active Adult Cohort Selection
The data contained within the EMR is in both a structured and unstructured format.14 Structured data include vital statistics, blood pressure readings, anthropomorphic measures, lab values and prescriptions. Unstructured data is the free-text data, and includes the cumulative patient profile (past medical history), progress notes, diagnostic tests, hospital discharge summaries and specialist consultation letters. Identification of MI patients relies solely on text recognition within the free-text as there are no standard lab test results contained within an EMR or a prescription for a medication that is exclusive to MI patients that would have allowed us to confine our search for MI patients within the structured data.

5.3.2.2 CIHI Discharge Abstract Database

The CIHI-DAD captures a patient’s date of admission and discharge, demographics, a most responsible diagnosis and up to 15 secondary diagnostic codes using the International Classification of Diseases and Related Health Problems, 9th and 10th Revisions (ICD-9 prior to fiscal year 2002 [410, 412] and ICD-10 after fiscal year 2002 [I21, I22, I252]) developed by the World Health Organization. We used the most responsible diagnosis or all of the secondary diagnostic codes to indicate that a patient had an MI. It has been demonstrated that administrative data are accurate in reflecting the incidence and prevalence of acute conditions that result in hospital admissions such as myocardial infarction15-18 The accuracy of AMI coding in the CIHI database has also previously been assessed using a multi-centre chart audit whereby hospital records were reviewed for a sample of patients coded as having had an MI as the most responsible diagnosis. The study reported a sensitivity of 95% and a specificity of 88% for capturing patients admitted to hospital for treatment of an acute myocardial infarction.19 Patients from the EMR were linked to the CIHI-DAD via encrypted health card numbers. Data within the EMR database were restricted to prior to March 31, 2007 in order to match the availability of CIHI data.
5.3.2.3 Ontario Health Insurance Plan (OHIP) Billing Records

The OHIP database contains information on most claims paid by the Ontario Health Insurance Plan to health care providers allowed to bill OHIP including physicians, laboratories, and out-of-province providers for care provided to registered Ontarians. Physicians using alternative billing systems are not captured by this database; however, approximately 95% of Ontario physician billings are captured in OHIP. These data are available from July 1991 to June 2007. We used OHIP data to measure the proportion of definite MI patients in our database that had blood glucose and cholesterol measurements taken as one indication of post-MI quality of care.

5.3.2.4 Ontario Drug Benefits (ODB) Claims

The Ontario Drug Benefit (ODB) database contains detailed information on prescription drugs covered by this health insurance program and dispensed to registered residents. This includes persons 65 years of age and older and those covered under ODB programs which are based on income insufficiency. ODB data is available from April 1990 to June 2007. This data source was used to identify drugs prescribed to MI patients ≥65 years of age for the measurement of selected quality indicators.

5.3.3 Identifying Myocardial Infarctions and Quality Indicators

5.3.3.1 Manual Abstraction – Gold Standard

We selected a random 5% sample of the 19,376 active adult patients in the EMR database (N =969). In addition, we saturated this sample with 324 additional patients with MI terminology in their EMR to ensure that an adequate number of potential events were included in our validation study. The additional MI patients were identified through a manual search of the EMR database using the find function of Microsoft ® Word Pad version 5.1. We searched for the presence of the terms “myocardial infarction”, “MI” and “heart attack” in the records of the 17
physicians that were included in our database. The resulting abstraction data set consisted of 1293 patients (Figure 5.3). These 1293 patient had their charts manually reviewed and the results were used as the gold standard against which the automated text mining tool was evaluated.

This abstraction data set was divided into a training set and a validation set. Fifty percent (50 %) of the patients from nine randomly selected physicians in our abstraction data set comprised the 316 MI patients training set. The purpose of this training set was to train the text mining tool to identify language and phrasing indicative of myocardial infarction diagnoses. The validation set consisted of the remaining 977 patients in the abstraction set. The validation data set was used to evaluate the validity and reliability of the text mining tool. It was also used to compare the performance of the text mining tool on patient records from physicians that were used to train the tool with its performance on patient records from “naïve” physicians in order to assess the adaptability of the text mining tool.

MI diagnoses were manually identified from the abstraction data set (N=1293) using a standardized data extraction manual developed by a family physician and a cardiologist, both study investigators. Three trained, independent and blinded EMR abstractors pre-tested the data abstraction rules using a random sample of 15 training EMRs and the necessary adjustments were made to these rules to enhance clarity and consistency of MI identification. The abstractors reviewed the entire EMR for each patient, comprised of a single cumulative patient profile (CPP) and a variable number of appointments. A patient was identified as having an MI if it was documented in the free-text of the cumulative patient profile, the progress notes, consultation letters, hospital discharge summaries or diagnostic tests. ECG evidence of MI was taken to be indicative of an MI, echocardiographic evidence alone described as wall motion abnormality or fixed defect with any qualifiers such as poor quality study or breast attenuation were taken to be possible MI.
Figure 5.3 Creation of Training and Validation Sets

19,376 Active pts

5% random sample from each physician

969 pts

324 pts w/ MI terms

17 MDs 1293 pts

50% Random Physician Sample

9 MDs 633 pts

50% Random Patient Sample

316 pts Training Set

8 MDs 660 pts

977 pts Validation Set
The CPP and each entry into the EMR were classified in one of the following categories: (1) unknown MI Status, (2) definite MI, or (3) possible MI. Family history of MI was also classified in combination with any of the categories above. Then each patient was scored globally as to whether they had an MI or not. If a patient had an entry for a definite MI, then the patient was considered to have had an MI. The patients in category 1,3 or family history constituted the “no MI” patients for the analysis. The manually abstracted results were used as the “gold standard” against which MI diagnoses obtained using the automated data mining tool was compared.

To assess inter-abstractor reliability, a random sample of 5% of the patients from the validation set was selected and the reviewers independently extracted information from the same records. To assess intra-abstractor reliability each abstractor re-abstracted approximately 2.5% of charts they previously completed.

5.3.3.2 EMR Text-Mining Strategy

For the automated data extraction, the diagnosis of MI was identified through a customized text mining tool. The tool was developed using the same clinical rules developed for the manual chart abstraction. The tool operated in multiple phases, the first involved the parsing of the text into individual sentences and the second involved fact extraction from these sentences and classification. Text classification used a support vector machine (SVM), which is a supervised learning method using regression. The SVM learns by analyzing the charts of patients with and without a specific disease condition. The MIs identified using this automated text mining tool was then compared with those manually abstracted from the EMR (gold standard).
5.3.3 CIHI-DAD – Linking Abstraction Set

We linked all of the patients in the abstraction set (N=1293) to the CIHI-DAD and searched the patients’ hospitalization records for the MI codes described previously. These results were then used as the gold standard against which we compared the results of the manual chart abstraction.

5.3.4 Assessing Discordance

We manually reviewed the EMR of all discordant records in the comparisons between the text mining tool and manual abstractions as well as the EMR MI diagnoses and CIHI MI diagnoses.

5.3.5 MI Quality Indicators

The following selected measures of quality of post-AMI care were evaluated in the year prior to March 31, 2007 for MI patients identified using manual EMR abstraction (i.e., the gold standard): the proportion prescribed (i) acetylsalicylic acid (ASA), (ii) angiotensin converting enzymes (ACE) inhibitors, and (iii) beta blockers; the proportion with a laboratory measurement of fasting or random blood sugar, and a total cholesterol measured.; and the proportion who had their blood pressure measured. The latter could only be assessed in the EMR data as this information is not available in administrative databases. We evaluated the prescription-based measures in the EMRs for all adult patients with an MI, as well as restricted the analysis to patients’ ≥ 65 years of age since prescription drug claims are universally available in the ODB only for this age group.

5.3.6 Analysis

We used a t-test for continuous variables, and a \( \chi^2 \) test for categorical variables, to compare (1) baseline physician and patient characteristics between the EMR database and the 2001 Ontario standard population, (2) sensitivity and specificity, and (3) differences in the
proportion of patients documented as “having an MI quality indicator in the EMR and administrative data. We also calculated the kappa statistic in order to evaluate inter-abstractor and intra-abstractor reliability.

The sensitivity of the text mining tool was calculated as the proportion of MI patients identified by the text mining tool who had an MI according to the “gold standard” of manual EMR abstraction (Table 5.1). Specificity was calculated in the same manner except that it was based on individuals without an MI. Ninety-five percent confidence intervals (95% CI) for these proportions were calculated using the binomial approximation method in SAS version 9.1. (SAS Institute, Cary, North Carolina). PPV was defined as the proportion of MI patients identified by the text mining tool that were confirmed by the gold standard method. NPV was defined similarly for patients that did not have an MI according to the text mining tool. All analyses were conducted using SAS version 9.1.

5.3.7 Ethical Consideration

This project received approval from both the Queen’s University Health Sciences and the Sunnybrook Health Sciences Research Ethics Boards prior to commencing (Appendix A).

5.4 Results

Our convenience sample of 17 physicians had a similar gender and urban rural split but had significantly more physicians in group practice compared with all Ontario family physicians (Table 5.2) The average years in practice was 20.5 years and the average length of time on the EMR was 7.4 years.

The age and gender distribution of the patients in our EMR database were similar to the Ontario 2001 standard population (using census data) with a slight overrepresentation of women and children. (Figure 5.4 a and b) The average age of our active adult cohort was 49.0 years.
Table 5.1 Description of Measures of Validity and Reliability

<table>
<thead>
<tr>
<th>Comparator</th>
<th>Gold Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI (+)</td>
<td>MI (-)</td>
</tr>
<tr>
<td></td>
<td>MI (+)</td>
</tr>
<tr>
<td>MI (+)</td>
<td>TP (true positives)</td>
</tr>
<tr>
<td>MI (-)</td>
<td>FN (false negatives)</td>
</tr>
</tbody>
</table>

Sensitivity = TP / (TP + FN)
Specificity = TN / (TN + FP)
Positive Predictive Value = TP / (TP + FP)
Negative Predictive Value = TN / (TN + FN)
Table 5.2 Physician Characteristics

<table>
<thead>
<tr>
<th>Physician Characteristics</th>
<th>Participating Physicians</th>
<th>All Ontario Family Physicians 2007†</th>
<th>P-value ‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>17</td>
<td>10,855</td>
<td>N/A</td>
</tr>
<tr>
<td>Mean Years since Graduation</td>
<td>20.5 yrs SD = 10.2</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Mean time on PS EMR®</td>
<td>7.4 yrs SD = 7.3</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Proportion Male</td>
<td>70.6 %</td>
<td>63.8 %</td>
<td>0.56</td>
</tr>
<tr>
<td>Proportion Urban/Suburban</td>
<td>58.3 %</td>
<td>58.4 %</td>
<td>0.97</td>
</tr>
<tr>
<td>Proportion in Group Practice</td>
<td>91.7 %</td>
<td>46.3 %</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

† Data obtained from The National Physicians Survey 2007 Results

‡ P-value is for two-sided chi square test with $\alpha = 0.05$

N/A = Not available
Figure 5.4 EMR Population Compared to 2001 Standard Ontario Population
which was slightly higher than the average age of adults ≥ 20 years in Ontario. As well, our active adult cohort had a slight under representation of males compared to the Ontario 2001 standard population (Table 5.3).

5.4.1 Intra-Observer and Inter-observer Reliability

The measurement of intra-observer reliability resulted in kappa values exceeding 0.80 indicating very good agreement. When measuring inter-observer reliability the kappa values exceeded 0.85 for all comparisons between the three abstractors also indicating very good agreement.

5.4.2 Text-Miner vs. Abstractor (Validation Set, N=977)

The trained manual abstractors identified 231 definite MIs of the 977 patients in the validation data set (Table 5.4). The text miner identified 255 definite MIs including 31 false positives, resulting in a sensitivity of 97.0% (CI 94.8–99.2%) and specificity of 95.8% (CI 94.3–97.3%) compared with EMR abstracted MIs. The text miner also had a sensitivity and specificity greater than 95.0% across the training and validation set. There was no statistically significant difference in validity and reliability of the tool between the training and validation set with respect to specificity and NPV. The tool had slightly improved sensitivity and specificity in the training set.

The tool identified 51/53 (96.2%) of the MIs in the charts of doctors it had been developed on versus 173/178 (97.2%) of the MIs of “new” physicians. There was no statistically significant difference between the two groups with the exception of the PPV which was 79.7% (CI 69.8–92.5%) for the “validation seen” cohort and 90.6% (CI 86.4–95.2%) for the “validation unseen” patients (p-value=0.02).
Table 5.3 Patient Characteristics for Active Patients

<table>
<thead>
<tr>
<th>Patient Characteristics</th>
<th>EMR 2007</th>
<th>Ontario 2001</th>
<th>P-value of $X^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number</td>
<td>40,228</td>
<td>15,035,295</td>
<td>N/A</td>
</tr>
<tr>
<td>Total Active Adults</td>
<td>19,376</td>
<td>8,783,527</td>
<td>N/A</td>
</tr>
<tr>
<td>Mean time on EMR*</td>
<td>4.9 yrs</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Mean number of visits *</td>
<td>4.0/person year</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Mean Age*</td>
<td>49.0 yrs SD=17.2</td>
<td>47.4 yrs SD=17.3</td>
<td>&lt;0.0001**</td>
</tr>
<tr>
<td>Proportion Male*</td>
<td>46.3 %</td>
<td>49.41 %</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

* For Active Adult Cohort

** p-value of t-test
Table 5.4 Text Miner vs. Abstractor*

<table>
<thead>
<tr>
<th>DATA SOURCE</th>
<th>Total</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation</td>
<td>977</td>
<td>224</td>
<td>31</td>
<td>7</td>
<td>715</td>
<td>97.0% (CI 94.8 - 99.2%)</td>
<td>95.8% (CI 94.3 - 97.3%)</td>
<td>87.8% (CI 83.8 - 92.5%)</td>
<td>99.0% (CI 98.3 - 99.2%)</td>
</tr>
<tr>
<td>Training</td>
<td>316</td>
<td>72</td>
<td>6</td>
<td>7</td>
<td>231</td>
<td>98.6% (CI 96.0 - 101.3%)</td>
<td>97.4% (CI 95.4 - 99.5%)</td>
<td>92.3% (CI 86.4 - 98.7%)</td>
<td>99.6% (CI 98.7 - 100.2%)</td>
</tr>
<tr>
<td>(1) Validation</td>
<td>317</td>
<td>51</td>
<td>13</td>
<td>2</td>
<td>251</td>
<td>96.2% (CI 91.1 - 101.4%)</td>
<td>95.0% (CI 92.4 - 97.8%)</td>
<td>79.7% (CI 69.8 - 92.5%)</td>
<td>99.2% (CI 98.1 - 99.8%)</td>
</tr>
<tr>
<td>seen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) Validation</td>
<td>660</td>
<td>173</td>
<td>18</td>
<td>5</td>
<td>464</td>
<td>97.2% (CI 94.8 - 99.6%)</td>
<td>96.2% (CI 94.4 - 98.0%)</td>
<td>90.6% (CI 86.4 - 95.2%)</td>
<td>98.9% (CI 98.0 - 99.3%)</td>
</tr>
<tr>
<td>not seen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall Abstraction</td>
<td>1293</td>
<td>296</td>
<td>37</td>
<td>14</td>
<td>946</td>
<td>97.4% (CI 95.6 - 99.2%)</td>
<td>96.2% (CI 94.9 - 97.4%)</td>
<td>88.9% (CI 85.5 - 92.7%)</td>
<td>99.1% (CI 98.6 - 99.3%)</td>
</tr>
</tbody>
</table>

* Gold Standard

Legend: TP= True Positive, FP= False Positive, FN= False Negative, TN= True Negative,
PPV= Positive Predictive Value, NPV= Negative Predictive values
5.4.2.1 Discordance

There were 51 discordant records when comparing the results of the text miner and the abstraction set. Of these discordant records, 9 were due to abstraction errors and 42 were due to text-mining errors. The reasons for the text-mining errors included misreading of the tool in interpreting the occurrence of the letters MI to mean myocardial infarction when they stood for something else 12/42 (28.6%), errors when the MI of a family member was mentioned 7/42 (16.7%) errors when MI was used in the context of a risk discussion of a procedure or therapy 4/42 (9.5%), and instances where the tool was not able to pick up and interpret the language correctly 19/42 (45.2%).

5.4.3 Manual Abstraction vs. CIHI (Abstraction Set, N=1293)

The abstractors identified 310 MIs in the abstraction set while the CIHI-DAD identified 234 MIs (Table 5.5). When evaluating the EMR against the administrative data as the gold standard it had a sensitivity of 94.9% (CI 92.0-97.7%), specificity of 91.7% (CI 90.0-93.5%), PPV of 71.6% (CI 66.6-79.2%), and NPV of 98.8% (CI 98.1-99.0%).

We then stratified the results according to patients with greater than four years of data and those with less than four years. There was no statistically significant difference between the sensitivities (p=0.68) NPV (p=0.82) and PPV (p=0.08) of the two groups. However, there was a statistically significant difference between the two specificities (p<0.001), with a result of 88.8% for those with ≥ four years of data and 95.1% for those with less than four years of data.

5.4.3.1 Discordance

There were 12 patients identified as having had an MI in CIHI but not identified by the abstractors. Of these 12 patients, 4/12 (33.3%) were abstractor errors, whereby there was in fact information relating to MI found in the patient EMR. Thus, in essence only eight patients had an
Table 5.5 EMR (Abstractor) vs. CIHI*

<table>
<thead>
<tr>
<th>COHORT</th>
<th>Total</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive Predictive Value</th>
<th>Negative Predictive Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Abstraction</td>
<td>1293</td>
<td>222</td>
<td>88</td>
<td>12</td>
<td>971</td>
<td>94.9% (CI 92.0 - 97.7%)</td>
<td>91.7% (CI 90.0 - 93.5%)</td>
<td>71.6% (CI 66.6 - 79.2%)</td>
<td>98.8% (CI 98.1 - 99.0%)</td>
</tr>
<tr>
<td>Patients ≥ 4 years of data</td>
<td>641</td>
<td>127</td>
<td>57</td>
<td>6</td>
<td>452</td>
<td>96.2% (CI 92.9 - 99.5%)</td>
<td>88.8% (CI 86.1 - 91.9%)</td>
<td>69.0% (CI 62.3 - 79.8%)</td>
<td>98.9% (CI 97.9 - 99.2%)</td>
</tr>
<tr>
<td>Patients &lt; 4 years of data</td>
<td>652</td>
<td>97</td>
<td>27</td>
<td>6</td>
<td>523</td>
<td>95.1% (CI 90.9 - 99.3%)</td>
<td>95.1% (CI 97.0 - 99.3%)</td>
<td>78.2% (CI 70.9 - 87.9)</td>
<td>99.0% (CI 98.2 - 99.3%)</td>
</tr>
</tbody>
</table>

* Gold Standard
MI identified in CIHI but not in the EMR record. Four of these patients had multiple visits and specialist consultation letters but no mention of MI, three had multiple visits but no hospital discharge summaries or specialist consultation letters and one only had very brief progress notes.

There were 88 patients classified as having an MI in the EMR with no evidence of an MI in CIHI. Two of these 88 patients, originally diagnosed as having had an MI, had subsequent documentation in their EMR that ruled out the MI (thus they were not truly false negatives missed by CIHI). Of the remaining 86 patients, 30/86 (34.9%) were clearly admitted to hospital with either an acute MI or past history of MI but did not appear in CIHI as having had an MI, and 13/86 (15.1%) had clear documentation of having had a silent MI, 13/86 (15.1%) had some elements that suggested that they had a silent MI (i.e., incidental ECG finding alone). The remaining 30 patients, had some mention of a past history of an MI, but from the information contained with the EMR distinguishing MIs that presented to hospital or silent MIs was not possible.

5.4.4 MI Quality Indicators: EMR vs. Administrative Data

Of the MI patients identified through manual abstraction, 75.3%, 69.7% and 71.4% received an ASA, beta blocker and ACE inhibitor prescription respectively (Table 5.6). When comparing the proportion of MI patients (as identified by manual abstraction) age 65 years and older who were on ASA, beta blockers and ACE inhibitors using EMR data versus the ODB, there was no statistically significant difference in the proportions captured by each data source with the exception of ASA (p<0.001). The EMR recorded 70.9 % of MI patients were on ASA while the ODB only recorded 13.9%. When evaluating the proportion of all active adult MI patients that had total cholesterol, blood sugar, or blood pressure measured when using EMR data versus OHIP, there was no statistically significant difference between the two data sources (Table 5.7).
Table 5.6 Selected MI Quality Indicators Using Prescriptions

<table>
<thead>
<tr>
<th>Prescriptions</th>
<th>Using EMR Data: All Active Adults</th>
<th>Using EMR Data: Age ≥ 65</th>
<th>Using ODB: Age ≥ 65</th>
<th>P-value of $\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASA</td>
<td>75.3%</td>
<td>70.9%</td>
<td>13.9%</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Beta Blocker</td>
<td>69.7%</td>
<td>71.5%</td>
<td>70.9%</td>
<td>0.9</td>
</tr>
<tr>
<td>ACE Inhibitor</td>
<td>71.4%</td>
<td>67.9%</td>
<td>67.9%</td>
<td>1.0</td>
</tr>
</tbody>
</table>

(N=310 for All Active Adult MIs patient, N=165 for MI patients 65 and older)
Table 5.7 Selected MI Quality Indicators Using Laboratory Values and Blood Pressure Measurements

<table>
<thead>
<tr>
<th>Tests</th>
<th>Using EMR Data: Age ≥ 20 yrs</th>
<th>Using OHIP: Age ≥ 20 yrs</th>
<th>P-value of $\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood Sugar</td>
<td>73.7%</td>
<td>78.6%</td>
<td>0.15</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>63.8%</td>
<td>70.4%</td>
<td>0.08</td>
</tr>
<tr>
<td>Blood Pressure</td>
<td>84.9%</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

(N=310 for All Active Adult MIs patient, N=165 for MI patients 65 and older)
5.5 Discussion

With a sensitivity and specificity exceeding 95% the text mining tool exceeded the performance of similar software programs\textsuperscript{23,24} for identifying patients with disease conditions such as diabetes. The improved results may be due to the way MI is recorded in the EMR. Some conditions have a greater number of synonyms or more ambiguity in how they are documented.\textsuperscript{25} In analyzing the reasons for discordance between the abstractors and the text miner we noted that the majority of the text miner errors were a result of the text mining tools inability to discern ambiguous language such as “best friend had a heart attack infarct” or the description of infarctions of another type. This type of error can be minimized by providing the text mining tool a greater number of clinical examples during the training phase. However, errors cannot be completely eliminated since it is impossible to provide a standardized computer program with infinite ways one can convey a complex concept.

The information used by the abstractors to identify myocardial infarction came from many different sources in the EMR. The most common sources were consultation letters from the cardiologist, ECG tests and the progress notes. If we had only used information in the EMR that was generated by the family physician (i.e., the progress notes, etc.) and excluded diagnostic tests and consultation letters, we would have identified only 154 (49.7%) of the 310 MIs. This demonstrates the importance of consultation letters and diagnostic tests which are external documents that are either scanned into the EMR in a picture like format or are converted to searchable free text using optical character recognition.

The fact that there was no clinically and statistically significant differences for the sensitivity and specificity of the tool used with the training EMRs and the “naïve” EMRs indicates that the text mining tool adapted well to new language patterns and variations in the documentation of clinical information that often existed between our sample of physicians.
We then evaluated the validity and reliability of MI diagnoses in the EMR by linking each patient to the CIHI hospitalization discharge database; the current standard for the identification of MI patients. The EMR had a sensitivity and specificity exceeding 90% suggesting that the majority of patients hospitalized for MI have the information recorded in their family physician’s record. This important finding would suggest that there is a good flow of information regarding MI from the acute care to the ambulatory setting.

When evaluating discordance between the EMR and CIHI there were 88 patients that were identified as having had an MI in the EMR that were not captured by the CIHI database. The fact that 15% of these MI cases had been diagnosed solely in the ambulatory setting supports our hypothesis that the EMR can serve as an additional data source for the identification of MI patients that do not present to hospital.

The comparison of EMR and administrative databases for the evaluation of selected MI quality indicators suggests that EMRs are at least as good as administrative data for the indicators tested, and offer the advantage of capturing information that is not contained in administrative databases including blood pressure readings and medications such as low-dose ASA that are not covered by ODB. These findings further validate the utility of EMRs as an additional data source.

This study has several limitations. The first is that we can not conclude that our results are generalizable as we used a convenience sample of physicians. We cannot exclude the possibility that physicians who agreed to participate may be quite different from those who did not. Indeed, we provided evidence that our study physicians were different with regards to a number of characteristics that could influence the extent and pattern of medical record documentation. However, the impact of this was addressed, at least in part, by the descriptive analysis of key physician and patient characteristics as mentioned previously.

Another limitation of our findings arises due to the artificial inflation of MI prevalence in our abstraction data set (N=1293) at a level of 24 %. The true prevalence of MI in the community
has been estimated to be between 1-2% as mentioned previously. This would decrease the PPV of our tool in settings where prevalence is substantially decreased.

In addition, the date of diagnosis is not consistently recorded in the EMR. For example, if the MI occurred prior to the time frame covered by the physician’s EMR, the physician may simply note that the patient had a previous MI. Consequently, the comparison of the EMR’s recording of MIs with that of the CIHI database may be limited by the inability to perfectly match MIs on their date of occurrence. The latter could lead to the inclusion of events recorded in the EMR that are beyond the time frame captured by the CIHI database or vice versa and result in misclassification of the outcome status. This potential source of information bias was considered during the detailed review of the discordant records at which time we identify some of these cases.

5.6 Conclusion

Automated data extraction tools such as the one evaluated in this study can be a valuable tool for accurately and efficiently extracting diagnostic information from the large volume of clinical information being generated and stored in EMRs. The data from EMRs contain more detailed clinical information and appears to be more comprehensive than health information obtained in administrative data. In future, the EMR will likely constitute an important additional database to obtain more complete case ascertainment for conditions such as MI and the measurement of quality of care.
5.7 References


(2) Mamlin BW, Heinze DT, McDonald CJ. Automated extraction and normalization of findings from cancer-related free-text radiology reports. AMIA Annu Symp Proc 2003;420-424.


(18) Austin PC, Daly PA, Tu JV. A multicenter study of the coding accuracy of hospital discharge administrative data for patients admitted to cardiac care units in Ontario. Am Heart J 2002; 144(2):290-296.
6.1 Main Findings

6.1.1 General Summary

In this validation study we found that our text mining tool was able to identify myocardial infarction (MI) diagnoses in EMRs at a level of accuracy comparable to manual abstraction. When comparing the MI diagnoses status of patients in our abstraction dataset with those obtained from the CIHI-DAD as the gold standard, the sensitivity, specificity and negative predictive value (NPV) all exceeded 90%, thus indicating good agreement between the two data sources. The positive predictive value (PPV) on the other hand, was only 71.6%. This would normally indicate that the EMR contained many false positives. However, after a manual review all of our discordant records, we concluded that the low PPV was the result of comparing the EMR to an imperfect gold standard. For example, the CIHI-DAD data does not capture silent MI diagnoses which are potentially only recorded in the ambulatory setting. Since this type of MI may be captured by the EMR, the latter may spuriously appear to be less accurate. In addition, there were cases where the EMR clearly documented a hospitalization for AMI and yet codes such as CAD were instead recorded as the primary reason for hospitalization in the CIHI-DAD. These findings suggests that the EMR is an important additional data source for complete case ascertainment of MI diagnoses.

When comparing the proportion of MI patients $\geq$ 65 years in our abstraction cohort who were on ASA, ACE inhibitors and beta blockers according to the EMR and ODB there was no difference between the two data sources with the exception of the EMR identifying significantly more users of ASA than the ODB database (70.9% vs. 13.9% respectively). However, this was an expected difference since most patients purchase low-dose ASA over the counter because it is not
covered under the ODB plan. In addition, the primary care physician is likely to document this information even when patient initiated. The same comparison was repeated for the proportion of MI patients ≥ 20 years who had a blood sugar, cholesterol or blood pressure measurement. Once again, there were no statistically significant differences between the two data sources. However, the proportions captured by the EMR were slightly lower, likely because OHIP also captures tests performed outside of the family practice setting such as a specialist’s office. In addition, the measurement of blood pressure was only captured the EMR as there is no billing code for this in OHIP.

6.1.2 Secondary Analyses

The text mining tool maintained validity and reliability on the medical records of physicians that were selected to be in the training set and those that had not. This suggests that it is able to adapt to the variations in the linguistic patterns that exist between physicians.

6.2 Consistency of the Evidence

Our results are consistent with the findings of other recent studies investigating text mining in the free text of medical records, with slightly improved sensitivity and specificity. The variation is likely a result of the differences in the way MI is documented in the EMR when compared with other disease conditions. The variation in the accuracy of identifying disease diagnoses between various conditions has been evaluated.1

6.3 Strengths and Limitations

A major advantage of our text mining tool was that it was developed using a relatively modest training set (N=316) which required less resource expenditure for the manual review of charts.2-4 A major strength in our approach to developing the text mining tool was the use of support vector machine technology. This method of text mining is founded upon the principle of providing the software program with a sample of EMRs that have both diseased and non-diseased
individuals. This means that you do not have to provide the tool with an exhaustive list of all the synonyms for a disease, as it can make predictions based on the types of words sentences and language structure in the EMRs of diseased individuals. However, a disadvantage to this approach is that you must insure that your training sample has adequate variation in the manners a disease is defined, which can increase the size of your training sample. In addition, our results are unlikely to be due to chance given our significant findings and 95% confidence intervals.

Because of our use of a convenience sample to identify physicians willing to contribute their anonymized data, we cannot exclude the possibility that physicians in our study differ from other physician users of EMRs and other primary care physicians in general. It is likely that our participants are early adopters of technology since their mean time on the EMR system was long (7.4 years). If physicians in this study differ systematically from other physicians with respect to the accuracy, completeness and comprehensive of diagnoses information documented in their EMRs, then our comparisons with administrative databases may overestimate the validity and reliability of this data source. Although we were able to compare study physicians with Ontario primary care physicians with regards to some important baseline characteristics, we were not able to obtain information on non-participating EMR users to assess this potential source of selection bias. This may also have affected our assessment of the text mining tool’s generalizability to different language and documentation patterns, as participating physicians are likely to be more similar with regards to documentation habits.

Our saturation of the abstraction data set with MI cases artificially inflated the prevalence of MI patients in this sample. This would affect the positive predictive value of the tool if it were run on a population of patients with a lower prevalence of disease. However, in an older and higher risk population the results may be comparable to those reported in this study.

By restricting our study to active patients only, we may have excluded individuals who were systematically different from those in our sample. Patients with more medical visits are more likely to be female and more likely to be sicker than those who do not. Our cohort
of 19,376 active adults had a sex distribution similar to that of the complete cohort (N = 40,228), with only a slight overrepresentation of women. On the other hand, our active adult cohort was slightly older than the Ontario 2001 population which was the opposite of our complete cohort (N = 40,228). This could affect the generalizability of our results if the data mining tool performed differently on the EMRs of patients with fewer medical visits and if the information was documented differently.

Another limitation of our study is the misclassification of a person’s MI status in both the EMR and CIHI-DAD. In the EMR, the date of MI occurrence is not consistently documented particularly when the MI occurred prior to the time period during which the patient was seen by the physician. In these instances there may only be a note of the patient’s self-reported MI status but not the event date. If the patient has no record of the MI in the CIHI-DAD it could be for one of the following reasons (1) The MI occurred prior to the time frame covered by the CIHI-DAD (2) The MI occurred out of province (3) The person was misclassified in CIHI (4) Abstractor error. For this reason, it is possible that there was misclassification of the outcomes status. However, the investigation of the discordant records demonstrated that this occurred in a small minority of cases.

6.4 Clinical Practice Implications

In the future physicians may be able to use this text mining technology to quickly identify patient cohorts of diseased individuals within their own practices and verify their level of adherence to appropriate treatment protocols. This could accelerate the uptake of new evidence into clinical practice and ultimately, improve health outcomes.

6.5 Future Research Directions

Our study needs to be replicated in a larger and more representative sample of EMR users, particularly with regards to possible variations in use of language and documentation
patterns. Furthermore, the results of this study are from a pilot version of the text mining tool and continued work is in progress to improve the results.

We plan to expand this work to encompass the identification of other disease conditions. We also expect to use the automated text mining tool developed for the current validation study for measuring the quality of care in Ontario primary care facilities.

6.6 Conclusions

While there are many barriers to accessing and utilizing EMR data, it is important that we continue to invest in the development of text mining methods to accurately and efficiently access this rich clinical data source in light of the tremendous investment that is being made to implement this technology across all sectors of the Canadian health care system. It is important that text mining methods be developed in conjunction with the large scale national initiatives overseeing the implementation of EMRs, or we could be left with a fragmented system that does not lend itself well to research and evaluation. Ensuring these abilities would maximize the impact of the EMR.
6.7 References


Appendix A

Ethics Approvals

QUEEN'S UNIVERSITY HEALTH SCIENCES & AFFILIATED TEACHING HOSPITALS RESEARCH ETHICS BOARD

May 8, 2008

Dr. Linda Levesque
Department of Community Health and Epidemiology
KFL&A Public Health
221 Desмонд Avenue
Kingston, ON K7M 4V5

Dear Dr. Levesque,

Study Title: Identifying Myocardial Infarction Patients Using Automated Data Extraction in Family Practice Electronic Medical Records
Co-Investigators: Dr. Karen Tu

I am writing to acknowledge receipt of your recent ethics submission. We have examined the protocol for your project (as stated above) and consider it to be ethically acceptable. This approval is valid for one year from the date of the Chair’s signature below. This approval will be reported to the Research Ethics Board. Please attend carefully to the following list of ethics requirements you must fulfill over the course of your study:

- Reporting of Amendments: If there are any changes to your study (e.g., consent, protocol, study procedures, etc.), you must submit an amendment to the Research Ethics Board for approval (see http://www.queensu.ca/vps/reb.htm).

- Reporting of Serious Adverse Events: Any unexpected serious adverse event occurring locally must be reported within 2 working days or earlier if required by the study sponsor. All other serious adverse events must be reported within 15 days after becoming aware of the information.

- Reporting of Complaints: Any complaint made by participants or persons acting on behalf of participants must be reported to the Research Ethics Board within 7 days of becoming aware of the complaint. Note: All documents supplied to participants must have the contact information for the Research Ethics Board.

- Annual Renewal: Prior to the expiration of your approval (which is one year from the date of the Chair’s signature below), you will be reminded to submit your renewal form along with any new changes or amendments you wish to make to your study. If there have been no major changes to your protocol, your approval may be renewed for another year.

Yours sincerely,

[Signature]
Chair, Research Ethics Board

[Date: May 8, 2008]

Study Code: EPID-260-08

Investigators please note that if your trial is registered by the sponsor, you must take responsibility to ensure that the registration information is accurate and complete.
RENEWED ETHICS APPROVAL

RESEARCH ETHICS BOARD
SUNNYBROOK
HEALTH SCIENCES CENTRE
C819, 2075 Bayview Avenue
Toronto, Ontario
M4N 3M5

Must be completed annually for all ongoing studies
Please return to room C819

Project Title: Canadian Cardiovascular Outcomes Research Team (CCORT) Team Grant in Cardiovascular-Outcomes Research

Project Identification Number: 271-2006

Original Approval Date: July 6, 2006

Principal Investigator: Dr. Jack V. Tu

Full Address Including Room Number: ICES G1 06 2075 Bayview Avenue, Toronto, Ontario M4N 3M5 Room G2 50

Full Board Review Required: Yes [ ] No [ ]

If industry sponsored, please note there is a $500 re-approval fee. Invoicing information including contact name and full mailing address is required for each renewal.
Project D: Access to primary and secondary prevention of heart disease in the community
Construction of the patient cohort has been completed. Methods of text mining are being pilot tested, and a data validation process is underway. Some preliminary analysis is in process.

Amendments to the study (must be submitted for approval): None

Changes in scientific knowledge that could impact on the study and action taken: None

Unexpected or adverse events and action taken: N/A

Protocol violations and actions taken: N/A

Expected date of completion: March 31, 2011
Sunnybrook Health Sciences Centre
Research Ethics Board
Renewed Ethics Approval

Please sign below:

My signature certifies the following information is correct and I will not use any procedures, which have not been approved by the Board.

Signature of Principal Investigator  

Date  

The above study is ethically acceptable and has received renewed ethics approval. This study may continue at Sunnybrook Health Sciences Centre.

Philip C. Hébert MD, PhD. FCAPC  
Chair, Research Ethics Board  

Date of Review  

Date of Full Board Review (if required):

Chair  

69