DESIGN AND EVALUATION OF AN ASSOCIATIVE CLASSIFICATION FRAMEWORK TO IDENTIFY DISEASE COHORTS IN THE ELECTRONIC HEALTH RECORD

by

Susan Rea Welch

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

May 2011

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Susan Rea Welch**

has been approved by the following supervisory committee members:

Stanley M. Huff , Chair **March 16, 2011**
Date Approved

Lewis J. Frey , Member **March 16, 2011**
Date Approved

Peter J. Haug , Member **March 16, 2011**
Date Approved

Scott P. Narus , Member **March 16, 2011**
Date Approved

Lucy A. Savitz , Member **March 16, 2011**
Date Approved

and by **Joyce A. Mitchell** , Chair of
the Department of **Biomedical Informatics**

and by Charles A. Wight, Dean of The Graduate School.

ABSTRACT

With the growing national dissemination of the electronic health record (EHR), there are expectations that the public will benefit from biomedical research and discovery enabled by electronic health data. Clinical data are needed for many diseases and conditions to meet the demands of rapidly advancing genomic and proteomic research. Many biomedical research advancements require rapid access to clinical data as well as broad population coverage. A fundamental issue in the secondary use of clinical data for scientific research is the identification of study cohorts of individuals with a disease or medical condition of interest. The problem addressed in this work is the need for generalized, efficient methods to identify cohorts in the EHR for use in biomedical research.

To approach this problem, an associative classification framework was designed with the goal of accurate and rapid identification of cases for biomedical research:

(1) a set of exemplars for a given medical condition are presented to the framework,

(2) a predictive rule set comprised of EHR attributes is generated by the framework, and

(3) the rule set is applied to the EHR to identify additional patients that may have the specified condition.

Based on this functionality, the approach was termed the ‘cohort amplification’ framework.

The development and evaluation of the cohort amplification framework are the subject of this dissertation. An overview of the framework design is presented. Improvements to some standard associative classification methods are described and validated. A qualitative evaluation of predictive rules to identify diabetes cases and a study of the accuracy of identification of asthma cases in the EHR using framework-generated prediction rules are reported. The framework demonstrated accurate and reliable rules to identify diabetes and asthma cases in the EHR and contributed to methods for identification of biomedical research cohorts.

“Capture everything, we’ll sort it out later.”

T. Allan Pryor, Ph.D. (1937 - 2009)

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Although domain experts are vital to the oversight of any disease case identification algorithm, the translation of the clinical and health care encounter characteristics of a phenotype to EHR data specifications can be improved. The cohort amplification framework may leverage the expert's time by providing information on the EHR data which best distinguishes disease exemplars.

Problem Statement

The problem addressed in this work is the need for generalized and computable methods to identify cohorts in the EHR for biomedical research. With the growing national dissemination of the electronic health record (EHR), there are expectations for enhanced secondary use of the EHR for purposes of biomedical research.[1, 2] Such functionality was explicitly defined as an objective in the developing national standards for meaningful use of the EHR.[3, 4] Clinical data are needed for many different diseases and conditions to meet the demands of rapidly advancing genomic and proteomic research.[5, 6] Other biomedical research to improve the general health status requires expeditious access to clinical data as well as general coverage of the population.[7, 8] To use the electronic health record data for research purposes, the first step is often the identification of study cohorts of individuals with a disease or medical condition of interest. Ideally, generalized criteria may be established for identification of cohorts in the EMR. This enables researchers to design studies that might be applied across the population for broad attribution of results, pooling of subjects and equitable access. The efficiency of biomedical research is improved when cohort identification logic can be shared and can be applied directly to the EHR.

BIOINFOMED, a study group funded by the European Commission (EC) addressed issues and challenges in correlating essential genotype information with expressed phenotype information.[9,10] They reported that genomic and proteomic data must be integrated with electronic health record data, which can be used as expressed phenotype information. Further, they reported that to obtain new knowledge, the phenotypes, genotypes and proteotypes of many patients from all over the world must be combined. To make this possible, descriptions of the phenotypes must be standardized. They proposed structured clinician entry or computerized interpretation of EHR content, including free text, or a combination of methods.

Current Solutions

Validated automated logic to identify disease-based cohorts in the medical record in the U.S. commonly uses International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes. ICD morbidity codes have been recorded in hospital records in the U.S. since 1944. They were originally collected for the systematic analysis of causes of morbidity and mortality. This followed a long tradition of international disease classification efforts begun before 1785, now formalized under the World Health Organization as the International Classification of Diseases (ICD).[11] ICD is intended as a classification for clinical, general epidemiological and many health management purposes, while explicitly not intended for financial applications.[12] In 1983, ICD-9-CM codes began to be used to determine reimbursement from healthcare payors in the inpatient setting in the United States. Consequently, ICD codes in the U.S. were expanded and detail added in the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM).[13] Subsequently, ICD-9-CM codes were used in the ambulatory setting to qualify the CPT procedure codes[14] submitted for reimbursement.

ICD-9-CM based algorithms to identify disease-based cohorts have variable accuracy rates.[15-21] Federal Health Insurance Portability and Accountability Act (HIPAA) guidelines for diagnostic coding have become complex, change several times per year, and require training for correct use.[22] In addition, the ICD-9-CM codes in the EHR are typically bound to billing processes. This leads to consistent recommendations that the use of ICD-9-CM codes to identify cohorts for biomedical research should be validated.[17, 23-25]

There is a general expectation that automated algorithms to identify disease-based cohorts can be improved by using additional EHR data rather than just ICD-9-CM codes alone. Logic to identify cohorts from clinical and administrative data in the EHR are usually defined by domain experts and analysts based upon specification and analysis of attributes in the EHR and/or billing claims data for particular diseases.[26-29] Such processes are often time-consuming for the experts. Such a process to identify phenotype cohorts from the EHR was described by Starren:[30]

Define Phenotype → Translate Definition to Data → Analyze Data →
→ Identify Subjects → Validate Algorithm → [repeat]

The expert's role shifts to refinement of the machine-generated knowledge instead of specifying and analyzing data definitions:

Define Phenotype Exemplars

→ ***Develop Rules Predictive of Exemplars of the Phenotype from EHR Data***

→ ***Rule refinement***

→ *Identify Subjects* → *Validate Algorithm* → *[repeat]*

Natural language processing (NLP) of health care providers' free-text documentation, a rich source of information in the EHR, is an active and promising area of research for purposes of disease case identification.[31-35] The cohort amplification framework is complementary to NLP methods and processes in providing domain knowledge as well as opportunities to combine coded and free-text data.

Cases might also be identified if diseases or conditions of interest were documented by clinicians in the coded Problem List structure of the standard EHR.[36-39] However, at this time, notation in Problem Lists is not commonly integrated into the routine data/work flow of clinical practices.[40, 41]

The Cohort Amplification Framework

A novel approach to identify cohorts in the EHR for biomedical research purposes was conceptualized, developed and evaluated. Design of the cohort amplification framework was motivated by the need to find phenotype cohorts in the EHR for genomics research at the University of Utah. The use case required the identification of many disease-related phenotypes of interest to researchers to support high-throughput familial clustering processes.[42, 43] The design accommodated these needs with a set of software components and processes that did not require reprogramming for a new disease or condition and required minimal domain expert input to generate classification rules for the disease. This development work resulted in a collection of original Java components and Structured Query Language (SQL) database procedures.

The general use case for the design was:

(1) A set of exemplars for a given medical condition are identified by a set of known rules, such as ICD-9-CM based rules.

(2) A clinical profile (predictive rule set) is generated from the exemplars' EHR data using the framework.

(3) The predictive rule set is applied to the entire patient population in the EHR to identify additional patients that may have the specified condition.

Exemplars

The FW takes two exemplar cohorts – referred to as cases and controls - as input. The framework generically exposes patterns in the EHR data that distinguish the exemplars with the condition of interest from exemplars without the condition. Although the scope of cohort amplification supported by the framework includes any medical condition for which health care services are typically sought, the term 'disease' is usually used throughout this dissertation. The condition of interest may be a syndrome or a subtype or subgroup of a disease. Exemplars of a disease may be specified by a set of known rules, such as ICD-9-CM based rules or could be a researcher's current list of known cases. Those without the condition are referred to as 'controls'. The control exemplars may be negative for the condition of interest, or they may represent any contrasting cohort such as those with less severe disease status, if two subtypes of one disease are compared. The size and the representativeness of the exemplars will affect the quality of the prediction rules that are generated. Given the large number of data elements used in the framework, both disease and control exemplar samples of at least 1,000 are recommended for generation of reliable rules.

EHR Data

The framework is distinguished by core candidate data attributes based on nationally required EHR observation categories. The core data attributes were based on requirements specified for the certification of ambulatory medical records by the Certification Commission for Health Information Technology.[44] This was the certification requirement for an EHR according to Centers for Medicare & Medicaid Services (CMS) when development of the framework commenced. Subsequently, the Health Information Technology for Economic and Clinical Health (HITECH) Act, which was part of the American Recovery and Reinvestment Act of 2009, authorized oversight for the national certification standards for EHRs.[4] The new standards designed for 'meaningful use' of the EHR also include the framework's core data attributes[3, 45]. The core attributes are typically populated in an EHR as a by-product of health care delivery and documentation processes.

There is no technical limitation to adding disease or site-specific content, but the focus of this research is standardized content for generalized application. The list of candidate data attributes are easily modified in one component using SQL, by design. Data observation categories used in the framework for this research are:

- Diagnosis and procedure codes (ICD-9-CM codes)
- Provider and ambulatory clinic procedure codes (CPT codes) [14]

- Provider specialty (local codes)
- Lab observations (CPT codes)
- Lab observations with results coded as ‘Abnormal’
- Imaging procedures (CPT codes)
- Medication list (FirstDataBank pharmacologic/chemical groups and ingredients)[46]
- Age > 64 (true)
- Female gender (true)

Support for attribute concept hierarchies was developed in order to address varying layers of granularity in native EHR data. Attributes from the EHR observations may be mapped to concepts at higher levels of abstraction. The framework uses a simple map of subsumption (‘Apple is a Fruit’) relations from an EHR attribute to a higher or subsuming concept. This functionality was treated at a very basic level in order to generate reasonable rules, given the degree of data granularity encountered during development. Semantic ontologies are the state-of-the-art knowledge engineering solutions to the variable granularity and relatedness of many concepts represented by native EHR data. Such comprehensive ontologies are highly valued as informatics infrastructure for many applications. They were out of the scope of the framework development reported. The framework development did prove the need for a semantic ontology in order to derive useful association rules directly from the EHR.

Given the possible candidate attributes per the national certification standards, iterative and detailed analysis of the EHR data content available in the study setting was conducted. Potential candidate data were analyzed for availability, consistency, quality and usefulness. Analysis included descriptive statistics and interaction with data stewards and domain experts. An approach to combine data across dimensions for presentation to the classification algorithm as a unified data set was challenging. The resulting methods were generalizable to many different categories of EHR data but represented the data at a very high level of abstraction. This enabled a core data set and methods that could accommodate multiple diseases or conditions and that could be used in any EHR setting. The design was modular and extensible to allow future enhancements.

Predictive Rule Set

Initially, proven associative classification methods were used to generate the predictive rule sets.[47-49] During development, rule sets were generated and evaluated repeatedly from different random samples. Rule sets were not as reliable as desired. Innovative methods to improve the generality of rule sets were developed. These improved the reliability. The development and research reported in the subsequent chapters focuses on the generation and testing of predictive rules sets.

Development/Research Setting

The cohort amplification framework was developed using data from a large, integrated health care delivery organization with a mature enterprise-wide, longitudinal EHR. The Intermountain Healthcare Enterprise Data Warehouse (EDW) provided the EHR data for secondary use that enabled this work. The EMR data of adult patients who visited an Intermountain Medical Group (IMG) central region Family Practice or Internal Medicine clinic at least once in 2005-2006 and at least once in 2007-2008 provided the target population for development and evaluation. This provided 106,250 eligible patients. Adult primary care[50] patients were chosen for the target population as Family Practice and Internal Medicine were visited by patients with a broad spectrum of conditions, and primary care was more frequently visited (~65% of adult patients in 2008) than any other IMG specialty among patients with diabetes and asthma. Primary care comprised 47% of all IMG adult ambulatory visits in 2008.

The required 2005-2006 visit was a ‘diagnosis period’ for the study. The required 2007-2008 visit was the ‘data mining period’ for the study. Two-year periods were used as this was the duration for many validated ICD-9-CM based algorithms to identify diabetes or asthma. A minimum of one visit in each of the study periods was required to provide a minimal amount of continuity of data. This requirement did not appear to create a biased study population. The average age and number of ambulatory visits/year were similar to the averages for Intermountain Healthcare ambulatory adult visitors.

Disease exemplars were identified from the coded Problem List during the diagnosis period. Using cases with previously documented disease, the classification rules were trained on their data in 2007-2008, the data mining period. Use of the Problem List has been integrated into the workflow in the central region IMG primary care setting. About 65% of all eligible patients have a coded Problem List. The Problem List was selected to identify disease exemplars because it includes both patients who present for treatment of the study disease, who are generally assigned an ICD-9-CM code, and those who present for other problems, in which case the study disease may not be assigned an ICD-9-CM code. The goal was to predict disease status, regardless of whether treatment was sought for that disease in the surveyed time period.

The disease-negative or control exemplars were generated from a random sample of all 106,250 eligible patients that had no evidence of the study disease by ICD-9-CM codes during the five-year period: 2003-2008 and no codes for the disease in the Problem List. The controls were not matched to cases on other demographic or risk factors because the prediction rules need to distinguish the cases from among all other patients in the EHR. Expected associations, such as a higher average age among diabetes cases, are absorbed into the associative classification rules-generation processes.

IRB approval was granted for this research from both the University of Utah and Intermountain Healthcare. The cohort amplification framework requires no protected health information.

Organization of the Manuscript

Associative classification was the approach used to develop prediction rules to identify new cases. Associative classification is described in Chapter 2, with emphasis on specific aspects relevant to this work. An overview of the functional processes of the cohort amplification framework and an evaluation of prediction rules generated for diabetes are presented in Chapter 3. Rules generated to identify diabetes mellitus were compared qualitatively to EHR-based rules published from other settings. The rules' accuracy was evaluated on test data. Chapter 3 was previously published. The development of rules to identify asthma, including original enhancements to the standard associative classification methods, is described in Chapter 4. Rule sets to identify asthma, generated by the both the standard and the improved methods, are compared on accuracy and generality. A comparative study of the accuracy of framework-generated rules to identify asthma cases in the EHR is reported in Chapter 5. Chapter 6 contains a summary discussion of this research.

References

1. Hersh WR. Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. Am J Manag Care. 2007 Jun;13(6 Part 1):277-8.

2. Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: An american medical informatics association white paper. J Am Med Inform Assoc. 2007 Jan-Feb;14(1):1-9.

3. Blumenthal D, Tavenner M. The "Meaningful use" Regulation for electronic health records. N Engl J Med. 2010 Jul 13.

4. Blumenthal D. Launching hitech. N Engl J Med. 2010 Feb 4;362(5):382-5.

5. Masys D. Extracting phenotypes from ehrs: The emerge consortium experience. 2009 Information Technology Roundtable Meeting. Washington, D.C.: The Clinical Research Forum; 2009.

6. Kullo IJ, Fan J, Pathak J, Savova GK, Ali Z, Chute CG. Leveraging informatics for genetic studies: Use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. Journal of the American Medical Informatics Association. September 1, 2010;17(5):568-74.

7. Kush R. Ehr-clinical research value/use case. 2009 Information Technology Roundtable Meeting. Washington, D.C.: The Clinical Research Forum; 2009.

8. Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL. Electronic medical records for clinical research: Application to the identification of heart failure. Am J Manag Care. 2007 Jun;13(6 Part 1):281-8.

9. Martin-Sanchez F, Iakovidis I, Norager S, et al. Synergy between medical informatics and bioinformatics: Facilitating genomic medicine for future health care. J Biomed Inform. 2004 Feb;37(1):30-42.

10. Maojo V, Iakovidis I, Martin-Sanchez F, Crespo J, Kulikowski C. Medical informatics and bioinformatics: European efforts to facilitate synergy. J Biomed Inform. 2001 Dec;34(6):423-7.

11. History of the development of the icd. World Health Organization.

12. Icd-10 : International statistical classification of diseases and related health problems : 10th revision. 2 ed. Geneva: World Health Organization; 2004.

13. Mullin R. A brief history of icd-10-pcs. J AHIMA. 1999;70(9):97-8.

14. Cpt - current procedural terminology. American Medical Association; 2008.

15. Aronsky D, Haug PJ, Lagor C, Dean NC. Accuracy of administrative data for identifying patients with pneumonia. American Journal of Medical Quality. 2005 November 1, 2005;20(6):319-28.

16. Ginde A, Blanc P, Lieberman R, Camargo C. Validation of icd-9-cm coding algorithm for improved identification of hypoglycemia visits. BMC Endocrine Disorders. 2008;8(1):4.

17. Jordan K, Porcheret M, Croft P. Quality of morbidity coding in general practice computerized medical records: A systematic review. Fam Pract. 2004 August 1, 2004;21(4):396-412.

18. Marklund B, Tunsater A, Bengtsson C. How often is the diagnosis bronchial asthma correct? Fam Pract. 1999 Apr;16(2):112-6.

19. Quan H, Li B, Saunders LD, et al. Assessing validity of icd-9-cm and icd-10 administrative data in recording clinical conditions in a unique dually coded database. Health Serv Res. 2008 Aug;43(4):1424-41.

20. Surján G. Questions on validity of international classification of diseases-coded diagnoses. International Journal of Medical Informatics. 1999;54(2):77-95.

21. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of icd-9-cm codes for identifying cardiovascular and stroke risk factors. Med Care. 2005 May;43(5):480-5.

22. Heubusch K. Coding's biggest challenges today. Journal of AHIMA. 2008;79(7):24-8.

23. Iezzoni LI. Risk adjustment for measuring healthcare outcomes. Second ed: Health Administration Press; 1997.

24. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: Icd code accuracy. Health Serv Res. 2005 Oct;40(5 Pt 2):1620-39.

25. De Coster C, Quan H, Finlayson A, et al. Identifying priorities in methodological research using icd-9-cm and icd-10 administrative data: Report from an international consortium. BMC Health Serv Res. 2006;6:77.

26. Miller DR, Safford MM, Pogach LM. Who has diabetes? Best estimates of diabetes prevalence in the department of veterans affairs based on computerized patient data. Diabetes Care. 2004 May;27 Suppl 2:B10-21.

27. Rector TS, Wickstrom SL, Shah M, et al. Specificity and sensitivity of claims-based algorithms for identifying members of medicare+choice health plans that have chronic medical conditions. Health Serv Res. 2004 Dec;39(6 Pt 1):1839-57.

28. Solberg LI, Engebretson KI, Sperl-Hillen JM, Hroscikoski MC, O'Connor PJ. Are claims data accurate enough to identify patients for performance measures or quality improvement? The case of diabetes, heart disease, and depression. Am J Med Qual. 2006 Jul-Aug;21(4):238-45.

29. Zgibor JC, Orchard TJ, Saul M, et al. Developing and validating a diabetes database in a large health system. Diabetes Res Clin Pract. 2007 Mar;75(3):313-9.

30. Starren J. Effective ehr phenotyping strategies. 2009 Information Technology Roundtable Meeting. Washington, D.C.: The Clinical Research Forum; 2009.

31. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: A review of recent research. Yearb Med Inform. 2008:128-44.

32. Chapman WW, Christensen LM, Wagner MM, et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing. Artif Intell Med. 2005 Jan;33(1):31-40.

33. Li L, Chase HS, Patel CO, Friedman C, Weng C. Comparing icd9-encoded diagnoses and nlp-processed discharge summaries for clinical trials pre-screening: A case study. AMIA Annu Symp Proc. 2008:404-8.

34. South BR, Chapman W, Delisle S, et al. Optimizing a syndromic surveillance text classifier for influenza-like illness: Does document source matter? AMIA Annu Symp Proc. 2008:692-6.

35. Turchin A, Kohane IS, Pendergrass ML. Identification of patients with diabetes from the text of physician notes in the electronic medical record. Diabetes Care. 2005;28(7):1794-5.

36. Meaningful use. Health Information Technology, U. S. Department of Health and Human Services; 2009.

37. Brown SH, Miller RA, Camp HN, Guise DA, Walker HK. Empirical derivation of an electronic clinically useful problem statement system. Ann Intern Med. 1999 Jul 20;131(2):117-26.

38. Burton MM, Simonaitis L, Schadow G. Medication and indication linkage: A practical therapy for the problem list? AMIA Annu Symp Proc. 2008:86-90.

39. Meystre SM, Haug PJ. Randomized controlled trial of an automated problem list with improved sensitivity. Int J Med Inform. 2008 Sep;77(9):602-12.

40. DesRoches CM, Campbell EG, Rao SR, et al. Electronic health records in ambulatory care--a national survey of physicians. N Engl J Med. 2008 Jul 3;359(1):50-60.

41. Wilcox A, Bowes WA, Thornton SN, Narus S. Physician use of outpatient electronic health records to improve care. AMIA Annu Symp Proc. 2008:809-13.

42. Albright LC. Computerized genealogies linked to medical histories for research and clinical care—a national view. AMIA Annu Symp Proc. 2006:1161-2.

43. Cannon Albright LA. Utah family-based analysis: Past, present and future. Human Heredity. 2008;65(4):209-20.

44. Certification commission for health information technology. 2010 [cited AMBULATORY EHR CERTIFICATION CRITERIA Jan 18, 2010]; Available from: http://www.cchit.org

45. 45 cfr part 170 health information technology: Initial set of standards, implementation specifications, and certification criteria for electronic health record technology; final rule (july 28, 2010). In: Services HaH, editor.: Federal Register; 2010. p. 44590-654.

46. Firstdatabank. [cited Feb 23, 2010]; Available from: www.firstdatabank.com

47. Agrawal R, Imielinski T, Swami A. Mining associations between sets of items in large databases. ACM SIGMOD Int'l Conf on Management of Data. Washington D.C.; 1993.

48. Li W, Han J, Pei J. Cmar: Accurate and efficient classification based on multiple class-association rules. First IEEE International Conference on Data Mining (ICDM'01); 2001; 2001. p. 369-76.

49. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. In: Intelligence AAfA, editor. KDD-98; 1998; New York; 1998.

50. Solutions to the challenges facing primary care medicine. Policy Monograph. Philadelphia: American College of Physicians; 2009.

CHAPTER 2

ASSOCIATIVE CLASSIFICATION

Associative classification (AC) is a data mining approach that uses two basic strategies. One is the deterministic and exhaustive generation of association rules between a predetermined outcome attribute and all other attributes, individually and combined, in a data set of training cases. Associations are co-occurrences of attributes within the same case. Exhaustive means that *all* associations are considered. The other strategy is classification, a general machine-learning task to assign a group status to cases in the target population based on patterns generated from training data with known group membership. Algorithms for classification – 'classifiers' - may be generated by many diverse methods including decision trees, Bayesian networks, statistical models, neural networks, support vector machines, covering rules, associative classification, and others.[1,2] In associative classification, the classifier is generated by the rigorous selection of a concise, general and accurate predictive subset of association rules from the exhaustive set of associations.

Data mining is one of the activities in the process of discovering knowledge from the large stores of data in databases. Fayyad et al. defined knowledge discovery from databases (KDD) as the "nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."[3] KDD has evolved from several fields including machine learning, databases, statistics, artificial intelligence, high performance computing, and data visualization. KDD has a unique goal within all of these, to find understandable patterns in the data that yield useful or interesting knowledge. The steps in the KDD paradigm are:

- Develop an understanding of the problem.
- Develop an understanding of the data.
- Prepare the data.
- Apply data mining methods.
- Evaluate and apply the discovered knowledge.

The data mining and evaluation stages may cycle back for a deeper understanding of the problem or the data.[4] Practically, the process is re-entered at any of the steps and flows downward. The steps are a ‘best practice’ model. Each is critical to a worthwhile project, but knowledge engineers accomplish them with different emphasis and methods. KDD is not a scientific method. It does not necessarily involve a specific hypothesis about the pattern-discovery outcomes, although it is purposed toward generating hypotheses for further scientific study. Rather, KDD is a disciplined approach and a collection of proven methods to provide useful knowledge from existing data repositories.

Han and Kamber[1] describe data mining as an evolution of the powerful databases that have become pervasive in the last decades. Powerful computers and advanced services for data analysis, coupled with volumes of data collected and stored in databases, led to increased interest in machine learning and pattern recognition. The data were available for mining for golden nuggets of new knowledge and useful information. They elaborated on the steps of KDD above:

- Preparation of the data includes
    - Gathering data from multiple sources.
    - Removal of outliers and inconsistent data.
    - Attribute selection and transformation.
    - Input to the data mining programs.
- Application of data mining methods includes
    - Selection of methods and algorithms.
        - More than one method may be applied on the same data.
- Evaluation of the knowledge includes
    - Presentation of results to the users
        - Visualization techniques.
        - Knowledge representation techniques.

Witten and Frank[2] described data mining as the search, discovery, and expression of patterns in the data. The data come from databases and are usually large data sets. The processes to find patterns are automated or semiautomated with computers. The patterns discovered should be meaningful or useful. They should help us understand the data and/or make predictions from them. They elaborated the KDD steps further:

- Preparation of the data includes explicit handling of missing values in the training data. Most data mining methods function under the assumption that missing values are random. If missing values are correlated with other data, appropriate assumptions should be accommodated in the preprocessing. Interestingly, they give an example of missing medical tests as highly correlated with disease since doctors do not order them if they were not related. They comment that the nonexistence of a test may be as predictive as the actual values. This foreshadows results reported in this research.

- Evaluation of the knowledge includes a broad array of methods. Data mining algorithms and their specific result are a ‘theory’ over the training data, and thus evaluation may take a philosophical tone. The entire KDD process must essentially be evaluated in context. The normative method for evaluation of classifiers is validation of the results on repeated random samples of the training and test data with descriptive and statistical measures of the accuracy of results.

The selection of the training data is an important aspect of the evaluation of knowledge gained in KDD. The larger the number of training cases, the more the reliability of results might be assumed. However, generalization of the results to a wider population depends upon the cases that were used for training. The selection of the training cases was noted as a critical step in machine learning.[5] Selection of training data is often a pragmatic choice, given domain-specific issues in accessibility of data. Competitor businesses may not agree to pool their data so training may describe only one company’s experience. In health care, training data may be limited to one provider organization due to data security and privacy concerns. Even within one health care organization, there are stringent policies for protection of medical records such that the process of obtaining access to data is a significant additional process step.[6]

There are many categorizations of data mining methods. They can be categorized by the purpose of the knowledge that is sought: one purpose is classification.[1] The purpose of classification is to assign a ‘case’, the object to be classified, to a group, using the available data that describes the case. Most classification data mining methods generate a ‘classifier’ from training data that has an explicit group ‘label’. The group label is called a ‘class’. This is called ‘supervised’ machine learning: the classifier was trained on data where the class was previously assigned. The purpose of the classifier is to ‘classify’ a new case, in which the class is not known from the data. Classifiers use many different algorithms to perform classification. They may be categorized by their approach to generating a classifier (Bayesian network, decision tree, neural network, and others), and further categorized on different algorithms employed to actually engage the data, render a classifier and apply the classifier to a new case. Associative classification is one type of approach, which is implemented using various algorithms.

Associative classification (AC) may also be called classification by association, classification association rule mining, and other derivative terms. It was also described as affinity analysis.[7] The concept was first described by Bayardo[8] as a “brute-force technique for mining classification rules from large data sets.” He introduced the idea of association rules between a predetermined class attribute and all other attributes as classifiers. It was called ‘brute force’ because of the exhaustive generation of associations. However, Liu et al.[9] are usually credited with the introduction of a more proper associative classification algorithm because they also used strategies to select the rules *most likely to be predictive* from all the association rules. The exhaustive set of associations captures the patterns that are unique to the training data, known as ‘noise’, as well as those more generally representative of a larger target population.

To generate classification rules from association rules, there must be a 'hypothesis' for the selection of general rules over rules more specific to the training sample. Classification rule discovery is an inductive task, predictive of the future, whereas association rule discovery is a deductive task, descriptive of the present.[10] The main strengths of AC are the global view of all associations in the training data and the use of combined attributes for pattern discovery. The main weakness is the generation of a large number of associations, which reflect the noise in the training data as well as the reliable associations.

Thabtah[11] described the steps in associative classification as:

(1) Discovery of associations among the training data attributes.

(2) Generation of association rules between the class and other attributes.

(3) Ranking and pruning of weak rules to form a classifier.

(4) Prediction on test data and evaluation of the classifier's accuracy.

In this chapter, an overview of association rule mining using the Apriori algorithm (steps 1 and 2 above) is presented. Improved methods for constraining and pruning association rules to form useful classifiers are an ongoing topic of computer science research. A novel pruning approach was developed in this research. Therefore, the background on associative rule classifiers (steps 3 and 4 above) are covered in more detail. Concept hierarchies in association rule mining and applications of associative classification in biomedical research are summarized.

A small example data set was created (Table 2.1) to clarify the technical explanations in this chapter. The data set will be referred to as the 'Disease Prediction' data.

Association Rule Mining

The foundational methodology of association rule mining (ARM) was first described by Agarwal et al.[12] Their approach was seminal in that *all* associations were measured between the attributes in a data set, configuring each individual attribute as a rule consequent and all remaining attributes as a candidate for union in the rule antecedent:

> Given the rule, 'If A, then C': 'A' is the antecedent, and
> 'C' is the consequent.

The problem domain was retail sales, and the objective was to discover purchasing patterns and express them as association rules. The rules were described as qualitative and deductive, as opposed to quantitative methods. Quantitative methods generate predictive models using an inductive approach. The methods vary greatly in their approaches to pattern finding, i.e., statistical regression analysis, Euclidean geometry, and Bayesian probabilities. Generally, they sequentially process the training data to build and test a predictive pattern that best fits the data. In distinction, ARM finds and reports all patterns found in the training data. The ARM approach was targeted at an uncontrolled setting, where many interdependencies coexist in the data. Domain knowledge was not a requirement to deduce the patterns, but may certainly be required to interpret and refine them for prediction.

The fundamentals of ARM were established. The attributes in the model were called 'items', and the combination or *association* of items, 'itemsets'. The itemsets of interest were limited to those that occurred frequently in the training data set. The user would specify the minimum frequency of interest. In the Disease Prediction data (Table 2.1), if a minimum frequency of occurrence over all training cases was specified as 30%, then all of the attributes are ‘frequent’ items. The frequent 2-item itemsets are ‘DrugA & TestC’, ‘DrugA & Class.PosDisease’, and ‘TestC & Class.PosDisease’. There is only one frequent 3-item itemset: ‘DrugA & TestC & Class.PosDisease’. Once all of the frequent itemsets are found, the rules are formed. In association mining, all frequent items in an itemset are permuted as the antecedents or consequents of rules. The rule ‘If DrugA & TestC, then Class.PosDisease’ would be one of the rules generated. All rules are then evaluated by a user-specified threshold for a measure of precision: the likelihood that the rule consequent occurred, given that the rule antecedent occurred. Since the rule consequent occurred in 30% of the cases and the antecedent occurred in 40%, the likelihood is 75%. These are the basic pattern-discovery methods for association rule mining.

In a subsequent paper, the Apriori algorithm was described.[13] The algorithm accomplished the same objectives for ARM as described but improved the computing efficiency. ARM, without constraints and smart computer programming algorithms, is of exponential complexity in the number of attributes. Much of the computer science literature on ARM deals with more efficient computing techniques. The size of the solution space is significant for its computing requirements as well as the large number of association rules that may be generated. Recall that one of the goals of data mining is understandable rules. The authors of Apriori acknowledged that ‘application-dependent’ constraints are necessary features of an association rule discovery system.

The Apriori ARM algorithm can be constrained to one specified attribute as the consequent of all rules. In this case, all other attributes are candidates to form the rule antecedent. For simplicity, assume that each attribute represents an item, and thus its value exists or it does not exist. Only the attributes that exist are populated in the data. This was the representation of retail purchase data used in the presentation of the Apriori algorithm. The Disease Prediction data (Table 2.1) also use this representation. Assume that the desired minimum frequency of interest is 'n', and the desired precision of the rule given the antecedent is 'm'. Apriori scans the data set, counts all attribute occurrences, and enumerates those attributes that occur among n% of the records. These are the frequent items. The next step is to find frequent two-item itemsets. Only frequent items can possibly combine to form frequent two-item itemsets. The combined itemset cannot occur more frequently than any one of its members. The two-item itemsets are combined into frequent three-item itemsets and this process repeats until there are no more frequent itemsets. After all frequent itemsets have been enumerated, the candidate rules are formed from the one specified consequent attribute and all other members of its frequent itemsets as the antecedents. All such frequent rules are then reduced to the set in which the precision of the rule is greater than or equal to 'm'.

The publication of the Apriori algorithm concluded with the need for two extensions: support for concept taxonomies and handling of discrete and continuous attributes. Current applications of ARM generally support discrete attributes. Continuous attributes must be discretized. The values of discrete attributes are considered as the items to be associated. Other methods to perform association rule mining have been developed subsequent to its introduction with the Apriori algorithm.

They mainly address improvements to computing complexity but may also include alternate methods to evaluate the precision of rules. ARM, by definition, is an exhaustive data scan for associations, given a frequent itemset threshold.

Association Rule Classifiers

Associative classification is a specialization of association rule mining and one of the data mining approaches used to generate classifiers. Freitas[10] presented a thorough discussion of the additional requirements for association rules to be considered as classifiers. The Disease Prediction data (Table 2.1) show an attribute 'Class' with the values 'PosDisease' and 'NegDisease'. AC must be constrained to the classifying attribute as the only rule consequent, in order to form rules predictive for the class. Another requirement is that the accuracy of the classification rules must be evaluated on test data. The more important differences are theoretical. The task of classification is inductive and nondeterministic. A classification hypothesis is generated from training data, and its predictive success is estimated on test data. Two major problems arise: (1) Overfitting or underfitting the classification hypothesis to the training data is a main concern in classification. ARM generates all rules, given constraints, over the training data. (2) All classifiers have an inductive bias: explicit or implicit criteria that influence the classifier to favor one hypothesis over another. The methods and the configuration of the classifier form the bias for one hypothesis to be preferred over another. Further, the bias is known to be domain-dependent: the classification methods must interact with a specific data set to form a classifier. ARM is deductive and deterministic. Both problems are addressed in AC by various constraints, pruning methods (algorithmic identification and elimination of weaker rules), and use of class-specific definitions of frequency for the selection of frequent itemsets. AC cannot be constrained and pruned for overfitting/underfitting avoidance or inductive bias in a general set of methods to suit all domains and data sets. However, such tuning can be customized for unique domains. The Pruning Methods section of this chapter presents general and domain-focused pruning methods used in health care data.

AC has been shown to perform as well as or better than decision trees, rule induction methods, and the naïve Bayes classifier on benchmark classification data sets.[9, 14-20] The potential advantages of association classification over other classifiers are:

- AC discovery is global: all interesting association rules are discovered and then pruned to a more concise and general set.
- Combinations (union) of attributes are used for pattern discovery.
- AC was designed for application in noisy, highly dimensional and interdependent data such as the operational transactions of an enterprise.
- Multiple hierarchical concept levels (taxonomies) can be mined for patterns in mixed or matched models.
- Missing values can be configured to participate in the discovered associations or not, at the discretion of a domain expert. If included, missing items participate in associations in the same way as a nonmissing item.
- Generated rules are understandable to users.
    - Rules are independent and can be modified by users or joined with other rule sets.

- Rules are amenable to interpretation as queries against the database mined, being "if-then" statements over arguments that represent attributes in the database.

AC has attracted researchers in the data mining and machine learning communities since it was first described in 1998. There were 66 academic publications on 'associative classification' in Scopus[21] for 2008-2009, which was higher than the previous two years at 42. PubMed[22] only listed 4 publications, all in recent years, but listed 83 publications on association mining since 2001. The keyword "data mining" was added to the Medical Subject Headings (MeSH)[23] in 2010, which suggests the growth of data mining research, in general, in the health care field.

Interestingness Metrics

'Interestingness' is a key data mining term. In classification, the hypothesis is guided and evaluated by a method's interestingness metrics and their thresholds, if specified for the application. Many statistical, mathematical and heuristic interestingness metrics have been used in data mining. Interestingness has at least three interpretations. One is the objective thresholds and parameters that are used to configure a data mining algorithm. Many, if not most, algorithms include user-specified parameters that bound the algorithm's functionality. For example, a statistical Type 1 error threshold is common in algorithms that use statistical comparisons over the mined data. In ARM, the frequency and precision thresholds for rule generation are examples. These are all considered 'interestingness' metrics because they constrain the algorithm's results to those perceived as useful by the user. A second interpretation is the criteria used to evaluate classification results. In health care applications, the sensitivity and specificity of a classifier may be used to compare two classifiers. A third interpretation is the subjective concept of true interestingness of the knowledge generated. These metrics or outcomes depend upon the goals of the consumers of the knowledge.[24, 25] Data mining is an engineering task and not a scientific method. Interestingness is an applied term and has been used somewhat ambiguously, although it is fundamentally an expression of the heuristic bias or merit of a data mining approach. In the following review, objective interestingness metrics used in association rule mining and associative classification will be explored.

Geng and Hamilton[26] surveyed interesting metrics for data mining and synthesized them into five objective criteria: conciseness, generality/coverage, reliability, peculiarity, and diversity. The first three are common to associative classification and germane to this research. Concise rules are valuable because they are more understandable to domain experts who may subjectively evaluate and refine the classifier.[27] Webb and Brain[28] provided rigorous proof for the preference of a more general rule to a more specific one, given all other evidence was equal. General rules and concise rules are intuitively related. A more general rule covers more of the training cases than a less general rule, since that is what defines a 'general' rule. Concise rules are the smallest set of rules that, together, achieve the best accuracy on test data. Therefore, assuming all other effects are equal, a set of more general rules should be a smaller or more concise set than a set of less general rules. Reliability is the accuracy of a set of rules that form a classifier. Measures of accuracy in AC commonly include 'confidence', a term that was described above by the synonymous term 'precision' to constrain and rank the rules. There are also measures of the predictive accuracy of the classifier on test data.

Measures of Generality

To generate prediction rules from the association rules, the most concise set of rules without loss of accuracy is desirable. General rules are more comprehensive and cover more of the dataset. Assuming general rules comprise the concise rule sets, the metrics for measuring generality will be addressed. The most commonly used measure of generality in AC is the concept of ‘support’. Support is a statistical measure of the frequency or likelihood of the occurrence of a rule. The concept of frequent itemsets in the Apriori algorithm uses a minimum support threshold to define ‘frequent’. As presented in the Association Rule Mining section, the minimum frequency of occurrence specified by the user forms a lower bound on the attributes and rules that will be considered ‘interesting’. This is a constraint used to guide the generation of an associative classifier. Assuming the rule antecedent is ‘A’ and the consequent is ‘C’, then

support (A → C) = support (A U C)[1] = count (A & C) /
count (training cases)

support (DrugB → Class.PosDisease) = 2 / 10 = 20%
(from Table 2.1)

Support is used to define the minimum threshold to allow large itemsets to participate in ARM, and is used in AC to rank rules. Gu et al.[29] introduced a specialization of support for associative classification in the health care environment. When mining health care data to classify a specific disease, the nondiseased population is usually greater so class sizes are often unbalanced. The common interestingness metrics fail to capture many interesting rules for the rarer class. A specialization of support was defined to represent support for each class fairly, regardless of the sample sizes. Such adjustments were recommended in order to apply association rules to prediction tasks[10] and were proven to be very effective in finding useful rules for rarer classes.[30] The local support (LSUP) for each class was defined as the support for the rule for the class. It also expresses the probability, based on the data, that the rule occurs when the class occurs. This metric is derived by dividing the support for the rule by the support for the class.

$$\text{LSUP}\ (C_i) = \text{support}\ (A \rightarrow C_i)\ /\ \text{support}\ (C_i) = \text{Probability}\ ((A\ \text{U}\ C_i)\ |\ C_i)$$

LSUP (DrugB → Class.PosDisease) = 20% / 50% = 40%

(from Table 2.1)

When mining a database directly, as in the reported application, one can adjust for the differences in prevalence of the two classes by using the local support metric. However, most off-the-shelf association rule mining software packages use the support metric and not the local support metric. In the case of a binary classifier, another way to approach the problem is to draw a balanced sample from the database so that the two classes are evenly distributed. When sample sizes are large, there are no statistical restrictions to the use of balanced random samples from unequal reference population sizes in terms of representativeness. A colorful analogy was drawn to illustrate this concept: a small spoonful of soup sampled from different sized pots will give the cook the same quality of information from each, provided the soups were stirred.[31]

With balanced sample sizes, the local support metric is roughly twice the support metric for each class of a binary classifier:

$$\text{LSUP}(C_i) = \text{support}(A \rightarrow C_i) / 0.5$$

Thus, local support and support can be approximated from each other, given a binary classifier.

Measures of Accuracy

Precision, or the likelihood that a rule consequent occurs given the rule antecedent occurs, was presented in the section on Association Rule Mining. This measure of precision is commonly called ‘confidence’ in association rule mining and associative classification. Confidence is used to define a user-specified minimum threshold to constrain the association rules that may be formed from the candidate frequent itemsets. After all frequent itemsets are found which satisfy the support interestingness metric, potential rules must satisfy the confidence or precision interestingness metric. Confidence is also used in the pruning processes as explained in the Pruning Methods section. In AC, confidence is a measure of the likelihood (the probability based on the given data) of a particular class occurring as the rule consequent, given the rule antecedent. This metric is derived by dividing the support for the rule by the support for the rule antecedent. The rule consequents, by definition in AC, are assignments to one of the class outcomes. The rule antecedents may occur in any or all class outcomes.

$$\text{confidence } (A \rightarrow C_i) = \text{support } (A \rightarrow C_i) \; / \; \text{support } (A) = \text{Probability}((A \text{ U } C_i) \mid A)$$

confidence (DrugB → Class.PosDisease) = 20% / 30% = 67%

(from Table 2.1)

Gu et al.[29] also introduced a specialization of confidence for situations where class sizes are unbalanced. Confidence is weakened for a rarer class by the 'support (A)' term in the denominator above. If the larger class is represented 10:1 for a disease with 10% prevalence, then 'support (A)' has 10:1 counts for the antecedent favoring the larger class of a binary classifier. As the ratio of larger class to rarer class grows, it can be seen the denominator in the formula above grows and thus confidence for C_i decreases. The exclusiveness (EXCL) metric was defined and proven to normalize the confidence metric for each class fairly. It was defined for the binary classifier (C_i, C_j) as:

$$\text{Exclusiveness } (C_i) = \text{LSUP } (A \rightarrow C_i) \; / \; \text{LSUP } (A \rightarrow C_i) + \text{LSUP } (A \rightarrow C_j)$$

Just as most off-the-shelf association rule mining software packages use the support metric rather than local support, they use confidence and not exclusiveness. In the case of the binary classifier and balanced class sizes, confidence and exclusiveness for a rule are equal. Referring to the conversion formula from local support to support when sample sizes are balanced given above:

$$\text{LSUP } (C_i) = \text{support } (A \rightarrow C_i) \; / \; 0.5$$

Inversely: $\text{support } (A \rightarrow C_i) = (.5) \text{ LSUP } (C_i)$

$$
\begin{aligned}
\text{confidence } (C_i) = \quad & \text{exclusiveness } (C_i) \\
= \quad & (.5)\ (\text{LSUP } (A \rightarrow C_i)\ / \\
& (.5)\ (\text{LSUP } (A \rightarrow C_i) + (.5)\ \text{LSUP } (A \rightarrow C_j) \\
= \quad & \text{support } (A \rightarrow C_i)\ / \\
& \text{support } (A \rightarrow C_i) + \text{support } (A \rightarrow C_j) \\
= \quad & \text{support } (A \rightarrow C_i)\ / \\
& \text{support } (A)
\end{aligned}
$$

In the denominator of the final equation, the total support for A was distributed across the two possible rule consequents for A. Added together, they comprise the total support for A. Thus, exclusiveness and confidence can be approximated from each other, given a binary classifier.

Table 2.2 shows the magnitude of difference of the support versus local support and confidence versus exclusiveness metrics for a binary classifier of disease with 10% prevalence between a hypothetical balanced versus representative class sampling strategy. The local support and exclusiveness are normalized and independent of the underlying prevalence in a representative sample. The support and confidence are affected by the differences in prevalence in the representative sampling strategy. The framework used balanced sample sizes to approximate the local support and exclusiveness from support and confidence metrics.

Measures of predictive accuracy in association classification are the same as those used for many other classifiers. The results of classification are commonly viewed in a ‘confusion matrix’,[32] as described in Table 2.3. Further, the confusion matrix may be repeated multiple times on separate random samples of the mined data, using different cases for the training and test sets for each run. With repeated samples, the accuracy metrics can be evaluated statistically. This evaluation approach is called cross-validation.[2] The metrics used in this research to evaluate predictive accuracy are sensitivity and specificity because these are commonly used in the health care domain. More specifically, they were used to evaluate other classification algorithms that will be compared to AC classification in this research. The sensitivity and specificity metrics are independent of the sample proportion. Sensitivity is the proportion of the classified population determined to be *positive by a reference standard* and *classified as positive*.

Sensitivity = True Positives /
True Positives + False Negatives

Specificity is the proportion of the same classified population determined to be *negative by the same reference standard* and *classified as negative*.

Specificity = True Negatives /
True Negatives + False Positives

Sensitivity and specificity are inversely related in all but the two boundary examples: all cases classified correctly or all incorrectly. This is demonstrated in Table 2.4. The possible outcomes that occur for positive determinations by the binary classifier are ‘true positive’ (TP) or ‘false positive’ (FP). The possible outcomes that occur for negative determinations by the binary classifier are ‘true negative’ (TN) or ‘false negative’ (FN). The binary classifier must generate a positive or a negative determination for each case. Assume that initially the sensitivity is zero, and the specificity is 100%. That is, no true positives have been identified and no true negatives have been misidentified. As cases are classified as positive, a true positive case will *increase* the sensitivity, and a false positive will *decrease* the specificity. At each positive classification determination, either the sensitivity or the specificity metric is affected in the opposite direction. Once all cases are classified, the sensitivity will have increased from zero, and the specificity will have decreased from 100%. The cumulative TP and FP proportions are reflected in receiver operator characteristic (ROC) curves, which permit visualization of the tradeoff between the two plotted on an X-Y axis.[33] Both a high sensitivity and a high specificity are desirable, in general. However, depending upon the purpose of the application of a classifier and the user's subjective interestingness preferences, a higher sensitivity or a higher specificity may be the preferred accuracy outcome.

Pruning Methods

Since associative classification (AC) is a specialization of association rule mining (ARM), it inherits the limitations of ARM. A large number of rules are generated since attributes are often highly correlated and, therefore, associated. The high correlation of attributes follows from direct mining of operational data, which is one of the fundamental objectives of ARM. The global nature of rule discovery casts a net for all potential interesting patterns, but many redundant rules are discovered as well as rules that reflect idiosyncrasies of the training data (overfitting). Therefore, it is necessary to apply 'pruning' methods to *eliminate* (prune) redundant and weak class association rules in order to develop a general and accurate associative classifier. Association rules have no basis for preference of one set of rules over another, other than the support and confidence thresholds. Multiple sets of rules might predict the class equally well on the training data. Pruning methods address the rigorous ranking and selection of class-constrained association rules in order to generate a concise, general and accurate classifier.

Thabtah[34] provided a recent review of the main pruning approaches used in AC. The initial AC model, known as 'classification by association' (CBA),[9] is still accepted and used as a benchmark for new AC methods. CBA was based on the Apriori algorithm for rules discovery. Subsequently a model, known as 'classification based on multiple association rules' (CMAR)[20] was introduced. CMAR is also a classic benchmark for AC methods. CMAR uses a different strategy for rules discovery, accommodating some of its pruning methods earlier in that step. Accuracy using CMAR was shown to be equivalent to CBA.

CBA Pruning Methods

CBA uses the *database coverage* method. More than one rule may cover the same case to be classified. The rules discovered by association rule mining are ranked in order of confidence, then support. Processing the rules in ranked order, the training cases that meet the rule are removed from further consideration by a subsequent rule. If no case meets a rule, the rule is pruned. This continues until all training cases have been covered or all rules were tested. Cases left uncovered are assigned to the class with the highest frequency in the training data. The database coverage method seeks the most accurate rules, by rank, and eliminates less accurate rules that cover the same cases.

An optional method is *pessimistic error* pruning, originally defined for decision trees. The method assesses if the error rate for the majority class at a node in the tree is less or equal to the classification error of its branch nodes. If so, the branch node is pruned. This method has not been implemented in AC as popularly as a similar method, redundancy pruning, described below. They both compare the gain in accuracy from a parent node to a child node, and prune the child node if it does not improve accuracy. A parent node is a more general rule than its child nodes and covers at least the same cases covered by the child node's rule.

CMAR Pruning Methods

CMAR uses the *redundancy* pruning method. Redundancy pruning is implemented before the rules are ranked. Multi-attribute rules that cover the same or fewer cases and do not improve the confidence of a more general rule, e.g., a subset of the multi-attribute rule, are redundant. If rule 'R_a' was met, then rule 'R_a U R_b' is redundant unless its confidence is greater. Rule 'R_a U R_b' (child rule) cannot have more support than 'R_a' (parent rule). At most, it can have equal support. The redundancy pruning method results in fewer, more general rules.

CMAR applies a chi-square test to prune rules before the rules are ranked. If the rule antecedent and the rule consequent are not positively correlated, the rule is pruned. This pruning method must be provided in the data mining software or programmed to use with statistical tables for the significance of the chi-square scores.

CMAR also uses a variation on the database coverage method in CBA. The rules are ranked by confidence and support. The rules are tested for coverage in the training data in ranked order, as in CBA. The difference is that a threshold on the number of times a case may be covered is set. A covered case can be re-covered multiple times, generating several potential rules that may cover one case. In the classification step, if all rules that cover a new case agree on the class, the class is assigned. If there is disagreement among the consequent among the set of classification rules that cover a new case, a normalized chi-squared test is used to assign the most accurate class. This pruning step was meant to overcome the problem of favoring only one most confident rule. Other slightly less or equally accurate rules may have higher support, and thus serve as better classifiers. The overall confidence of the ranked rules and the user-specified threshold on the number of rules that can cover each case will vary the effects on accuracy.

The pruning methods from CBA and CMAR are objective, general and effective methods. Pruning methods may be mixed from among these and may also be combined with other approaches.[35, 36] New pruning approaches may be compared with one or more of these basics.[27, 34] Pruning may have a domain-specific rationale. The two objectives are to remove redundant and misleading rules for the classification task at hand.

Pruning methods often include the ranking of rules. Most AC pruning algorithms that depend upon rule ranking use confidence (descending), support (descending) and then cardinality (ascending). The cardinality ranking supports rules that are more general. Thabtah[37] proposed and tested two additional rankings: the frequency (descending) of the rule consequent, then precedence of the antecedent in the training data. These slightly improved the average accuracy on highly dense datasets. Rule preference affects the accuracy of the classifier. Improvements in rule ranking was listed as one of the interesting research directions in associative classification.[38]

The purposes of pruning may be accomplished by constraints on the rules generation process. Ordonez et al.[36] introduced a constraint on the number of attributes that may participate in the association rule antecedent of an associative classifier for heart disease. After assessment of the impacts of the constraint at sizes from one to five, they used a limit of four attributes per rule antecedent in the rules discovery process for the classifier. Because the high dimensionality in medical data results in many associations, this constraint disabled many redundant rules from forming in the first place. It is straightforward to constrain the size of the large itemsets in the Apriori algorithm, as they are generated in cycles of ascending size. This has the effect of global redundant rule pruning, given that additional attributes after the specified size do not add accuracy to the model. This constraint also enables the Apriori algorithm to generate associations at lower support levels since the computing complexity is reduced to a small exponent.

Concept Hierarchies

Concept hierarchies have been discussed in the ARM literature since its inception.[13] With the objective of ARM to discover rules in large databases directly, the fine granularity of data comprising an enterprise’s operational transactions might obscure the interesting patterns. For example, in the medical data domain, there could be many drug formulations in the same therapeutic class that might each exhibit very similar association patterns but may each be diluted by low support. Assume that the drug formulations are stored in the training dataset, but they can be linked to a therapeutic class in a taxonomy. If the associations with the therapeutic class rather than the formulations were exposed, there may be greater support for an interesting association at the therapeutic class level.

Using the Disease Prediction data (Table 2.1), assume that DrugA and DrugB are both used to treat the disease and share the same therapeutic class. Individually, the only interesting rule selected would be:

DrugA → Class.PosDisease confidence 4/5 = 80%, support 40%

A more interesting rule would be generated from the therapeutic class containing both:

DrugA or DrugB → Class.PosDisease confidence 5/6 = 83%, support 50%

Operational data are often categorized into larger concepts for analysis and reports in an enterprise data warehouse. A data element may belong to multiple categories. In the drug ingredient example, there may be taxonomy of drug ingredients. A formulation may belong to a therapeutic class and have one or more ingredients. The optimal concept level is generally unknown when the rules discovery task commences.

Han and Fu[39, 40] described approaches to managing multiple concept levels in association rule mining. Initially, Han emphasized user interaction to resolve the complex, domain-specific concept levels. Subsequently, methods were developed to encode the concept levels into the transaction data. The original methods were difficult to accomplish technically and supported only hierarchical concepts. A transaction item could belong to a taxonomy of concepts; like ‘chocolate milk’ is ‘2% milk’ is ‘milk’. Methods for associating the concept levels ranged from associating across one level at a time to cross level associations. There was thought toward varying support and confidence thresholds by level, and a pruning technique to recognize and disallow a rule to contain an item and its conceptual ancestor (‘2% milk’ and ‘milk’). The technology frontier has advanced considerably in the past decade. Current research and development in data mining includes linking of the entire knowledge discovery lifecycle to domain-specific ontologies. Two recent projects describe association mining with semantic links to the Unified Medical Language System (UMLS) ontologies for data selection, pattern mining, data visualization and to provide some guidance normally provided by domain experts.[41, 42] The linkage of databases to semantic networks opens up a completely new dimension to knowledge discovery.

The data mining software package used in this work was the Waikato Environment for Knowledge Analysis (WEKA) toolkit.[2] It does not offer tools to accomplish even a simple concept taxonomy linkage as described above. In the research reported in subsequent chapters, associative classification using multiple concept levels was implemented with a simple set of mappings from EHR-stored data to more general concepts for the purpose of aggregating very granular data. Cross-level associations were allowed. The more general concept was ranked higher and, therefore, less general concepts were pruned. The research focus was the effect of aggregations on results and not the technologies used. The data mining methods and linkages to accommodate the complex semantic relationships of stored health care data were beyond the scope of this research.

Applications in Biomedical Research

Longitudinal electronic health records (EHR) are a perfect setting for association rule mining. The longitudinal EHR covers both ambulatory and hospital care. At this time, an EHR contains records of care for a particular healthcare delivery system only.

The EHR contains highly dimensional data. There are many choices of disease labels, treatments and diagnostic test ‘items’ attached to a patient encounter. The large set of possible ‘items’ are sparsely populated in each patient record. The patient typically has a different set of ‘items’ at each encounter. There is noise among these data. The health care providers who assign the ‘items’ sometimes make intelligent guesses about diagnosis and treatment, and these may prove to be mistakes in terms of the true disease status. There are also individual and group provider biases in the ‘item’ choices, such as medications prescribed for a particular indication. The patient may attend another health care facility, so there may be an incomplete record of care. However, there is a rich tapestry of associated ‘items’ that describe patterns across many patients’ encounters. Further, the ‘items’ reflect the decision making of providers, which form dimensions of interest. For example, given incomplete individual patient data, the patterns may emerge for differences in treatments given by primary care practitioners versus treatments given by specialists. Association rule mining describes multidimensional aggregate patterns, and, therefore, may provide new knowledge from sparsely populated, noisy and incomplete electronic health records.

Researchers have applied various association rule mining approaches to an assortment of problems using electronic health records (EHR) and secondary, anonymous patient data repositories. McAullay et al.[35] developed a framework for end users to mine the predictive attributes for adverse drug events directly from EHR data. They used static and sequential association rule methods, developing a classification model focused on rare events. Ordonez et al.[36] developed methods for association mining over a cardiovascular clinical database to classify heart disease. Li et al.[43] developed an association rule mining method based on frequent itemsets and relative risk to discover risk patterns in EHR data. They applied the method to risk of hospital admission from the emergency department (ED) based on data routinely collected in the ED. Mahamaneerat et al.[44] used ‘Domain Concept Mining’ to discover clinically meaningful associations in the 2005 Nationwide Inpatient Sample (n = 8,000,000 admissions). Elfangary and Atteya[45] used association rules to discover novel patterns in an inpatient nephrology clinical data system. They reported that the patterns were expressed in a manner that physicians could understand, and the mined rules were accepted by nephrology specialists. Wright and Sittig[46] used association rule mining to develop content for order sets in an ambulatory computerized physician order entry system. Tai and Chiu[47] applied association rule mining to study comorbidities of attention deficit disorder in the National Health Insurance Database of Taiwan. They generated new knowledge on developmental delay and associations with progression to other psychiatric illnesses.

Associative classification has not been applied to the problem of generating rules to identify cohorts of patients with particular conditions from secondary EHR data. Secondary data are those collected for patient care purposes and subsequently used for other legal, ethical and beneficial purposes.[48, 49] Although the task is to classify patients who have or have had a particular condition or disease from those who have not, this problem and solution space are unique from the problem of predicting disease for health care purposes. Some conceptual differences are listed below:

• **Identifying cohorts with disease**	**Predicting disease for care**
• QI, research, public health	Patient care decision support

• Data by-product of care processes	Concurrent operational care data
• Enterprise data warehouse	EHR transaction repository
• As accurate as possible	Exacting accuracy required
• Validation of cohort prediction	Validation of patient prediction
• Accept available input data	Input data desired from workflow
• Retrospective data view	Concurrent data view
• Deductive knowledge value	Inductive knowledge value
• Target user is not a clinician	Target user is a clinician

These differences are not absolute. They are listed to point out distinctions in a primary patient care use case and a secondary data use case for identification of disease status. The secondary data consumer must abide the quality of data that are available from the care processes that have already occurred, sometimes years earlier. Association rule mining is well suited to the task. The data mining method will expose the patterns that are there. Associative classification will expose the patterns associated with a health condition or disease state, provided there are representative training cases where the disease status can be inferred. If the disease statuses were accurately known for a large population of patients, one might question the need to develop a classifier. That is not the case. Patient records have not been consistently and accurately labeled for diseases treated, and less so for comorbidities. The current research describes and evaluates an associative classification framework to identify cohorts of patients with particular conditions from secondary EHR data.

Table 2.1 ‘Disease Prediction’ Data Set Example

	Attributes			
	DrugA	**DrugB**	**TestC**	**Class**
Case 1	Yes		Yes	PosDisease
Case 2	Yes	Yes		PosDisease
Case 3	Yes		Yes	PosDisease
Case 4		Yes	Yes	PosDisease
Case 5	Yes		Yes	PosDisease
Case 6				NegDisease
Case 7				NegDisease
Case 8				NegDisease
Case 9	Yes	Yes	Yes	NegDisease
Case 10				NegDisease

Table 2.2 Associative Classification Interestingness Metrics
in Representative and Balanced Class Samples

Hypothetical Disease with 10% Prevalence				
Representative Sample Sizes				
	Rule +	**Rule -**		
Class 1	160	40	200	
Class 2	90	1710	1800	
	250	1750	2000	2000
	Support (Rule + -> Class 1)			8
	LSUP (Rule + -> Class 1)			80
	Confidence (Rule + -> Class 1)			64
	EXCL (Rule + -> Class 1)			94
Balanced Sample Sizes				
	Rule +	**Rule -**		
Class 1	800	200	1000	
Class 2	50	950	1000	
	850	1150	2000	2000
	Support (Rule + -> Class 1)			40
	LSUP (Rule + -> Class 1)			80
	Confidence (Rule + -> Class 1)			94
	EXCL (Rule + -> Class 1)			94

Table 2.3 Confusion Matrix: Classification Accuracy Metrics

		Reference Standard			
		Pos	***Neg***		
Classification Algorithm	***Pos***	True Positive	False Positive	*Positive Predictive Value (PPV)*	*TP / TP+FP*
	Neg	False Negative	True Negative		
		Sensitivity	*Specificity*		
		TP / TP+FN	*TN / TN+FP*		

Table 2.4 Derivation of Sensitivity and Specificity Metrics

Case #	**Classification Result**	**Cumulative TP Count**	**Sensitivity**	**Cumulative FP Count**	**Specificity**
			0%		100%
1	TP	1	12.5		
2	FP			1	87.5
3	TN				
4	TP	2	25.0		
5	FN				
6	TP	3	37.5		
7	TN				
8	TN				
9	FP			2	75
10	TN				
11	TP	4	50.0		
12	TN				
13	TP	5	62.5		
14	FP			3	62.5
15	TP	6	75.0		
16	TN				
Reference Standard Positive	(TP + FN)	8	Sensitivity	Cum TP Count / 8	75.0%
Reference Standard Negative	(TN + FP)	8	Specificity	1 - (Cum FP Count / 8)	62.5%

References

1. Han J, Kamber M. Data mining: Concepts and techniques: Morgan Kaufmann Publishers; 2001.

2. Witten IH, Frank E. Data mining: Practical machine learning tools and techniques with java implementations: Morgan Kaufmann Publishers; 2000.

3. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery: An overview. Advances in knowledge discovery and data mining: AAAI/MIT Press; 1996.

4. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Magazine. 1996(Fall):37-54.

5. Mitchell TM. Machine learning. New York: McGraw-Hill; 1997.

6. Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med. 2002 Sep-Oct;26(1-2):1-24.

7. Shmueli G, Patel NR, Bruce PC. Data mining in excel: Lecture notes and cases. Arlington: Resampling Stats, Inc.; 2005.

8. Bayardo RJ. Brute-force mining of high-confidence classification rules. Proc 3rd Int Conf on Knowledge Discovery & Data Mining (KDD-97) AAAI Press; 1997. p. 123-6.

9. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. In: Intelligence AAfA, editor. KDD-98; 1998; New York; 1998.

10. Freitas AA. Understanding the crucial differences between classification and discovery of association rules: A position paper. SIGKDD Explor Newsl. 2000;2(1):65-9.

11. Thabtah F. A review of associative classification mining. Knowledge Engineering Review. 2007 Mar;22(1):37-65.

12. Agrawal R, Imielinski T, Swami A. Mining associations between sets of items in large databases. ACM SIGMOD Int'l Conf on Management of Data. Washington D.C.; 1993.

13. Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases: Morgan Kaufmann Publishers Inc.; 1994.

14. Liu B, Ma Y, Wong C. Classification using association rules: Weaknesses and enhancements. In: Vipin Kumar ea, editor. Data mining for scientific applications; 2001.

15. Mutter S, Hall M, Frank E. Using classification to evaluate the output of confidence-based association rule mining. Ai 2004: Advances in artificial intelligence; 2005. p. 538-49.

16. Rodda S, Shashi M. An improved associative classifier. Conference on Computational Intelligence and Multimedia Applications, 2007 International Conference on; 2007; 2007. p. 286-90.

17. Rutkowski L, Tadeusiewicz R, Zadeh L, et al. An efficient association rule mining algorithm for classification. Artificial intelligence and soft computing – icaisc 2008: Springer Berlin / Heidelberg; 2008. p. 717-28.

18. Thabtah FA, Cowling P, Yonghong P. Mmac: A new multi-class, multi-label associative classification approach. Data Mining, 2004 ICDM '04 Fourth IEEE International Conference on; 2004; 2004. p. 217-24.

19. Yin X, Han J. Cpar: Classification based on predictive association rules. Proceedings of the 2003 SIAM International Conference on Data Mining; 2003; San Francisco, CA; 2003.

20. Li W, Han J, Pei J. Cmar: Accurate and efficient classification based on multiple class-association rules. First IEEE International Conference on Data Mining (ICDM'01); 2001; 2001. p. 369-76.

21. About scopus. 2010 [cited Sept 21, 2010]; Available from: http://www.info.sciverse.com/scopus/about/

22. Pubmed. 2010 [cited; Available from: http://www.ncbi.nlm.nih.gov/pubmed/

23. Medical subject headings: Entry terms and other cross-references. 2009 Sept 01, 2009 [cited Oct 8, 2009]; Available from: http://www.nlm.nih.gov/mesh/intro_entry.html

24. Ohsaki M, Abe H, Tsumoto S, Yokoi H, Yamaguchi T. Evaluation of rule interestingness measures in medical knowledge discovery in databases. Artificial Intelligence in Medicine. 2007;41(3):177-96.

25. Silberschatz A, Tuzhilin A. What makes patterns interesting in knowledge discovery systems. Knowledge and Data Engineering, IEEE Transactions on. 1996;8(6):970-4.

26. Geng L, Hamilton HJ. Interestingness measures for data mining: A survey. ACM Comput Surv. 2006;38(3):9.

27. Zaïane OR, Antonie M-L. On pruning and tuning rules for associative classifiers. Knowledge-based intelligent information and engineering systems; 2005. p. 966-73.

28. Webb G, Brain D. Generality is predictive of prediction accuracy. Data mining: Springer Berlin / Heidelberg; 2006. p. 1-13.

29. Gu L, Li J, He H, Williams G, Hawkins S, Kelman C. Association rule discovery with unbalanced class distributions. Ai 2003: Advances in artificial intelligence; 2003. p. 221-32.

30. Liu B, Hsu W, Ma Y. Mining association rules with multiple minimum supports. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. San Diego, California, United States: ACM; 1999.

31. Tardanico R. Sample size and population size. [cited Sept 12, 2010]; Available from: http://www.fiu.edu/~tardanic/size.pdf

32. Provost FJ, Fawcett T, Kohavi R. The case against accuracy estimation for comparing induction algorithms. Proceedings of the Fifteenth International Conference on Machine Learning: Morgan Kaufmann Publishers Inc.; 1998.

33. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform. 2005 Oct;38(5):404-15.

34. Thabtah F. Pruning techniques in associative classification: Survey and comparison. Journal of Digital Information Management. 2006;4:202-5.

35. McAullay D, Williams G, Chen J, et al. A delivery framework for health data mining and analytics. Proceedings of the Twenty-eighth Australasian conference on Computer Science - Volume 38. Newcastle, Australia: Australian Computer Society, Inc.; 2005.

36. Ordonez C, Ezquerra N, Santana CA. Constraining and summarizing association rules in medical data. Knowl Inf Syst. 2006;9(3):259-83.

37. Thabtah F. Rule preference effect in associative classification mining. Journal of Information and Knowledge Management. 2006;5(1):13-20.

38. Thabtah F. Challenges and interesting research directions in associative classification. Sixth IEEE International Conference on Data Mining Workshops; 2006; 2006. p. 785-92.

39. Han J. Mining knowledge at multiple concept levels. Proceedings of the fourth international conference on Information and knowledge management. Baltimore, Maryland, United States: ACM; 1995.

40. Han J, Fu Y. Mining multiple-level association rules in large databases. Knowledge and Data Engineering, IEEE Transactions on. 1999;11(5):798-805.

41. Kuo Y-T, Lonie A, Sonenberg L, Paizis K. Domain ontology driven data mining: A medical case study. Proceedings of the 2007 international workshop on Domain driven data mining. San Jose, California: ACM; 2007.

42. Svátek V, Rauch J, Ralbovský M. Ontology-enhanced association mining. Semantics, web and mining: Springer Berlin / Heidelberg; 2006. p. 163-79.

43. Li J, Fu AW-c, Fahey P. Efficient discovery of risk patterns in medical data. Artificial Intelligence in Medicine. 2009;45(1):77-89.

44. Mahamaneerat WK, Kobayashi T, Green JM. Domain-concept mining on the 2005 nationwide inpatient sample data. American Medical Informatics Association (AMIA) 2007 Annual Symposium, Knowledge Discovery and Data Mining Working Group (KDDM-WG); 2007; 2007.

45. Elfangary L, Atteya WA. Mining medical databases using proposed incremental association rules algorithm (pia). Proceedings of the Second International Conference on Digital Society: IEEE Computer Society; 2008.

46. Wright A, Sittig DF. Automated development of order sets and corollary orders by data mining in an ambulatory computerized physician order entry system. AMIA Annu Symp Proc. 2006:819-23.

47. Tai Y-M, Chiu H-W. Comorbidity study of adhd: Applying association rule mining (arm) to national health insurance database of taiwan. International Journal of Medical Informatics. 2009;78(12):e75-e83.

48. Hersh WR. Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. Am J Manag Care. 2007 Jun;13(6 Part 1):277-8.

49. Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: An american medical informatics association white paper. J Am Med Inform Assoc. 2007 Jan-Feb;14(1):1-9.

CHAPTER 3

OVERVIEW OF THE COHORT AMPLIFICATION FRAMEWORK AND EVALUATION OF PREDICTION RULES GENERATED FOR DIABETES

Reprinted from Proceedings of the American Medical Informatics Association Annual Symposium Proceedings, 2010. Welch SR, Huff SM. Cohort amplification: An associative classification framework for identification of disease cohorts in the electronic health record. Pages 862-866, ©2010, with permission from the American Medical Informatics Association.

Cohort Amplification: An Associative Classification Framework for Identification of Disease Cohorts in the Electronic Health Record

Susan Rea Welch, MS, BSN, Stanley M. Huff, MD
University of Utah, Intermountain Healthcare, Salt Lake City, UT

Abstract

With the growing national dissemination of the electronic health record (EHR), there are expectations that algorithms to identify disease-based cohorts for health services research will be deployable across health care organizations. Toward that goal, a novel associative classification framework was designed to generate prediction rules to identify cases similar to the exemplar cases on which it was trained. It processes exemplars for any medical condition without modification. The framework is distinguished by core candidate data attributes based on common EHR observation categories, application of associative classification methods to cull disease-specific attributes and predictive rules from the core attributes, and support for attribute concept hierarchies to manage the various layers of granularity in native EHR data. The framework processes and an evaluation of prediction rules generated to identify diabetes mellitus are presented.

Introduction

Functionality to identify disease-based cohorts has been explicitly defined as an objective in the developing national standards for meaningful use of the EHR.[3] A framework (FW) was developed to generate prediction rules to identify research cohorts for various medical conditions using a generalized approach from coded EHR content. The general use case for the design was: (1) a set of exemplars for a given medical condition are identified, (2) a clinical profile (predictive rule set) is generated from the exemplars' EHR data using the FW, and (3) the rules are applied to the entire patient population in the EHR to identify additional patients with the specified condition. Since the objective was to identify new cases based on data patterns of known cases, it was called a 'cohort amplification' framework.

Background

Algorithms have been developed from EHR data, including standard diagnosis and procedure codes, to identify disease cohorts for research. A typical process was described by Starren:[30] (1) define a cohort by clinical characteristics, (2) translate to EHR data, (3) analyze the data, (4) identify subjects, (5) validate the algorithm, and (6) iterate. The FW may leverage the experts' time by providing information on EHR data content and distinguishing features among cohort exemplars early in the process: (1) define characteristics, (2) *identify exemplars of the cohort*, (3) *expose EHR data availability and predictive value for exemplars*, and then formulate an algorithm, validate and iterate.

Natural language processing (NLP) of free-text provider documentation, a rich source of information in the EHR, is an active and promising area of research for purposes of disease case identification [31]. The FW complements NLP efforts with domain knowledge and an opportunity to combine evidence.

The FW was developed from data in a large, integrated health care delivery organization with a mature enterprise-wide, longitudinal EHR. The Intermountain Healthcare Enterprise Data Warehouse provided the EHR data for secondary use that enabled development. Three diseases were selected for the focus of development: diabetes mellitus (DM), asthma, and clinical depression. These diseases are significant health problems and have established health care guidelines, which provided a source of domain knowledge for development and testing of the FW. IRB approval was granted for this research from both the University of Utah and Intermountain Healthcare. Study data contained no protected health information.

Methods

1. Description of the Cohort Amplification Framework

The cohort amplification FW was designed to meet the use case depicted above the dashed line in Figure 1. The processes and their implementation artifacts are depicted below the dashed line. Each process is described below:

EHR attribute selection

The basis of the candidate attribute categories was coded attributes defined by a national EHR certification organization as content requirements for ambulatory EHRs.[44] There were no technical limitations to adding disease or site- specific content, but the FW focus was standardized content for generalized application. Data used in the FW included diagnosis and procedure codes, provider and ambulatory clinic procedure codes, clinical lab tests performed, lab tests coded as abnormal, imaging procedures performed, medications in the EHR Medication List, and other demographic and encounter features. Exemplar patient attributes were populated once for each unique coded observation that occurred. Continuous observations were discretized as binary attributes: 'age > 47 = true.' There was no treatment for missing values. Attributes represented data that were populated in the EHR. An implicit attribute was the 'class' designation, which was generated from the exemplar lists, e.g., case or control. A class attribute is a fundamental of associative classification, serving as the consequent of all rules.

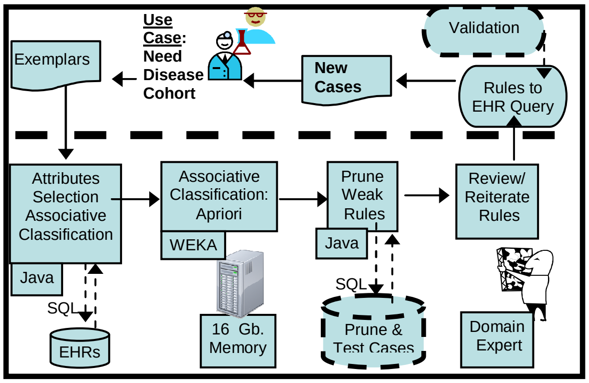

Figure 1 – Cohort Amplification Framework

Attributes from the EHR observations were mapped to concepts at higher levels of abstraction. The FW used a simple map of 'is-a' relations from an EHR attribute to a list of abstract concepts. When an attribute was instanced, all mapped concepts were also instanced. For example, URINE MICROALBUMIN and SERUM ALBUMIN were mapped to ALBUMIN. The FW considered all three as attributes to generate prediction rules from cross-level associations.[86]

The initial step in associative classification was applied to all EHR and derived attributes. A Java component developed for the FW was used. A data set containing attributes at a specified frequency of occurrence among disease cases was transferred to the data mining software.

Associative classification mining in the Waikato Environment for Knowledge Analysis

The next step was to use the Waikato Environment for Knowledge Analysis (WEKA) toolkit for association mining.[52] The Apriori algorithm constrained to the class attribute consequent was used.[47] Thabtah[60] described the steps in associative classification (AC) as discovery of associations among the training attributes, generation of rules associating other attributes with the class attribute, ranking and pruning rules to form a predictive rule set, and testing it on unseen data. WEKA Apriori was used for discovery and generation of the class association rules. The single attribute rules input from the EHR were selected according to standard AC interestingness metrics specialized for this application. In AC, the most common measures of interestingness are ‘support’ – the frequency of a rule in the data set – and ‘confidence’ – the likelihood of a particular class occurring, given the rule. These metrics only describe the training data from which they were generated.

In the FW, ‘local support’ for the disease class (LSUP), rather than support, was used. This is the frequency of the rule (a single attribute or a combination) among the disease case exemplars. For example, if 82% of DM exemplars have the attribute ‘Abnormal_HbA1c’, the LSUP is 82. A specialization of confidence was used: ‘exclusiveness.’ Exclusiveness of the disease class (EXCL) is its local support divided by the sum of both the disease and the control classes’ local support. For example, if the LSUP of ‘Abnormal_HbA1c’ among control exemplars is 0.6, then the EXCL is 82/(82+0.6) = 99 (expressed as a percentage).[76]

The WEKA Apriori association program was modified to use a variable threshold on the maximum number of attributes combined in a rule antecedent. Exponential combinations of attributes can accrue, if unbounded.[83] Antecedents with three or more attributes did not contribute to predictive accuracy in testing of configuration choices for the DM data presented. Constraining the number of attributes per rule also enabled the Apriori algorithm to run in the available processor memory (16 Gb.) at the desired minimum support threshold (2.5%).

Prune weak rules

The next step was to prune weak rules. Association rules must be pruned (generalized) for prediction purposes. The most concise set of rules without loss of accuracy is desirable.[59] Three pruning methods were implemented in Java components using the interesting attribute sets and metrics from WEKA. Redundant rule pruning[48] and database coverage pruning[49] were performed. An additional pruning method was developed to improve the generality of rules in this EHR data setting. The methods are not detailed in this overview. Functionally, configuration choices in the new pruning method were designed for specification on an application basis: the desired specificity threshold for the prediction rule set, the minimal number of pruning data set cases each rule must cover, and the minimum positive predictive value (PPV) of each rule on pruning set cases. PPV is the proportion of case coverage by the rule to the total coverage by the rule.

The new pruning method was also designed to manage multiple concept levels in the candidate rules. Higher order concept levels were preferred, given two concept levels for the same attribute. For example, a drug observation represented by an ingredient, ‘Insulin’, was mapped to a drug class, ‘Antihyperglycemic.’ If both presented as single attributes in the final pruning process, ‘Insulin’ was pruned.

The final pruned rule set was executed against a separate test database. The sensitivity and specificity of the prediction rule set on test data was calculated.

Domain knowledge

The final pruned rules should be examined for concordance with domain knowledge. Evaluation of the machine-generated knowledge and iteration of the process steps are cornerstones of knowledge discovery from data mining.[54]

2. Qualitative Evaluation of the Cohort Amplification Framework

Three parameters for successful prediction rules were defined for this evaluation: (1) accuracy on test data, (2) consistency with domain knowledge, and (3) conciseness and generality. Training and test data were sampled from the EHR data of adult patients who visited an Intermountain Medical Group (IMG) Family Practice (FP) or Internal Medicine (IM) clinic in Salt Lake County at least once in 2005-2006 and at least once in 2007-2008. DM case exemplars were a random sample of those with DM coded in the Problem List prior to 2007. Control exemplars (CTLs) were a random sample of those with no Problem List DM codes prior to 2009 and no ICD-9-CM DM codes assigned during 2004-2008. The sample sizes were 4,001 DM cases and 4,019 CTLs. The data mining timeframe was 2007-2008.

Four data sets for evaluation were generated in two random, stratified, two-fold cross-validation runs using the Knowledge Flow interface in WEKA. Each of the 8,020 patient records was randomly assigned in each of two runs. Each fold numbered ~2000 records for training data, ~1000 for pruning data, and ~1000 for test data. EHR attributes were selected if they had a frequency of at least 5% among cases: there were 472 qualified attributes for the study population. The main 3-digit ICD-9-CM code for DM (250) was not used.

The AC processes were configured as follows: The maximum number of attributes per rule antecedent was 2. The Apriori minimum support threshold was 0.025; minimum confidence 0.95. The specificity threshold was 98, chosen because no viable rules were ranked below 98 in the pruning data. Rules with pruning data set case counts < 3 and PPV < 80 were pruned. The prediction rules generated in the pruning step were evaluated for each of the four sets.

Results

The sensitivity, specificity and their 95% confidence intervals for each set and the average are shown in Table 1. For a common ICD-9-CM algorithm to identify DM (1 inpatient or 2 outpatient codes/2 years), sensitivity (average = 95.8%) was higher than those reported in two large studies: 72% and 80%.[26, 95] Specificity (average = 98.8%) was the same as those reported: 98% and 99%.

SAM-PLE	FOLD	SENS	Lo CI	Hi CI	SPEC	Lo CI	Hi CI
1	1	95.6	94.1	96.7	98.9	98.1	99.4
1	2	95.8	94.4	96.9	98.7	97.8	99.2
2	1	96.4	95.1	97.4	98.0	96.9	98.7
2	2	95.4	93.9	96.5	99.5	98.8	99.8
AVG		**95.8**			**98.8**		

Table 1 – Sensitivity and Specificity of Rule Sets with Confidence Intervals (CI)

The single attributes that participated in any of the four rule sets are shown in Table 2. These were contrasted with single attributes from three published projects that described the identification of DM cases from EHR data.[96-98] Other than ICD-9-CM codes, all three used elevated laboratory glucose assay results that were consistent with national diagnosis guidelines. One used elevated HbA1c results, and one used both HbA1c test orders and elevated results. One used all antihyperglycemic medications, one used all but metformin, and one used only three classes: metformin, insulins, and sulfonylureas (insulin release stimulants). In contrast, the FW did not select laboratory glucose assays but identified additional laboratory parameters that discriminated DM patients. Laboratory glucose assays were done in 94% of cases, but an abnormal value resulted in only 19% of cases, with exclusiveness of only 87%. The blood glucose measured by professional glucometers during office visits was more predictive. Tests for blood glucose, urine microalbumin and HbA1c were nearly as predictive as their respective abnormal results. Metformin, insulins and insulin release stimulants were found to be the strongest rules among antihyperglycemics as was noted by Wilke.[98] The novel strong rule 'Diabetic supplies (Pharm orders)' was formed by an aggregate concept over several pharmacy orderables including home glucometers, lancets, test strips, diabetic ulcer preparations, and other blood monitoring supplies. This rule identified additional DM patients beyond the lab test and medication rules in all sets.

The FW brings a new dimension to disease identification rules with combined attributes (Table 3). Metformin is known to generate false positives because it may be used in pre-diabetic conditions. However, it is the most common medication for DM. Table 3 shows metformin combined with several other attributes, as it did not form a strong enough rule alone. The combination with an HbA1c test and with a diagnosis or medication for dyslipidemia was consistent across rule sets. Both HbA1c tests and abnormal HbA1c results were strong single attributes, but they combined with other laboratory tests and attributes to form even stronger rules.

The conciseness and generality of the rules is shown in Table 3. Each set had 10 or 11 rules. Eight rules (bold font) were shared in 3 or 4 sets. Since each set started with ~2,000 rules after redundancy pruning, the generality of these machine-generated rules was encouraging. The rules can be further generalized based on the FW's exposure of patterns in the data, coupled with domain knowledge. For example, separate rule sets can be generated using only HbA1c tests or abnormal HbA1c, as these two single attributes cancelled each other out in the pruning and test data. Similarly, urine microalbumin, abnormal urine microalbumin and blood or urine albumin covered many of the same cases. The final step in the FW processes, refinement and iteration, was not implemented for this evaluation.

Category Attribute Description	**LSUP**	**EXCL**	**# OF SETS**
Dyslipidemia			
Hmg Coa Reductase Inhibitors	72	79	3
ICD9 272 Disord lipoid metab	82	71	2
Diabetic supplies & services			
Diabetic supplies (Pharm orders)	40	100	4
Antihyperglycemics			
Insulin Release Stimulant	46	100	4
Insulin Response Enhancer	38	99	2
Insulins	31	100	4
Metformin	64	98	4
Diabetes-related laboratory tests			
Glycosylated Hemoglobin (HbA1c)	94	96	4
ABN Glycosyl Hemoglobin (HbA1c)	82	99	4
Microalbumin, Urine	81	97	2
ABN Microalbumin, Urine	62	98	1
Creatinine, Urine	78	96	2
Glucose, Glucometer (Prof)	45	95	3
ABN Glucose, Glucometer (Prof)	42	96	1
Creatinine, Blood or Urine	81	90	4
Albumin, Blood or Urine	81	97	3
Demographic			
Age > 47	86	62	1

Table 2 - Single Attributes in All Rule Sets

	Sample 1 Fold 1	**Sample 1 Fold 2**	**Sample 2 Fold 1**	**Sample 2 Fold 2**
MICROALB_URINE & HbA1c	1		1	
ABN_HbA1c & Albumin	10	1		1
Metformin & HbA1c	2	2	2	2
ABN_HbA1c & Creatinine	3	5	6	
ABN_MICROALB_URINE & HbA1c		3		
Disord_lipoid_metabol & Metformin		11	3	
Insulin_Releas_Stimulators	4	4	4	3
Insulin_Resp_Enhancers		9		4
Insulins	5	7	10	6
Hmg_Coa_Reductases & ABN_HbA1c			5	
Metformin & Hmg_Coa_ Reductases	6		9	5
Metformin & Age_GT_47		6		
GLUC_GLUCOMETER & ABN_HbA1c	7	10	11	
DiabSupplies_Pharmacy	8	8	7	7
CREAT_URINE & HbA1c			8	
CREAT_URINE & ABN_HbA1c				8
Metformin & Albumin	9			
ABN_GLUC_GLUCOMETER & ABN_HbA1c				9
Metformin & Creatinine				10

Table 3 – Order of Rules in All Rules Sets

Discussion

The cohort amplification FW offers potential for an efficient generalized approach to derive cohort identification rules from EHR data. The strength of the approach is its ability to discover the patterns in the trail of data left by health care providers, who use an incalculable amount of professional knowledge to make diagnostic and treatment decisions. The framework indirectly taps into that knowledge. For diseases and conditions with less organized or shared care guidelines, the trail may not be as straightforward as it is for DM. On the other hand, patterns of care for DM are so consistent that one must discard the stronger, dominant attributes to expose potential novel associations. The merit of the rules is based upon many factors including the representativeness of the exemplars, the availability of candidate EHR data elements, the coverage of relevant evidence for a disease in the EHR (i.e., smoking history), and the strength of class association patterns found.

The highly correlated, sparsely populated EHR attributes are well suited to associative classification methods. There are known limitations in the AC methodology. No causal or inductive reasoning is used to form associations. There are many unexamined correlations among the attributes. Over-fitting rules to exemplars used for training can limit accuracy in the prediction task. Pruning weaker rules to gain a concise, general set of prediction rules helps to minimize this problem. A limitation with the reliability and generalizability of prediction rules is that the interestingness metrics, on which rule generation is based, may vary by health care setting, data quality, and choice of exemplars.

The FW was designed to be modular and extensible. There are many potential improvements to the processes and algorithms in the FW. These include linkage to standardized terminologies and concept hierarchies, support for sequential patterns, and evaluation of other methods for more efficient association, rule discovery and pruning. Further research on portability across organizations would inform our assumption of standardized EHR content. Further study of the reliability and accuracy of the cohort amplification framework applied to asthma is in progress.

Conclusion

The cohort amplification framework processes and rules generated for identification of DM were presented. Evaluation results were successful.

References

1. Blumenthal D, Tavenner M. The "Meaningful use" Regulation for electronic health records. N Engl J Med. 2010 Jul 13.

2. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: A review of recent research. Yearb Med Inform. 2008:128-44.

3. Certification commission for health information technology. 2010 [cited Jan 18, 2010]; [ambulatory EHR certification criteria].Available from: http://www.cchit.org.

4. Han J. Mining knowledge at multiple concept levels. Proceedings of the fourth international conference on Information and knowledge management. Baltimore, Maryland, United States: ACM; 1995.

5. Witten IH, Frank E. Data mining: Practical machine learning tools and techniques with java implementations: Morgan Kaufmann Publishers; 2000.

6. Agrawal R, Imielinski T, Swami A. Mining associations between sets of items in large databases. ACM SIGMOD Int'l Conf on Management of Data. Washington D.C.; 1993.

7. Thabtah F. A review of associative classification mining. Knowledge Engineering Review. 2007 Mar;22(1):37-65.

8. Gu L, Li J, He H, Williams G, Hawkins S, Kelman C. Association rule discovery with unbalanced class distributions. Ai 2003: Advances in artificial intelligence; 2003. p. 221-32.

9. Ordonez C, Ezquerra N, Santana CA. Constraining and summarizing association rules in medical data. Knowl Inf Syst. 2006;9(3):259-83.

10. Freitas AA. Understanding the crucial differences between classification and discovery of association rules: A position paper. SIGKDD Explor Newsl. 2000;2(1):65-9.

11. Li W, Han J, Pei J. Cmar: Accurate and efficient classification based on multiple class-association rules. First IEEE International Conference on Data Mining (ICDM'01); 2001. p. 369-76.

12. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. American Association for Artificial Intelligence. KDD-98; New York; 1998.

13. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Magazine. 1996(Fall):37-54.

14. Lix L, Yogendran M, Burchill C, et al. Defining and validating chronic diseases: An administative data approach. Winnipeg: Manitoba Centre for Health Policy; July 2006.

15. Miller DR, Safford MM, Pogach LM. Who has diabetes? Best estimates of diabetes prevalence in the department of veterans affairs based on computerized patient data. Diabetes Care. 2004 May;27 Suppl 2:B10-21.

16. Flood G, Pyenson BS, Rosenberg M. Health outcomes: The use of electronic medical records in the id and management of patients with diabetes. Society of Actuaries 2009 Health Spring Meeting. Toronto, Ontario: Society of Actuaries; 2009.

17. Toh MP, Leong HS, Lim BK. Development of a diabetes registry to improve quality of care in the national healthcare group in singapore. Ann Acad Med Singapore. 2009 Jun;38(6):546-6.

18. Wilke RA, Berg RL, Peissig P, et al. Use of an electronic medical record for the identification of research subjects with diabetes mellitus. Clin Med Res. 2007 Mar;5(1):1-7.

CHAPTER 4

RANKING AND PRUNING METHODS DEVELOPED FOR THE COHORT AMPLIFICATION FRAMEWORK

Introduction

In this chapter, the process for generating associative classification rules for the identification of asthma cases is presented. The best set of rules was selected on their generality and accuracy across ten random training/testing samples. Novel ranking and pruning methods were used to generate the classification rules. The ranking and pruning methods will be described and results compared to a standard method of ranking and pruning association rules, known as classification by association.[1] Methods for evaluating generality and accuracy are described. The best rules were used in a study of the accuracy of identification of asthma cases in a random sample of the EHR, which is reported in Chapter 5.

Background

Ranking and pruning are critical methods to generate an associative classifier. In association rule mining, a deterministic and deductive task has been performed. An exhaustive set of all associations in the training data, given the constraints, are generated. Classification rule discovery is an inductive task, predictive of associations in data not yet seen. Rigorous ranking and pruning strategies are used to select a concise, general and accurate predictive subset of association rules from the exhaustive set. General and concise rules go hand-in-hand because general rules, by definition, cover more cases. Therefore, fewer rules are needed. Both accurate and general rules are preferred in most ranking and pruning strategies. More accurate and more general rules help avoid overfitting to the training data. Rules that are more accurate are more highly associated and less likely to be artifacts. Rules that are more general are more representative of the theoretic target population represented by the training data.

In associative classification, the goal is to take the association rules and form a classifier. To develop the associative classifier, first association rule mining is used to discover the rules with the class as the consequent, and then ranking and pruning is performed over those rules to form the subset of classifying rules. Further, their predictive accuracy must be evaluated on test data.[2] A benefit of associative classification is that the ranking and pruning processes and interim results are understandable. The entire process of developing an associative classifier is transparent. This enables understanding of the knowledge generated, development of domain specific ranking and pruning strategies, and the ability to revise the input data or configuration parameters in order to improve the next version of the classifier.

Associative classifiers were also proven in some studies to be as accurate as the classic classification methods. Chapter 2 contains an extensive literature review on associative classification. Of particular relevance to the pruning strategies described and compared in this chapter, Chapter 2 details the classic pruning methods.

Methods

Rules Development

The study population was described in Chapter 1. In summary, the population sampled for all studies reported in this dissertation was adult patients who had at least two health care visits to Intermountain Medical Group family practice or internal medicine ambulatory clinics in Salt Lake County at least twice during a four year period, 2005-2008. The electronic health records (EHR) of this study population, as stored in the Intermountain Health Care Enterprise Data Warehouse, were the source for all data. The EHR covers all health care given by Intermountain employed providers as well as some documentation of health care visits to providers affiliated with Intermountain.

During development of the cohort amplification framework, the number of disease-positive training cases required to generate reliable rules was analyzed. Training set sizes of approximately 1,000 showed too much variability, while sizes of approximately 2,000 were as consistent as sizes of 4,000. In order to generate disease exemplar training sets of size ~2,000 with ~1,000 each for pruning and testing of the association rules, 3,938 subjects with active asthma coded in the Problem List prior to the year 2007 were selected as the disease exemplars. The asthma codes were developed by and for the clinicians using the Problem List. An approximately equal sized control group of 3,948 subjects was randomly selected among the target population having no codes for asthma in the Problem List or the hospital or ambulatory encounter records during the years 2004-2008. For the intended purpose of classification of asthma cases versus no asthma in the EHR data, the control group was defined broadly in order to represent all adult patient records without evidence of asthma. The training sample sizes are balanced so that differences between them are normalized for rules discovery.

A description of the core candidate electronic health record data used in the cohort amplification framework was presented in Chapter 1. Direct coded observations, aggregate concepts and derived concepts were used from the following EHR data sources:

- Diagnosis and procedure codes (ICD-9-CM codes)
- Provider and ambulatory clinic procedure codes (CPT codes) [3]
- Provider specialty (local codes)
- Lab observations (CPT codes)
- Lab observations with results coded as ‘Abnormal’
- Imaging procedures (CPT codes)
- Medication list (FirstDataBank pharm/chemical groups and ingredients) [4]
- Age > 64 (true)
- Female gender (true)

Single attributes were selected using Apriori association mining methods for frequent one-item sets. This was accomplished in a Java class designed to accept a list of observation codes for each exemplar. The output was a WEKA file containing the attributes that met a frequency threshold of 5% among diseases exemplars. In other words, attributes that were populated in less than 5% of the disease exemplars were pruned. A distinction that runs through the entire process of rules generation in the cohort amplification framework is a focus solely on rules that predict the disease class versus a comparison class. There were 414 one-item attributes selected for association mining in WEKA and then further ranking, pruning, and testing to form an associative classifier.

The process of one-item attribute selection using the Apriori algorithm from the EHR required manual curation of the attributes. The EHR query swept all observation codes within the categories listed. Some observation codes were not germane to the rules, such as Healthcare Common Procedure Coding System (HCPCS) codes that are stored with the CPT codes. This process was run twice, on both halves of a randomized split of the entire data set with a balance of disease and control exemplars. The differences in the frequencies of the 414 one-item attributes were compared by a Paired t test. The differences were not statistically significant (2 tailed p value = .71, 95% confidence interval of the difference -0.08 to 0.12). At sample sizes of ~2,000, the frequencies of one-item attributes discovered in the EHR data among asthma exemplars were stable. These two transfers from the EHR to a format suitable for data mining were used for the remainder of the rules discovery processes. Randomly selected examples from the 414 one-item attributes are shown in Table 4.1.

The 3,938 cases and 3,948 controls were randomly sampled ten times into training sets of 1,969 asthma exemplars and 1,974 control exemplars. The Waikato Environment for Knowledge Analysis (WEKA) Knowledge Flow user interface[5, 6] was used to randomly split the subjects into two sets, stratified on the exemplar status. One, the training set, was used to generate the association rules using the Apriori algorithm[7] in WEKA version 3.6.2 on a computer with 16 Gb of memory. The other set was for pruning and testing and was stored in a local SQL Server Express database by a Knowledge Flow component. Both sets had an equal number of cases and controls. The work flow is shown in Figure 4.1. This process was repeated ten times to generate ten random training/pruning/testing samples.

The first five samples were analyzed for the maximum number of attributes to combine in the antecedents of the association rules. Since the Apriori algorithm is exponential in computing complexity with respect to the number of combinations performed, the lowest maximum number of attributes which achieves accurate rules is best.[8] Redundant rules are not formed, and lower minimum support thresholds can be reached by the Apriori algorithm. The five samples were consistently more accurate, by sensitivity and specificity on test data, when combinations were constrained at three attributes versus two. This computing complexity is estimated as $(414)^3$. On the other hand, the best rules by rank, sensitivity gain, specificity loss, and agreement across samples had one or two attribute rule antecedents. In addition, rules having one or two attribute antecedents are more general rules, and therefore, preferred. Four-attribute combinations are of estimated computing complexity $(414)^4$, which is difficult to accommodate in computer memory for WEKA Apriori association mining. To accomplish it, the frequency threshold for the algorithm must be set higher. The trade-off in frequency to gain four- attribute antecedents with low rank and low generality was not attempted. The lowest frequency among disease exemplars that was reached with the three attribute maximum was 8%. A comparison with rules generated at a two attribute maximum, which reached 5% frequency among disease exemplars, showed that no rules were missed by losing attributes with only 5-7% frequency among disease exemplars.

Five more samples with three attribute association rules were generated. The ten sets of random samples of the exemplar data, each with candidate rules generated with a maximum of three attribute rule antecedents in WEKA, were then used to create associative classification rules using ranking and pruning methods developed for the cohort amplification framework. The methods are shown to perform better than classic ranking and pruning methods for this application. One best set of rules to identify asthma cases in the EHR was selected from the rules generated in the ten random samples.

Ranking and Pruning

The Apriori algorithm implemented in WEKA was used to perform the association rule mining step, constrained to a maximum of three attributes per rule antecedent and constrained to the exemplar status or ‘class’ as the rule consequent. The minimum frequency or support was set at 8% for the disease class (4% overall). WEKA provided an output file with all frequent-item sets with their total count and their class count. In the framework, three downstream associative classification pruning steps were implemented using the output. Ranking and pruning were performed on separate training data from the set used to create the rules.

Redundant Rule Pruning

Classification based on multiple association rules (CMAR)[9], a classic associative classification model, introduced the redundancy pruning method. Redundancy pruning is implemented before the rules are ranked. Multiattribute rules that cover the same or fewer cases and do not improve the confidence of a more general rule, e.g., a subset of the multi-attribute rule, are redundant. If rule ‘R_a’ was met, then rule ‘R_a U R_b’ is redundant unless its confidence is greater. Rule ‘R_a U R_b’ (child rule) cannot have more support than ‘R_a’ (parent rule). At most, it can have equal support. The redundancy pruning method results in fewer, more general rules.

In the framework, a Java class parsed the WEKA output file and created a directed acyclic graph of the single attribute relations to their combinations and calculated and stored the local support and exclusiveness metrics for each. Local support is the frequency of occurrence in the disease class, and exclusiveness is the normalized likelihood of the disease class occurring as the rule’s consequent.[10] The exclusiveness metric is equal to the confidence metric for equal training sample sizes. Although the exclusiveness metric is not representative of the population to which the rules will be applied, its relative values are consistent with the confidence metric and valid for ranking purposes. The Java class pruned redundant rules and generated an output file with the remaining rules and their metrics for the next pruning steps.

Database Coverage Pruning

The initial associative classification model, known as ‘classification by association’ (CBA), introduced the database coverage method.[1] More than one rule may cover the same case to be classified. The rules discovered by association rule mining are ranked in order of confidence, then support. Processing the rules in ranked order, the training cases that meet the rule are removed from further consideration by a subsequent rule. If no case meets a rule, the rule is pruned. This continues until all training cases have been covered or all rules were tested. Cases left uncovered are assigned to the class with the highest frequency in the training data. The database coverage method seeks the most accurate rules, by rank, and eliminates less accurate rules that cover the same cases.

Database coverage was implemented in the framework on the set of rules pruned for redundancy. In the EHR data used in the framework, as is true in transaction data in general, there are naturally co-occurring attributes in the data because they describe the same processes. This was one of the main assumptions of the original association rule mining approach, but it leaves a conundrum for associative classification rules. Small differences in rank can generate different sets of rules over samples of the same training data. This is because selected rules in rank order will cover others that follow, which will be pruned if all their cases were previously covered. The varying rules might predict the class equally well on the test data, but inconsistent rule sets are not appreciated by users. Otherwise, the associative classification rules are quite amenable to domain expert understanding. Since one of the goals of the framework is to generate understandable as well as concise, general and accurate classifiers, a new ranking algorithm was developed to overcome this inconsistency.

Framework Ranking and Pruning Improvements

The ranking was based on the principle of generality and also requires specification of a lower bound on the acceptable specificity for the set of rules. Sensitivity and specificity are trade-offs for accuracy of the rules. The threshold could be placed on either. However, it made more sense to specify the tolerance for incorrect classification, depending upon the intended application of the rules. An abbreviated example is shown in Table 4.2. With the rules ranked according to CBA, the specificity is calculated on a pruning data set. The rules at or above the specificity threshold are reranked according to their absolute frequency (descending) among disease exemplars in the pruning data set. Within absolute frequency, rules are ranked by exclusiveness (descending), then cardinality (ascending). The CBA database coverage pruning algorithm is executed on this reranked subset of rules, with further constraints to avoid over-fitting. Rules that covered less than five disease exemplars were pruned, rather than the CBA algorithm's acceptance of a rule that classified even one case. Rules with a positive predictive value (PPV) less than 70 on the pruning data set were pruned.

The reranking by absolute frequency among disease exemplars promotes generality of the rules. The higher ranking of more frequent attributes also results in attributes of more general concept levels covering all cases of any child concept levels. The duplicative and less general attribute will be pruned from the rules. Since the rules most likely to cover disease exemplars are executed first, it was hypothesized that fewer, more frequent rules would be generated. Pruning lower ranked rules was hypothesized to reduce misclassifications, resulting in better accuracy. The requirement for a rule to cover at least five disease exemplars also supports generality. The PPV constraint supports accuracy. Generality will be measured by conciseness of the rule sets based on the number of rules. Accuracy will be measured by the specificity of the rule sets.

Rules that survive pruning become the rule sets for each of the ten training sets. The sensitivity and specificity of the rules sets were assessed on test data set aside for each of the ten training sets. The pruning methods developed for the framework were compared to the CBA methods on conciseness and sensitivity. Conciseness is defined as the total count of rules in the set. The framework pruning method was based on a threshold set on the specificity derived by CBA pruning methods and was the point at which the two processes deviated. The specificities after the framework pruning process were usually improved above the thresholds, even though the subsequent specificity was measured on separate test data. Specificity and sensitivity are bound, moving in opposite directions. For comparison of accuracy, the framework rule set sensitivity will be compared to the CBA sensitivity at the same specificity.

Selection of the Best Rules

The best rules were selected from the ten training sets to identify asthma cases in the EHR based on their consensus across samples (generality) and their accuracy. It was preferable to take a general set of rules from all ten training sets than to take the rules from one set, which performed well on the test data.   The ratio of the average sensitivity gain to the average specificity loss of the rule was calculated on test data. The order of each rule in its rule set influences the actual contribution to sensitivity and specificity. Higher ordered rules cover more cases and leave fewer for subsequent rules to classify. There may be further bias in the uncovered cases. The stronger associations execute first and the weaker lower ordered rules get the left-over cases. The ratio of average sensitivity gain to average specificity loss is independent of the order and is an index of correct prediction to misclassification for each rule on the test data.

Results

The five most accurate training sets on test data are shown in Table 4.3. These had both the best sensitivity and the best specificity. The remaining five sets are shown in Table 4.4. For each ordered rule in each training set, the accruing sensitivity and specificity loss (the false positive accrual as a percentage of the true negatives) are shown as well as the contribution by each rule.

Table 4.5 shows the generality and accuracy measures for the ten training sets. The number of rules and sensitivity are shown for the rule sets pruned by framework methods and pruned by CBA. The specificity attained by the framework methods is shown. The sensitivities for both methods are reported at this specificity. Conciseness of the rules was dramatically improved with the framework pruning methods. At a constant specificity, the sensitivity in the framework pruning methods was higher in seven training sets, equal in one, and slightly lower than the CBA pruning method sensitivity. The average sensitivity using framework pruning was 58.2%. The average for CBA pruning was 56.4%. The difference in the sensitivity over the ten training sets was statistically significant (p = .028, 95% confidence interval 0.24-3.3) by a paired t test. The framework pruning methods resulted in a modest improvement in accuracy.

The ten sets generated a fairly consistent sets of rules (Table 4.6). The rules shown in italicized font were covered by a more general rule in the collection. Usually these were three-attribute rules that were covered by a two-attribute rule. However, ‘Asthma Procedures’ (CPT 94010, BREATHING CAPACITY TEST; CPT 94640, AIRWAY INHALATION TREATMENT) was a conceptual subset of ‘Other Pulmonary Procedures’ (CPT 94010; CPT 94640, CPT 94240, RESIDUAL LUNG CAPACITY; CPT 94060, EVALUATION OF WHEEZING,BRONCHODIL RESPN PRE&POST DILAT; CPT 94720, MONOXIDE DIFFUSING CAPACITY). No general-specific pair was generated in the same rule set since the ordering by absolute frequency forced the more general concept first. The more general rule of a pair was preferred for the best set.

Rules were selected for the best set if the average sensitivity gain to average specificity loss ratio was 4 or greater because the most consistent rules, those that occurred in at least four rule sets, had a ratio greater than 4. This ratio corresponds to an

80% likelihood that the rule, given its order, improved the overall accuracy of classification on test data.

The best rules are shown in Table 4.7. They will be used in Chapter 5 to classify asthma cases in a random sample from the study population.

Discussion and Conclusions

The processes used to generate a set of associative classification rules to identify asthma cases in the electronic health record were described. An overview of the cohort amplification framework processes and workflow was described in Chapter 3. In this chapter, the focus was the ranking and pruning processes developed for the framework in order to gain more general rules than were generated using the classic CBA methods. Compared to CBA, the framework ranking and pruning strategies improved both generality and accuracy of the rules on test data.

Improvements in rule ranking was listed as one of the interesting research directions in associative classification.[11] Rule preference affects the accuracy of the classifier. Two novel ranking approaches were introduced in this study. The calculation of the sensitivity and specificity of the CBA-ranked rules on training data presented a metric that is familiar in the application domain. Secondly, rules at or above a lower bound on the acceptable specificity, a parameter setting for the application at hand, were re-ranked according to the absolute frequency at which the rule was satisfied in the pruning data set. This forced the most general rules to execute first. This not only solved the problem of generating rules that are more consistent across training samples, it also automated the selection of the most general concept if multiple concept hierarchies were present. The rules ‘Asthma Procedures’ and ‘Other Pulmonary Procedures’, discussed in the Results section, are examples of the latter.

More conservative pruning strategies than those in CBA were used. Pruning may have a domain-specific rationale. The two objectives are to remove redundant and misleading rules for the classification task at hand. Redundant rule pruning was used as described in the classic CMAR method.   Since the candidate EHR data represented routine processes of health care, the data were inherently highly associated. To avoid over-fitting in this domain, higher thresholds were set on the number of training cases covered (generality) and the positive predictive value (accuracy) of each rule. As a further step to avoid over-fitting, a separate slice of the training data was used to perform the ranking and pruning steps than the training data used to generate the rules.

Table 4.1 Random Examples of 414 EHR Candidate Attributes Frequencies Among Asthma Exemplars

Attribute	Frequency (%)	
	Sample 1	Sample 2
c82947_GLUC_BLD_LAB__QUANT	79	78
c84520__UREA_NITROGEN_ASSA	78	77
c84295__SERUM_SODIUM_ASSAY	78	77
c84075_ASSAY_ALKALINE_PHOS	71	70
c82247_BILIRUBIN__TOTAL_BI	71	70
isFemale_NO_DESCRIPTION__	67	66
c85025_COMPLETE_CBC_W_AUTO	67	65
c80061_LIPID_PANEL_LIPID_P	61	59
GT_5_FF_Vis_Per_Yr_NO_DES	59	57
age_GT_47_NO_DESCRIPTION_	59	60
c80061A__ABN_LIPID_PANEL__	58	57
c84443_ASSAY_THYROID_STIM_	57	56
c85025A__ABN_COMPLETE_CBC_	54	54
Urinalysis__by_dip_stick__L186	53	52
c3000250272_Analgesics__Narcotic	47	46
c272_Disord_lipoid_metabol	46	46
c401_Essential_hypertensio	45	46
c3000253044_Fluticasone__H	44	45
c90658_FLU_VACCINE__3_YRS_	39	40
c3000250264_Antidepressant	38	37
c3000252433_Salmeterol__Hi	32	33
c3000508986_Proton_Pump_In	31	31
c461_Acute_sinusitis__LSup	31	28

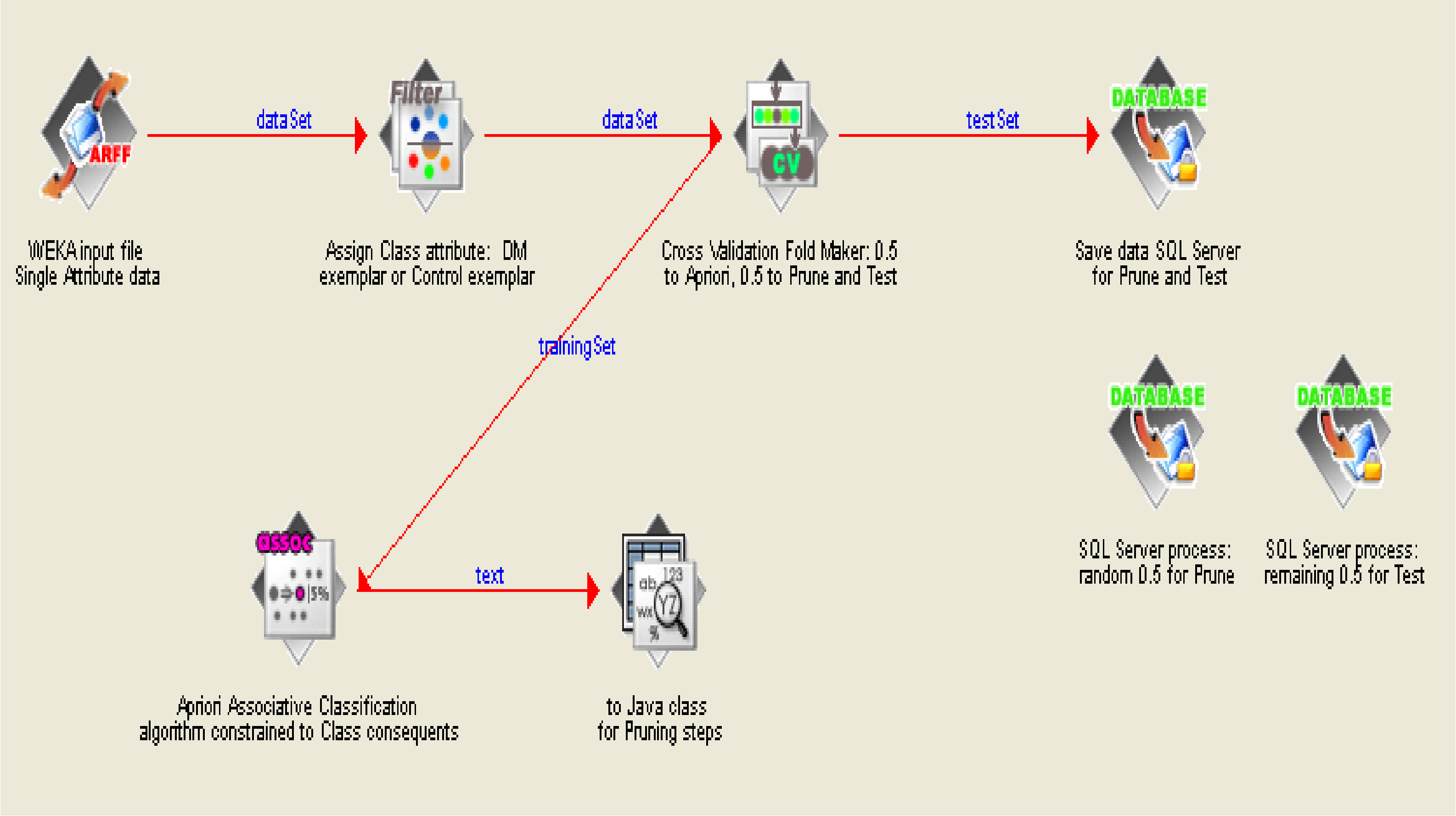

Figure 4.1 Data Flow and Components to Generate Associative Classification Rules

Table 4.2 Example of Specificity Based Pruning (control sample size = 987)

Rule	**CBA Rank**	**Control Count**	**Cumulative Control Count**	**Specificity**	**Absolute Frequency Disease Exemplars**
ALLERGY_SRVC AND Glucocorticoid AND Other_Pulmonary_Procedure	1	3	3	99.7	71
Salmeterol AND Diagnostic_Radiology	2	3	6	99.4	81
Glucocorticoid AND Albuterol AND Allergic_rhinitis	3	2	8	99.2	70
Leukotriene_Receptor AD Albuterol	4	0	8	99.2	85
Albuterol AND Montelukast	5	0	8	99.2	83
Salmeterol AND Need_for_prophylactic_vac	6	4	12	98.8	132
Salmeterol AND FLU_VACCINE	7	1	13	98.7	137
Albuterol AND Salmeterol	8	0	13	98.7	126
Antihistamines AND Salmeterol	9	1	14	98.6	123
Salmeterol AND IMMUNZATN_ADMIN	10	0	14	98.6	122
Salmeterol AND Hmg_Coa_Reducta	12	0	14	98.6	94
Salmeterol AND Oth_and_unspecified	13	0	14	98.6	109
Albuterol AND Fluticasone AND isFemale	14	0	14	98.6	118
Albuterol AND ABN_LIPID_PANEL AND isFemale	15	2	16	98.4	106
Leukotriene AND age_GT_47 AND isFemale	16	1	17	98.3	99

Table 4.3 Five Most Accurate Rule Sets of the Ten Training Sets

Set	Rule	Rule Order	Sensitivity Accrual	Sens. Gain This Rule	Specificity Accrual	Spec. Loss This Rule
1	c3000252433_Salmeterol	1	32.5	32.5	99.0	1.0
	c3000250386_Glucocorticoid AND c3000252425_Albuterol	2	44.2	11.7	98.3	0.7
	c3000250652_Leukotriene_Rec_Antag	3	50.8	6.6	98.0	0.3
	c3000250386_Glucocorticoid AND AsthmaProcedures_cpt94010	4	54.2	3.4	97.9	0.1
	c3000252425_Albuterol AND c82947_GLUC_BLD_LAB__QUANT	5	57.0	2.8	97.3	0.6
	BetaAdrenergHic3NotAlbutO_L32	6	58.7	1.7	97.3	0.0
2	c3000252433_Salmeterol	1	31.9	31.9	99.3	0.7
	c3000250386_Glucocorticoid AND c3000252425_Albuterol	2	44.1	12.2	98.7	0.6
	c3000250386_Glucocorticoid AND AsthmaProcedures_cpt94010	3	48.3	4.2	98.3	0.4
	c3000250652_Leukotriene_Rec_Antag	4	53.6	5.3	97.9	0.4
	BetaAdrenergHic3NotAlbutO_L32	5	55.7	2.1	97.9	0.0
	c3000252425_Albuterol AND c780_General_symptoms	6	57.7	2.0	97.2	0.7
	c3000250386_Glucocorticoid AND Other_Pulmonary_Procedure AND isFemale	7	58.0	0.3	96.9	0.3
3	c3000252433_Salmeterol	1	33.9	33.9	99.3	0.7
	c3000250386_Glucocorticoid AND c3000252425_Albuterol	2	45.0	11.1	98.5	0.8
	c3000250652_Leukotriene_Rec_Antag	3	52.0	7.0	97.9	0.6
	c3000250386_Glucocorticoid AND AsthmaProcedures_cpt94010	4	54.8	2.8	97.4	0.5
	c3000252425_Albuterol AND isFemale	5	57.2	2.4	97.0	0.4
	BetaAdrenergHic3NotAlbutO_L32	6	58.9	1.7	96.9	0.1
	c82947_GLUC_BLD_LAB__QUANT AND AsthmaProcedures_cpt94010 AND asthmaComorbids_473	7	59.1	0.2	96.8	0.1
	c3000252425_Albuterol AND Urinalysis__by_dip_stick	8	59.8	0.7	96.8	0.0
4	c3000252433_Salmeterol	1	33.3	33.3	98.7	1.3
	c3000250386_Glucocorticoid AND c3000252425_Albuterol	2	44.5	11.2	98.0	0.7
	c3000250652_Leukotriene_Rec_Antag	3	49.9	5.4	97.6	0.4
	c3000250386_Glucocorticoid AND AsthmaProcedures_cpt94010	4	53.1	3.2	97.1	0.5
	c3000252425_Albuterol AND isFemale	5	55.2	2.1	96.4	0.7
	BetaAdrenergHic3NotAlbutO_L32	6	57.1	1.9	96.4	0.0
	c3000252425_Albuterol AND c85025A__ABN_COMPLETE_CBC	7	57.4	0.3	96.4	0.0
	c84443_ASSAY_THYROID_STIM AND AsthmaProcedures_cpt94010	8	58.7	1.3	95.8	0.6
5	c3000252433_Salmeterol	1	31.0	31.0	98.9	1.1
	c3000250386_Glucocorticoid AND c3000252425_Albuterol	2	43.7	12.7	98.1	0.8
	c3000250652_Leukotriene_Rec_Antag	3	49.0	5.3	97.5	0.6
	c3000250386_Glucocorticoid AND Other_Pulmonary_Procedure	4	53.4	4.4	96.9	0.6
	c3000252425_Albuterol AND Urinalysis__by_dip_stick	5	55.4	2.0	96.4	0.5
	BetaAdrenergHic3NotAlbutO_L32	6	57.6	2.2	96.3	0.1
	AsthmaProcedures_cpt94010 AND V04_Need_for_prophylactic_vaccine	7	58.1	0.5	95.8	0.5

Table 4.4 Five Least Accurate Rule Sets of the Ten Training Sets

Set	Rule	Rule Order	Sensitivity Accrual	Sens. Gain This Rule	Specificity Accrual	Spec. Loss This Rule
6	c3000252433_Salmeterol	1	31.0	31.0	99.1	0.9
	c3000250386_Glucocorticoid AND c3000252425_Albuterol	2	42.7	11.7	98.2	0.9
	c3000250652_Leukotriene_Rec_Antag	3	49.1	6.4	98.0	0.2
	c3000250386_Glucocorticoid AND Other_Pulmonary_Procedure	4	52.5	3.4	97.3	0.7
	c3000252425_Albuterol AND isFemale	5	54.9	2.4	96.7	0.6
	BetaAdrenergHic3NotAlbutO_L32	6	57.1	2.2	96.7	0.0
	c3000252425_Albuterol AND c82947_GLUC_BLD_LAB__QUANT AND GT_5_FF_Vis_Per_Yr	7	57.5	0.4	96.5	0.2
7	c3000252433_Salmeterol	1	32.1	32.1	99.0	1.0
	c3000250386_Glucocorticoid AND c3000252425_Albuterol	2	43.5	11.4	98.5	0.5
	c3000250652_Leukotriene_Rec_Antag	3	49.5	6.0	97.9	0.6
	c3000250386_Glucocorticoid AND c82947_GLUC_BLD_LAB__QUANT AND Other_Pulmonary_Procedure	4	52.6	3.1	97.3	0.6
	c3000252425_Albuterol AND c82947_GLUC_BLD_LAB__QUANT AND	5	53.9	1.3	97.1	0.2
	BetaAdrenergHic3NotAlbutO_L32	6	56.0	2.1	97.0	0.1
	c3000253044_Fluticasone AND c782_Symptoms_involving_skin AND GT_5_FF_Vis_Per_Yr	7	56.5	0.5	96.5	0.5
	c2_ALLERGY SERVICE AND Other_Pulmonary_Procedure	8	57.1	0.6	96.4	0.1
8	c3000252433_Salmeterol	1	31.4	31.4	99.2	0.8
	c3000250386_Glucocorticoid AND c3000252425_Albuterol	2	43.5	12.1	97.9	1.3
	c3000250386_Glucocorticoid AND Other_Pulmonary_Procedure	3	48.0	4.5	97.0	0.9
	c3000250652_Leukotriene_Rec_Antag	4	54.1	6.1	96.4	0.6
	c3000252425_Albuterol AND isFemale	5	56.7	2.6	95.7	0.7
	c3000252425_Albuterol AND GT_5_FF_Vis_Per_Yr	6	57.9	1.2	95.4	0.3
	BetaAdrenergHic3NotAlbutO_L32	7	58.3	0.4	95.4	0.0
9	c3000252433_Salmeterol	1	30.8	30.8	99.0	1.0
	c3000250386_Glucocorticoid AND c3000252425_Albuterol	2	42.2	11.4	97.9	1.1
	c3000250652_Leukotriene_Rec_Antag	3	49.6	7.4	97.2	0.7
	c3000252425_Albuterol AND c82947_GLUC_BLD_LAB__QUANT	4	51.4	1.8	96.5	0.7
	c3000250386_Glucocorticoid AND c82947_GLUC_BLD_LAB__QUANT Other_Pulmonary_Procedure_L26	5	55.8	4.4	95.5	1.0
	BetaAdrenergHic3NotAlbutO_L32	6	58.3	2.5	95.4	0.1
10	c3000252433_Salmeterol	1	31.1	31.1	98.5	1.5
	c3000250386_Glucocorticoid AND c3000252425_Albuterol	2	42.8	11.7	97.6	0.9
	c3000250652_Leukotriene_Rec_Antag	3	49.7	6.9	96.7	0.9
	c3000250386_Glucocorticoid AND Other_Pulmonary_Procedure	4	53.3	3.6	95.9	0.8
	SerumElectrolytes AND AsthmaProcedures_cpt94010	5	54.3	1.0	94.6	1.3
	c3000252425_Albuterol AND c85025A__ABN_COMPLETE_CBC	6	55.8	1.5	94.4	0.2
	BetaAdrenergHic3NotAlbutO_L32	7	57.6	1.8	94.4	0.0

Table 4.5 Generality and Accuracy of Rules
Sensitivity Compared at Framework Specificity

	Framework Pruning			**CBA Pruning**	
Set	**Number of Rules**	**Speci-ficity**	**Sensi-tivity**	**Number of Rules**	**Sensi-tivity**
1	6	97.3	58.7	33	53.2
2	7	96.9	58.0	52	57.1
3	8	96.8	59.8	52	58.6
4	8	95.8	58.7	48	55.5
5	7	95.8	58.1	57	57.5
6	7	96.7	57.1	61	54.7
7	8	96.4	57.1	54	57.9
8	7	95.4	58.3	70	58.3
9	6	95.4	58.3	45	53.5
10	7	94.4	57.6	61	57.7
Average Sensitivity:			58.2		56.4
Paired t Test of Sensitivity Differences in Sets					
95% Confidence Interval:				0.24 - 3.3	
2-tailed p value:				0.03	

Table 4.6 All Rules Over Ten Training/Validation Sets

Rule	Num. of Rule Sets	Local Support	Avg Sensitivity	Avg Specificity	Ratio Sens. to Spec.
c3000252433_Salmeterol__Hi	10	32	31.9	1.0	31.9
c3000250386_Glucocorticoid **c3000252425_Albuterol__Hic3**	10	25	11.7	0.8	14.1
c3000250652_Leukotriene_Rec_Antag	10	20	6.2	0.5	11.8
BetaAdrenergHic3NotAlbutOrSalmerol	10	9	1.9	0.0	46.5
c3000250386_Glucocorticoid AND Other_Pulmonary_Procedure_L26	4	21	4.0	0.8	5.3
c3000250386_Glucocorticoid AND AsthmaProcedures_cpt94010	4	18	3.4	0.4	9.1
c3000252425_Albuterol__Hic3 AND isFemale	4	20	2.4	0.6	4.0
c3000252425_Albuterol__Hic3 AND c85025A__ABN_COMPLETE_CBC	2	17	0.9	0.1	9.0
c3000252425_Albuterol__Hic3 AND Urinalysis__by_dip_stick	2	16	1.4	0.3	5.4
c3000250386_Glucocorticoid AND c82947_GLUC_BLD_LAB__QUANT AND Other_Pulmonary_Procedure	2	18	3.8	0.8	4.7
c3000252425_Albuterol__Hic3 AND c82947_GLUC_BLD_LAB__QUANT	2	22	2.3	0.7	3.5
c2_ALLERGY_SERVICE AND Other_Pulmonary_Procedure_L26	1	9	0.6	0.1	6.0
c3000252425_Albuterol__Hic AND c82947_GLUC_BLD_LAB__QUANT AND GT_5_FF_Vis_Per_Yr_NO_DES	1	17	0.4	0.2	2.0
c3000252425_Albuterol__Hic3 AND c82947_GLUC_BLD_LAB__QUANT AND isFemale	1	17	1.3	0.2	6.5
c3000252425_Albuterol__Hic3 AND GT_5_FF_Vis_Per_Yr	1	19	1.2	0.3	4.0
c3000252425_Albuterol__Hic3 AND c780_General_symptoms	1	14	2.0	0.7	2.9
c82947_GLUC_BLD_LAB__QUANT AND AsthmaProcedures_cpt94010 AND asthmaComorbids_473	1	9	0.2	0.1	2.0
c3000250386_Glucocorticoid AND Other_Pulmonary_Procedure_L26 AND isFemale	1	15	0.3	0.3	1.0
SerumElectrolytes AND AsthmaProcedures_cpt94010	1	18	1.0	1.3	0.8
c84443_ASSAY_THYROID_STIM_ AND AsthmaProcedures_cpt94010	1	13	1.3	0.6	2.2
AsthmaProcedures_cpt94010 AND V04_Need_for_prophylactic_vaccine	1	11	0.5	0.5	1.0
c3000253044_Fluticasone AND c782_Symptoms_involving_skin AND GT_5_FF_Vis_Per_Yr	1	8	0.5	0.5	1.0

Table 4.7 Best Rules Selected Among Ten Training Sets

Rule Number	**Rule**	**Data Source**
1	Salmeterol	Med ingredient
2	Glucocorticoid AND Albuterol	Med class
3	Leukotriene Receptor Antagonist	Med class
4	Beta Adrenergic Agent Not Albuterol or Salmeterol	Med class
5	Glucocorticoid AND Other_Pulmonary_Procedure	Med class, CPT aggregate*****
6	Albuterol AND Female	Med ingredient, Demographic feature
7	Allergy_Specialist_Visit AND Other_Pulmonary_Procedure	Visit feature, CPT aggregate
8	Albuterol AND Abnormal_CBC	Med ingredient, Lab abnormality
9	Albuterol AND Urinalysis_by_dip_stick	Med ingredient, Lab order
	***** Breathing capacity test, airway inhalation treatment, pulse oximetry, monoxide diffusing capacity, residual lung capacity, bronchodilator response evaluation	

References

1. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. In: Intelligence AAfA, editor. KDD-98; 1998; New York; 1998.

2. Thabtah F. A review of associative classification mining. Knowledge Engineering Review. 2007 Mar;22(1):37-65.

3. Cpt - current procedural terminology. American Medical Association; 2008.

4. Firstdatabank. [cited Feb 23, 2010]; Available from: www.firstdatabank.com

5. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The weka data mining software: An update. SIGKDD Explorations. 2009;11(1):10-8.

6. Witten IH, Frank E. Data mining: Practical machine learning tools and techniques with java implementations: Morgan Kaufmann Publishers; 2000.

7. Agrawal R, Imielinski T, Swami A. Mining associations between sets of items in large databases. ACM SIGMOD Int'l Conf on Management of Data. Washington D.C.; 1993.

8. Ordonez C, Ezquerra N, Santana CA. Constraining and summarizing association rules in medical data. Knowl Inf Syst. 2006;9(3):259-83.

9. Li W, Han J, Pei J. Cmar: Accurate and efficient classification based on multiple class-association rules. First IEEE International Conference on Data Mining (ICDM'01); 2001; 2001. p. 369-76.

10. Gu L, Li J, He H, Williams G, Hawkins S, Kelman C. Association rule discovery with unbalanced class distributions. Ai 2003: Advances in artificial intelligence; 2003. p. 221-32.

11. Thabtah F. Challenges and interesting research directions in associative classification. Sixth IEEE International Conference on Data Mining Workshops; 2006; 2006. p. 785-92.

CHAPTER 5

ACCURACY OF COHORT AMPLIFICATION FRAMEWORK RULES TO IDENTIFY ASTHMA CASES IN THE EHR

Introduction

The proposed value of the general disease cohort amplification framework was to learn the rules from known cases and apply them to identify additional cases. A validation study of the rules generated for asthma was performed to test the accuracy of the rules to identify new cases in the EHR. Validation is a recommended step in the application of framework rules in an EHR setting because it is a standard practice when introducing any new algorithm to identify cases for research. The specific objectives of such validation studies may vary. There may be use-case specific requirements for proof of accuracy. Various reference standards may be available or preferred. The current study was not intended for generality of the design. The purpose was to demonstrate the value of a set of predictive rules generated by the cohort amplification framework to identify additional asthma cases in the same EHR setting.

The accuracy of the rules was tested on a random sample of 992 subjects in the original study population, described in Chapter 1. These subjects were not used to generate the rules. The validation set subjects were classified as positive or negative according to a composite reference standard developed for this study. The rules were executed against the validation set. Accuracy was assessed by sensitivity and specificity of the rules. A practical value of the rules was explored by evaluation of the sensitivity gain when rules were combined with a validated ICD-9-CM based algorithm to identify cases.

There was no existing gold standard for identification of asthma patients in the study population or the resources to generate one. A composite reference standard (CRS) was developed from a combination of two imperfect but accepted standards and clinician review of provider documentation of care. The CRS consisted of various versions of asthma documentation in the EHR. By design, the framework used standard EHR content generated by routine health care delivery processes and documentation. Rules to identify additional cases used the same EHR content. Therefore, the reference standard positives for the validation study were patients who were considered to have asthma according to the medical records. The three components of the composite reference standard were asthma Problem List codes, asthma ICD-9-CM codes, or statements interpreted as a probable diagnosis of asthma in the clinical text documentation in the Intermountain Healthcare EHR.

Because the prevalence of asthma was estimated to be 11 % in the study population, approximately 89 % of the test set was expected to be negative. Many of the positive cases were determined by the coded evidence. There were not resources to review the clinical text documents of all those with no coded evidence of asthma to find the additional asthma cases. Text mining was used to select the most likely positive cases among those with no coded evidence. The likely cases were reviewed by clinicians. The text mining method was validated, and the expected error in the composite reference standard was estimated accordingly.

Background

Reference Standards

The evaluation of the accuracy of a diagnostic test or predictive algorithm in clinical and epidemiologic research requires a standard for comparison. The perfect standard is commonly called a 'gold' standard and is always preferred for research purposes. However, there are often circumstances that prevent the use of a gold standard. The gold standard level of proof may be autopsy or an invasive test that cannot ethically be performed on all subjects. The restrictions may be costs and resources associated with a thorough clinical review of cases or issues with access to the definitive records. In research situations where a less than perfect standard of comparison must be used, the standard is called a 'reference' standard. Methods that accommodate imperfect reference standards have been reported in the statistical and epidemiologic literature. Two recent review articles offered guidelines for appropriate methods.[1, 2] A composite reference standard was used to evaluate the accuracy of framework-generated rules to identify new asthma cases in the EHR because it enabled the use of multiple sources of evidence found in the electronic medical records. Statistical adjustment of one component of the composite reference standard was used to accommodate known error in the text mining methods used to select the most likely potential positives for manual review by clinical experts.

Composite Reference Standards

The composite reference standard was described for situations when multiple imperfect reference standards may be combined to form a better, although still imperfect, reference standard. The composite reference standard is an empirical methodology to leverage the available evidence. The level of proof stands on the acceptability of the combined evidence to consumers of the research.

Alonzo and Pepe[3, 4] described the composite reference standard (CRS) for binary outcomes: positive or negative. It is a staged approach in which one reference standard is applied. Cases not covered are then subjected to another reference standard. The examples of the use of this methodology were situations in which the first reference standard had good specificity, e.g., provided a believable level of proof that a positive determination was truly positive. However, the first reference standard had unacceptable sensitivity. Examples were diagnostic tests that would have been routinely applied if there were a clinical suspicion for the positive outcome and were not practical or ethical to apply when there was no evidence. Those that are covered by the first reference standard are considered resolved. The subsequent reference standard may be considered the ‘resolver’. The generic model is diagrammed in Table 5.1.

Other than enabling the evaluation of a new test where no gold standard exists, the CRS provides a straightforward, deterministic reference standard based on observable evidence. Several sources of evidence can be used, which is a practical approach to the sparse nature of many observations in the electronic health record. In addition to the common clinical workflow of performing diagnostic tests differentially based on prior evidence, the evidence may simply not be stored in a consistent manner. For example, some providers may document a diagnosis using structured data and others may document via dictated textual notes only. The CRS provides for a statistically unbiased comparison of a new test. No statistical manipulations are required.

Statistical Adjustment for Reference Standards with Known Error Estimates

In the case where there are imperfect reference standards with quantifiable error, there are statistical methods to adjust the sensitivity and specificity metrics for the new test. Staquet's equations [5] estimate the sensitivity and the specificity of a new test when a reference standard is imperfect but with known sensitivity and specificity and the new test and the reference standard are otherwise independent. With reference to Table 5.2, the equations are

SENSITIVITY NEW TEST =
Equation 5.1

$$(A + C)\ \text{specificity}_{RS} - C\ / \\ N\ (\text{specificity}_{RS} - 1) + (A + B)$$

SPECIFICITY NEW TEST =
Equation 5.2

$$(B + D)\ \text{sensitivity}_{RS} - B\ / \\ N\ (\text{sensitivity}_{RS}) - (A + B)$$

The equation for sensitivity was given in Equation 2.8, and the equation for specificity in 2.9. Staquet's adjustment can be applied to either the sensitivity or the specificity or both. Table 5.3 shows a hypothetical confusion matrix for a new test for a disease with 10% prevalence with the assumption that the reference standard is perfect. Table 5.4 shows the adjusted specificity under the assumption that the sensitivity of the reference standard was 85% but the specificity was perfect. The cell contents of Table 5.4 show the logic and the adjusted *specificity* of the effect of Equation 5.2 on Table 5.3.

Since the specificity of the reference standard was perfect, the sensitivity was not adjusted, as follows from Equation 5.1.

Application of Staquet's adjustment has often been reported in the literature and was recently reported to be the preferred adjustment when contrasted with two other methods.[6]

Reference Standards for Asthma

Diagnosis of Asthma

An expert panel, commissioned by the National Asthma Education and Prevention Program Coordinating Committee and coordinated by the National Heart, Lung, and Blood Institute of the National Institutes of Health, developed guidelines for asthma assessment, treatment and control.[7] They defined asthma as a "common chronic disorder of the airways that is complex and characterized by variable and recurring symptoms, airflow obstruction, bronchial hyperresponsiveness, and an underlying inflammation." Airway obstruction or narrowing, with subsequent airflow interference, is the dominant event leading to the typical clinical symptoms: wheezing, breathlessness, chest tightness, cough and mucous production. The bronchoconstriction may occur quickly, in response to variety of allergens or irritants. These symptoms require and respond to bronchodilator therapy. Airway inflammation is variable in intensity, cellular biology, and response to therapy and may distinguish asthma subtype phenotypes. As the disease progresses, swelling of the airways, hypersecretion of mucus and structural changes may occur and may not respond to treatment. Permanent structural changes are associated with a progressive loss of lung function.

What causes asthma? The complex disease has been characterized by abnormal immune system physiology, genetic predisposition, higher frequency among women in adult onset asthma, allergies and/or exposure to allergens, and certain respiratory viruses. Less well established environmental associations include tobacco smoke, air pollution, occupations, and diet. Current treatment can control symptoms but not prevent progression to the individual's underlying severity of asthma. In the opinion of the Expert Panel, there is insufficient evidence to recommend any specific strategies to prevent the development of asthma.

The Panel provided recommendations to establish a diagnosis of asthma. The clinician should assess for episodic symptoms and partially reversible airflow obstruction and exclude alternative diagnoses. Methods to establish the diagnosis are detailed medical history and physical exam focused on the upper respiratory tract, chest, and skin. Spirometry is required to demonstrate obstruction and assess reversibility. Reversibility is determined either by an increase in forced expiratory volume of ≥12% from baseline or by an increase ≥10% of predicted forced expiratory volume after inhalation of a short-acting bronchodilator. Additional studies should be performed as necessary to exclude alternate diagnoses.

In primary care, asthma may be suspected based on symptoms and history. Patients presenting with symptoms of airway obstruction are uncomfortable, and the provider may try drug therapy to open their airways. The suspicion of asthma may not be resolved in the medical record. Even if the trial is successful symptomatically, a reliable diagnosis of asthma should be supported by spirometry. A lack of objective testing and follow-up to determine asthma according to practice guidelines has been described.[8-10]

Incorrect diagnosis of asthma was studied in several European primary care settings.[11-14] A substantial rate of misdiagnosis of COPD as asthma among those greater than forty years of age has been reported.[15] Airway obstruction of unknown etiology may be diagnosed as ‘reactive airway disease’ and treated with the same medications as for asthma. This nonspecific diagnosis may prevent an appropriate workup to determine if the adult patient has asthma.[16, 17]

Gold Standard for Asthma

The gold standard for asthma diagnosis is, by definition, the expert consensus criteria to diagnose asthma.[7] The problem with the gold standard was described in a verification study of administrative disease codes to identify pediatric asthma cases using a gold standard based on the Canadian Asthma Consensus Report as reflected in the medical record. The criteria “had to be modified to reflect the cursory level of information available based on chart abstraction.”[18] Some studies claim the acceptance of encounter diagnosis codes as reference standards.[19, 20] The Problem List has been used as a reference standard. Others have used methods of inferring a diagnosis of asthma from the clinical documentation.

ICD-9 based reference standard

The ICD-9-CM category ‘493 Asthma’ and all subclassifications were used in the reported reference standards for asthma. The ICD-9-CM Official Guidelines for Coding and Reporting [21] address the overlap in COPD, asthma and bronchitis. Generally, the coding should follow the terms that were documented by the treating clinician. The guidelines contain some arbitrariness for coding acute exacerbations of asthma comorbid with chronic obstructive bronchitis or COPD as the primary code, while the bronchitis or COPD would be primary in a nonacute episode. The ICD-9-CM index defines 'Asthma, unspecified' as an asthma subclassfication. Coders may use '493.9' to code 'reactive airway disease' in adults. There is inherent uncertainty in the ICD-9-CM codes as a reference standard for asthma.

Three ICD-9-CM algorithms to ascertain asthma among adult patients using ambulatory ICD-9-CM codes were considered for this study. Blais et al.[22] validated 392 family practice subjects coded as asthma in the ambulatory, fee-for-service billing records against provider documentation. They estimated that approximately 77% of the ambulatory asthma diagnoses in the population-based billing system were provided by family physicians. Sensitivity was 85.5% and specificity was 88% for at least two asthma codes in a one-year period. Lix et al.[23] validated 529 subjects coded as asthma in ambulatory and inpatient encounters against population-based health surveys. Sensitivity fell between the 95% confidence interval from 43.9% to 51.3%. Specificity fell between 97.3% and 98.3% for algorithms for at least two ambulatory or one inpatient asthma code in a five-year period. Cases identified by inpatient-only codes contributed little to the population studied (0.1%).

Specificity was the important metric for the components of the composite reference standard developed for this study. The specificity may have been underestimated in the Blais study. They discussed the limitation of incomplete medical documentation and did not report the exact statements required to substantiate a diagnosis of asthma. At face value, it appears they required a diagnostic statement, which others have found to be insufficient. The Lix study validated the billing claims by the patient's independent report of having asthma in a population-based survey in the coverage area. In that context, the low sensitivity is not surprising, as patients may not have had an asthma-related visit during the five years or the asthma may have been comorbid with and coded as other respiratory illness. The rigorous study investigated several algorithms over varying time periods. The five-year period was consistent with the duration of evidence used for the other components of the reference standard for this study. An algorithm requiring two asthma codes was substantiated by Pacheco et al.,[10] who found that at least two asthma-related events were required to accurately classify asthma patients. The Lix algorithm for two visits over five years was selected as the reference standard for ICD-9-CM codes for the current study.

Problem List

Although Problem List in the EHR is known to be incomplete, it has been shown to be reliable at >98% specificity.[24, 25] Szeto et al.[25] studied the accuracy of the ambulatory Problem List compared to chart review for 148 patients attending a general medicine clinic at a Veterans Administration (VA) hospital. Sensitivity ranged from 42 to 81%, while specificity ranged from 98 to 100%. The Problem List has been used as a reference standard for positive disease in recent studies.[26-28] An algorithm to identify asthma cases for genomewide association studies in the Electronic Medical Records and Genomics (eMERGE) network[29] described the Problem List codes for asthma as a medical diagnosis of asthma.[10]

The literature supports high specificity of an asthma diagnosis in the Problem List. It is consistent with the expected specificity using the ICD-9-CM algorithm selected for this study. Further, and more importantly, the uncertainties of the asthma diagnosis have been noted. A coded entry for asthma in the Problem List may be incorrect by validation standards, but asthma may have been suspected at some visit and used for treatment decisions by the provider. In addition, notation in the EHR Problem List is considered part of the legal medical record.

Expert Review of Clinical Documentation

Various methods to review the clinical documentation for an asthma diagnosis have been reported. Wilchesky et al.[30] used trained study personnel to review a five year period in each chart and record whether asthma, as well as 25 other medical conditions, were present or absent. Blais et al.[22] reported study nurse review of the chart for a diagnosis of asthma. Neither reported further on the words or conditions constituting an asthma diagnosis. Three studies published criteria to determine asthma from the clinical documentation as shown in Table 5.5.[18, 31, 32] Review of clinical documents is time consuming, and the cost of labor can be high. Only one of these studies[18] used physician reviewers. Two used nurses, one used medical records technicians, and one used trained abstractors.

Text Mining

Text mining methods were used to limit the number of cases requiring expert review of the clinical documentation. Cohen and Hersh[33] described text mining as a way to examine the relationships of specific information both within and between documents. Text mining techniques used in the current study were information extraction, specifically named entity recognition, and data mining.[34] Named entity recognition identifies the terms of interest in the text. Named entity recognition can be difficult to accomplish, given that the terms may have multiple meanings, may consist of multiple words and may be misspelled. Misspellings in the clinical documentation are reportedly in the range of 10%.[35] In the current study, the named entities searched for a direct mention of 'asthma' in the clinical documents were singular terms and rarely used with a different meaning. In text mining, conventional data mining methods may be applied to text features. A text feature of interest for the current study was the frequency of documents having a named entity for asthma for each subject. The Oracle data mining manual presented the use of frequencies of named entities to characterize documents and the exploration of their patterns as a common text mining task.[36]

Methods

The study was designed to evaluate the sensitivity and specificity of rules generated by a general disease classification framework to identify asthma cases in the electronic health record. The validation set consisted of 992 randomly selected subjects from the target population as described in Chapter 1. The validation subjects were assigned as positive or negative for asthma. A small number of negative cases had 'possible' asthma, which did not meet the criteria for 'probable' asthma. The rules were executed against the 2007-2008 EHR data content for the validation subjects. The sensitivity, specificity, and their 95% confidence limits were described for the rules and for the union of subjects identified by either the rules or the ICD-9-CM algorithm used in the study.

Framework Generated Rules

The rules to identify new asthma cases were generated using the cohort amplification framework. This associative classification framework was designed to use generalized EHR candidate data and processes to generate prediction rules to classify multiple health conditions in the EHR. The framework was described in Chapter 3. The development of the rules used in the current study was presented in Chapter 4. The rules were selected from ten sets of rules generated on separate random samples of the training data based on the individual rule's generality and accuracy.

The rules are expressed as attributes or combinations of attributes present in the EHR during the two year data mining period, 2007-2008. The expressed rules represent the full rule syntax: 'if <attributes> is true, then classify as an asthma case'. The rules are shown in Table 5.6.

Validation Set

The validation set (n=992) were randomly selected from the target population described in Chapter 1 and were not used in the development of the rules. The validation set were annotated as probable asthma cases or negative for asthma. The validation cases were classified from the available coded evidence for disease in the electronic health record as well as expert review of the clinical documents to determine additional probable asthma patients that may not have been coded as such. The negative group included those with suggested evidence not meeting the coded standards or the documentation criteria (Table 5.5). The validation set was created and classified before the rules were executed. The rules were executed against data for subjects in the validation set during the years 2007-2008.

Composite Reference Standard for the Study

The three components of the composite reference standard (CRS) for positive disease were Problem List codes, ICD-9-CM codes, or statements indicating asthma in the clinical text documents as interpreted by physicians expert in the diagnosis of asthma. CRS negative disease was determined if there was no coded evidence, and either physicians determined asthma was not probable from clinical documentation or asthma was not mentioned more than once in five years of clinical notes. The composite reference standard is shown in Table 5.7.

The validated ICD-9-CM based algorithm used was two ambulatory visits or one inpatient visit, including emergency department, over a two year period, 2007-2008.[23] Either the ICD-9-CM based evidence or an active coded Problem List entry for asthma determined a positive case, but did not resolve whether the case was negative. The 'Resolver' was the Clinical Text Documentation. Assignments were made to probable or negative asthma status based on statements about asthma in clinical text documents stored for study patients in the Intermountain Enterprise Data Warehouse (EDW) over a five-year period, 2004-2008.

A text mining approach was used to identify clinical documents with mentions of asthma terms and to develop a probability-based classifier to distinguish the most likely cases for further manual review to identify asthma cases without coded evidence. Details of the development and verification of the text mining approach are described in Appendix A. The result was a simple Bayes rule classifier based on the total number of clinical documents with mentions of asthma, including ten variations of the term and its common misspellings, per subject over a five year period. The model demonstrated a probability of 1.3% that a subject with zero or one documents containing the ten asthma terms was likely to have asthma. It was developed on 10,448 random subjects from the target population described in Chapter 1. The Bayes negative asthma status prediction model is shown in Table 5.8. Verification of the model's prediction of negative asthma status was conducted by manual review of a convenience sample of 160 test subjects (2%) with zero or one clinical documents containing asthma terms. The verification process used a liberal interpretation of a single probable asthma statement as an error. The error rate was 2.5%.

The 'negative asthma status' prediction model was populated with data from the validation set with similar results (Table 5.9). Thirteen percent of the predicted negative asthma status validation subjects' clinical documents were reviewed by an independent nurse reviewer. Clinical document review methods are described in Appendix B. The error rate was 1%. Seventy-eight percent of the validation sample was covered by the negative asthma status prediction model, with an estimated chance of falsely labeling a positive case as a negative one of 1-2.5%. Assuming the highest error estimate (2.5%) and using Staquet's adjustment, the composite reference standard has an estimated sensitivity of 87% (Table 5.10). Assuming the lowest estimate (1%), the estimated sensitivity is 94% (Table 5.11).

Cases in the resolver component of the composite reference standard with more than one document with an asthma term (n=98) were manually reviewed (Appendix B) in two stages. First, expected mentions among negative cases were screened out (n=52). These consisted of hypothetical statements about the disease such as "at risk for", references about someone else and the disease such as in "family history of", or assertions that the patient did not have asthma.[37, 38] Cases with a statement concerning the possibility of a diagnosis or history of asthma were reviewed by physician experts (n=46). Of these, 28 subjects were judged as probable asthma cases. For purposes of this study, the composite reference standard was assumed to have perfect specificity. Codes or statements defined as positive for the reference standard were accepted as a provider's best judgment at the time, given the uncertainty and misdiagnosis of asthma previously described.

Statistical Analysis

The sensitivity and specificity with 95% confidence intervals were computed for the classification of the validation data using the framework rules. Standard algorithms for sensitivity (Equation 2.8) and specificity (Equation 2.9) were used, reporting both adjusted and nonadjusted specificity, and the 95% confidence intervals were computed using the Wilson score method.[39] This sensitivity was also computed and reported for the cases covered by either the framework or the validated ICD-9-CM algorithm. The specificity was not affected by the ICD-9-CM algorithm. The ICD-9 algorithm was a component of the positive reference standard, and its absence did not determine a negative case. The purpose of the joint sensitivity statistic was to evaluate the contribution of the framework rules beyond cases known by a validated ICD-9-CM based identification algorithm.

Frequencies of occurrence of the single attributes that formed the rules as well as some interesting profile characteristics among composite reference standard positives and negatives and true positives, false negatives, true negatives and false positives classified by the framework rules were described. Differences in the frequencies between CRS positives and negatives were calculated with 95% confidence intervals using Newcombe's method. Confidence intervals that include zero are not statistically significant differences.[40] Ratios of the frequencies, also known as risk ratios, with 95% confidence intervals based on the Cox-Hinkley-Miettinen-Nurminen method,[41] were also calculated for some asthma comorbidities.

Results

The sensitivity of the framework generated rules was 54% with 95% confidence interval 46.0% to 62.4%. The most conservatively adjusted specificity, under the assumption that the error in the negative asthma prediction model was 1%, was 97.1% with 95% confidence interval 95.8% to 98.1% (Table 5.12). When cases identified by the ICD-9 algorithm of two asthma codes over five years were combined with cases identified by the rules, the sensitivity was 83% (76.2% to 88.6%) (Table 5.13). The specificity was unchanged. Using the ICD-9 algorithm alone, the sensitivity was 70% (62.2% to 77.3%) (Table 5.14). The rules alone did not identify as many CRS positive cases as two ICD-9 codes over five years alone. However, it did contribute an additional 13% of the CRS positive cases. The difference in the proportions contributed by both versus by ICD-9 codes alone was statistically significant, with the 95% confidence interval of the difference from 3.1% to 22.7%.

Table 5.15 shows the sequential rules, the number of true positives (TP) and false positives (FP) covered, and the Positive Predictive Value (PPV) of each rule. Subjects often met multiple rules. Table 5.16 shows the number of subjects that met each combination of rules, with no distinction as to the classification accuracy. This shows how the rules clustered. The first covering rules are the focus of analysis because they perform the classification. Table 5.15 shows that two of the nine rules covered no CRS positives, one had no unique coverage, and five rules had poor (< 70%) positive predictive value on the validation data.

The diagnosis of asthma may be confounded by other respiratory problems that may be comorbid with asthma or mistaken as asthma: bronchitis, sinusitis, rhinitis and chronic airway obstruction (COPD).[7] All of these conditions were found to be more frequent among the CRS positive subjects (Table 5.17). Since similar medications may be used as well, frequencies among these respiratory problems among the true positive (TP), false negative (FN), true negative (TN), and false positive (FP) rule classifications were explored and are shown in Table 5.18. Bronchitis and COPD appeared most frequently among the incorrect classifications. Sinusitis, acute bronchitis and allergic rhinitis occurred with false negative classifications more frequently than false positives. Chronic bronchitis occurred equally with both. COPD occurred with false positive classifications more frequently than false negatives.

Differences in demographic, health care encounter and EHR documentation characteristics between the composite reference standard (CRS) positives and negatives are shown in Table 5.19. Characteristics with no statistically significant difference are italicized. There is virtually no difference in the ages. There are more females (64.5% vs. 53.5%) in the positive group than the negative. The positives have more obesity (7.2% vs. 5.2%) and more pain (14.5% vs. 10%), as documented by ICD-9 codes, but neither was a statistically significant difference. Positives appear to have more health problems as evidenced by statistically significant differences in those with more than six ambulatory provider visits in one year (58% vs. 40.7%) and at least one emergency department visit over two years (31.2% vs. 18.9%). Positives were more likely to have a populated Problem List (79% vs. 66.3%) or Medication List (97.8% vs. 90.6%) in the electronic health record. Since all subjects had at least two visits to Primary Care during the study period, where providers typically use the Medication and Problem Lists, this may reflect more patients among the negatives with acute or short term problems, such as the common cold or health checkups, that providers have no need to track in the Problem List. Similarly, the negatives may be less likely to have any prescribed medications. The CRS asthma-positive group were no older, but were generally less healthy and more likely to be female than the CRS asthma-negative group.

Table 5.20 shows how these characteristics aligned with the classification outcomes: TP, FP, TN, and FN. Obesity documented in ICD-9 codes was associated more frequently with true and false positive classifications compared to the overall classification distribution. ICD-9 codes for pain, more than six health ambulatory provider visits in one year, and at least one emergency department visit over two years were associated with true positive predictions. Female gender was associated with false negative classifications.

The associative classification metrics used to generate the rules may also provide new knowledge about the targeted disease or condition. Some previously described associations – allergies, eczema, gastric esophageal reflux disease (GERD),[42] sleep apnea [7] - and unexpected associations between asthma and other medical conditions were noted in the training data. Associations that persisted in the validation data between CRS positive and CRS negative subjects are shown in Table 5.21. The first group contains GERD and allergy-related attributes. These attributes all had statistically significant differences in their frequencies between CRS positives and negatives. They also had risk ratios, the ratio of proportions, with 95% confidence intervals greater than unity. The second group contains sleep disorder, often related to obesity (Table 5.19), along with diabetes, hypertension, and cardiac problems. Diabetes and hypertension were not found to be highly associated with asthma in either the training data or the validation data, but they were included here because they are also known to be related to obesity and cardiac problems. Both cardiac symptoms and cardiologist visits were significantly higher among CRS asthma subjects even though their age was no different from the CRS negative subjects. The third group shown is arthritis and fibromyositis or neuritis ICD-9 codes along with a higher frequency of narcotic analgesics use. The ICD-9 codes defined by the National Arthritis Workgroup[43] were combined to form the arthritis grouping. ICD-9 729 covers unspecified rheumatism, fibrositis, myalgia, myositis, neuritis and other inflammatory conditions of related tissues. This was interesting because fibromyalgia syndrome – characterized by arthritis, generalized muscular pain, sleep disorders, and other associations discovered among the asthma subjects – also occurs more often in women but has not been described as a comorbidity with asthma. The CRS positive asthma group had a significantly higher proportion of women (64.5% versus 53.5%), as described above. The use of narcotic analgesics could not be explained by any other associations in the data other than the pain inherent in the inflammatory joint and muscle conditions included in this group. The fourth group shows associations among an ICD-9 code for nonspecific findings on imaging and other diagnostic procedures, and the use of antipsychotic medications. The use of antidepressant medications was higher among the CRS positive subjects, but the stronger association was the combination of antidepressant and antipsychotic agents. These are often used together in more severe cases of clinical depression.

Discussion and Conclusions

The accuracy results showed that the cohort amplification process was successful in learning patterns among a standard EHR-based data set using exemplars of a target cohort to identify additional members of the cohort directly in the EHR. Asthma was one of the first medical conditions chosen to test the cohort amplification process. A fair amount of difficulty in identifying asthma cases using retrospective EHR data was described in this chapter. The inaccuracy of the classification rules reflect some of the same problems: uncertain diagnoses, misdiagnoses due to confounding respiratory problems, similar treatments and medications used in related respiratory problems, and the episodic nature of the disease itself. Nonetheless, the rules learned on exemplar data of subjects with asthma noted in the Problem List were useful in identifying additional asthma subjects beyond those that could be identified using the ICD-9 codes in the EHR.

Further analysis of the number of subjects that met each rule demonstrated poor predictive value by five of the nine rules (Table 5.15). This could be due in part to an incorrect reference standard. The reference standard developed for asthma was flawed by the same problems in discerning true asthma cases from retrospective EHR documentation, even when interpreted by experts. In a real-world application of the cohort amplification framework, the false positive and false negative cases may be further reviewed by experts. Those resources were beyond the scope of this study. Reference standards for validation of retrospective case identification methods for secondary uses of the EHR are a difficult problem. If a good reference standard exists, it implies there is already a case identification method. As the expectations for secondary use expand with the wider adoption of the EHR in practice and the promise of standardized health care data, computer-based methods to ascertain cases must be used. Perhaps ‘triangulation’ strategies will prove successful, in which machine learning and classification methods use coded data and natural language processing to compare evidence from multiple sources in the EHR and focus expert review on the marginal cases. A repository of such evidence may also prove useful so it does not have to be rediscovered and clinicians may authenticate or reject a machine generated disease label.

Another plausible reason for the poorer predictive value than was generated from the training data is the variation that may exist among exemplars as compared to the population of true cases in the EHR. In this study, the asthma exemplars were drawn from those coded in the Problem List. Frequencies of rule component attributes were checked between CRS positive cases that were also coded in the Problem List (PL) (n=60) versus the entire group (n=138). The largest differences found were a lower proportion of female subjects (57% versus 64%) and less use of albuterol (17% versus 25%) in the PL group. Smaller differences in other rule components and a similar distribution of differences in demographic and comorbidity characteristics showed the expected sampling variation and perhaps some bias of an exemplar cohort. Unfortunately, we cannot assess the differences before the rules are generated so two remedies are suggested. First, the standard processes for knowledge discovery from databases include refinement and iteration of the machine-learning steps. In actual application, one may correct for the less useful rules or biases discovered in the training data and repeat the rules generation. Secondly, more conservative criteria for the selection of rules was used in the cohort amplification rule pruning methods than in the classic methods. With the pattern-learning focus on disease processes described by highly related EHR data, the pruning constraints may need to be tightened. For example, for the rules used in this study, a minimum of five training cases had to be covered by a rule or it was pruned. In addition, rules were pruned if the positive predictive value on the training data was less than 70%. The results of this study suggest that differences in the exemplar and target populations as well as random variation may necessitate stricter pruning criteria and thus more general rules.

The analysis of false positives and false negatives among the subjects with associated respiratory problems was useful. There did not appear to be remarkable differences in the distribution of classification outcomes for subjects with sinusitis or allergic rhinitis. However, bronchitis, in particular chronic bronchitis, and chronic obstructive pulmonary disease (COPD), with or without comorbid asthma, resulted in less accurate classification by the rules. These conditions are known to be confounders in asthma diagnosis and case finding. For some research purposes, cases with these complications have been excluded.[10] For many purposes, such as clinical quality improvement, they may be important to capture because patients may be at higher risk. In this study, less than 5% of the validatation cohort had evidence of chronic bronchitis or COPD. These may need focused expert review or development of a classifier using specific EHR data and methods relevant to this problem.

Other demographic, general health and comorbidity characteristics presented in the results may provide knowledge to improve the identification of cases. The rules may be refined and reiterated based on the additional domain information directly or further domain analysis may be inspired. A suggested increase in false positive classifications correlated with the ICD-9 code for obesity might invoke further review of the data for these subjects. The same may be said for the higher frequency of female subjects with a false negative classification. For purposes of this study, a benefit of the associative classification method was shown. The frequencies of all attributes and attribute combinations that are used to form the rules are exposed in the training data. In this study, selected attributes of interest in the training data were generated and described for the validation set to help understand the classification outcomes.

Another benefit of the associations exposed by this classification method is the potential for serendipitous knowledge discovery. In this study, focused on the development of rules to identify asthma, several associations that appeared in the training data and persisted in the validation data were described. These characterized some known comorbidities of asthma, such as gastric reflex disease, sleep disorder and symptoms of skin (allergic reactions). However, associations with cardiac problems, arthritis and other inflammatory conditions of the connective tissue have not been described as comorbidities with asthma. They may not be. The associations discovered among existing data do not imply causality nor rule out a shared dependency on some other causal factor. However, the associations reported in Table 5.21 were originally noted in the training data and persisted in the validation data as statistically significant associations among asthma-positive cases compared to asthma-negative cases. Further review of these associations with domain experts is planned.

Table 5.1 General Composite Reference Standard

		Reference Standard 1		**Reference Standard 2**	
		Positive	***Not Covered***	**Resolver**	
			→ → →	**Positive**	**Negative**
New Test	**Positive**	+			
				+	
					-
	Negative	+			
				+	
					-
		Final Determination shown as ‘+’ or ‘-‘			

Table 5.2 Reference for Equations 5.1 and 5.2

		Reference Standard **(RS)**		
		+	-	Total
New Test	+	A True Pos	C False Pos	
	-	B False Neg	D True Neg	
	Total			N
Sensitivity		TP/(TP+FN)		
Specificity			TN/(TN+FP)	

Table 5.3 Unadjusted Sensitivity and Specificity

		Reference Standard (RS)		
		+	-	Total
New Test	+	90	40	108
	-	10	860	892
	Total	100	900	1000
Sensitivity		90/100 = .90		
Specificity			860/900 = .96	

Table 5.4 Adjusted Specificity from Table 5.3
Given Sensitivity$_{RS}$=.85, Specificity$_{RS}$ =1

		Reference Standard (RS)			
New Test		+	0.15 Pos Missed	-	Total
	+	90	(-16)	24	130
	-	10	(-2)	858	870
	Total	100	(-18)	882	1000
Sensitivity		90/100 = .90			
Specificity			858/882 = .97		

Table 5.5 Criteria to Determine Asthma in Clinical Documents

First Author	Asthma Definition	Used This Study
Vollmer	**Probable**	
	2 or more asthma care (AC) visits	x
	Single AC visit and prior history	x
	Single AC visit for active symptoms	
	SingleAC visit and response to meds	x
	Possible	
	Patient reported history only	x
	Uncorroborated emergency diagnosis	
	Suspected with no clear resolution	x
Twiggs	**Definite**	
	Clinical dx and 2+ visits acute wheezing	x
	Possible	
	Asthma symptoms + history allergy, wheezing	x
To	2 visits for wheezing	x
	1 visit wheezing + risk factor	x
	1 visit wheezing + response meds	x

Table 5.6 Asthma Identification Rules

Rule Number	**Rule**	**Data Source**
1	Salmeterol	Med ingredient
2	Glucocorticoid AND Albuterol	Med class
3	Leukotriene Receptor Antagonist	Med class
4	Beta Adrenergic Agent Not Albuterol or Salmeterol	Med class
5	Glucocorticoid AND Other_Pulmonary_Procedure	Med class, CPT aggregate*****
6	Albuterol AND Female	Med ingredient, Demographic feature
7	Allergy_Specialist_Visit AND Other_Pulmonary_Procedure	Visit feature, CPT aggregate
8	Albuterol AND Abnormal_CBC	Med ingredient, Lab abnormality
9	Albuterol AND Urinalysis_by_dip_stick	Med ingredient, Lab order
	***** Breathing capacity test, airway inhalation treatment, pulse oximetry, monoxide diffusing capacity, residual lung capacity, bronchodilator response evaluation	

Table 5.7 Diagram of the Composite Reference Standard

		Problem List or ICD-9-CM codes		**Clinical Text Documents**	
		Positive	***Negative***		
			→ → →	**Positive**	**Negative**
Frame-work Rules	**Positive**	+			
				+	
					-
	Negative	+			
				+	
					-
		Final Determination shown as ‘+’ or ‘-‘			

Table 5.8 Negative Asthma Status Prediction Model Test Data

Named Entity Count	Coded Evidence case count	Likelihood Coded Evidence	Conditional Probability Coded Evidence	No Coded Evidence case count	Likelihood No Coded Evidence	Conditional Probability No Coded Evidence	Likelihood NE count
0-1	103	0.010	0.013	8037	0.769	0.987	0.779
2+	1057	0.101	0.458	1251	0.120	0.542	0.221
Totals	1160	0.111		9288	0.889		1.000

Table 5.9 Negative Asthma Status Prediction Model Validation Data

Named Entity Count	Coded Evidence case count	Likelihood Coded Evidence	Conditional Probability Coded Evidence	No Coded Evidence case count	Likelihood No Coded Evidence	Conditional Probability No Coded Evidence	Likelihood NE count
0-1	8	0.008	0.010	784	0.783	0.990	0.798
2+	102	0.102	0.510	98	0.098	0.490	0.202
Totals	110	0.110		882	0.881		1.000

Table 5.10 Adjusted Sensitivity Composite Reference
Standard w/ 2.5% NEG prediction error

			Truth Assumption for this Study	
			POS	**NEG**
Composite Reference Standard	**POS**	Coded	110	
		Reviewed	28	
	NEG	Reviewed		70
		0-1 NE Asthma		784
		Prob Pos | 0-1 NE ~ 2.5%	*(+20)*	*(-20)*
		TOTAL	**158**	**834**
		Sensitivity	0.87	
		Specificity		1.00

Table 5.11 Adjusted Sensitivity Composite Reference Standard w/ 1% NEG prediction error

			Truth Assumption for this Study	
			POS	**NEG**
Composite Reference Standard	**POS**	Coded	110	
		Reviewed	28	
	NEG	Reviewed		70
		0-1 NE Asthma		784
		Prob Pos | 0-1 NE ~ 1%	*(+9)*	*(-9)*
		TOTAL	**147**	**845**
		Sensitivity	0.94	
		Specificity		1.00

Table 5.12 Sensitivity and Specificity of Framework Rules

		Composite Reference Standard			
		POS	**NEG**		
Framework Rules	**POS**	75	29		
	NEG	63	825		
	TOTAL	138	854		
				95% Confidence Interval	
Sensitivity		54.3	%	46.0 -	62.4
Specificity			96.6		
*****Specificity Adjusted 1% Error CRS			97.1	95.8 -	98.1
Specificity Adjust. 2.5% Error CRS			97.8		
*The most conservative adjustment was used.					

Table 5.13 Sensitivity of Framework Rules OR 2 ICD-9 Codes

		Composite Reference Standard			
		POS	NEG		
FW Rules OR 2+ ICD9 Codes	POS	115	29		
	NEG	23	825		
	TOTAL	138	854		
				95% Confidence Interval	
Sensitivity		83.3	%	76.2 -	88.6

Table 5.14 Sensitivity of 2 ICD-9 Codes Over Five Years

		Composite Reference Standard			
		POS	NEG		
2+ ICD9 Codes	POS	97	NA		
	NEG	41	NA		
	TOTAL	138	NA		
				95% Confidence Interval	
Sensitivity		70.3	%	62.2	77.3

Table 5.15 First Ordered Rule Met by True Positives and False Positives and Positive Predictive Value (PPV) of the Rule

		True Positive n=75	False Positive n=29	
Rule Number	**Rule**	**Number Met Rule**	**Number Met Rule**	**PPV (%)**
1	Salmeterol	41	6	87.2
2	Glucocorticoid AND Albuterol	17	2	89.5
3	Leukotriene Receptor Antagonist	5	2	71.4
4	Beta Adrenergic Agent Not Albuterol or Salmeterol	3	3	50.0
5	Glucocorticoid AND Other_Pulmonary_Procedure	5	6	45.5
6	Albuterol AND Female	3	7	30.0
7	Allergy_Specialist_Visit AND Other_Pulmonary_Procedure	1	0	100.0
8	Albuterol AND Abnormal_CBC	0	3	0.0
9	Albuterol AND Urinalysis_by_dip_stick	0	0	

Table 5.16 Number of Subjects Meeting Each Rule

Number of Cases	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Rule 7	Rule 8	Rule 9
895									
18	xxx								
3		xxx							
2			xxx						
2	xxx		xxx						
3				xxx					
1	xxx			xxx					
2			xxx	xxx					
1	xxx		xxx	xxx					
10					xxx				
3	xxx				xxx				
1		xxx			xxx				
2	xxx	xxx			xxx				
1			xxx		xxx				
3	xxx		xxx		xxx				
2				xxx	xxx				
1	xxx			xxx	xxx				
3						xxx			
1		xxx				xxx			
1	xxx	xxx				xxx			
1			xxx			xxx			
1	xxx	xxx		xxx		xxx			
1		xxx			xxx	xxx			
2							xxx		
1					xxx		xxx		
2	xxx				xxx		xxx		
1		xxx			xxx		xxx		
1			xxx		xxx		xxx		
1	xxx		xxx		xxx		xxx		
2								xxx	
1		xxx						xxx	
2						xxx		xxx	
1		xxx				xxx		xxx	
1		xxx	xxx	xxx		xxx		xxx	
1		xxx			xxx	xxx		xxx	
1	xxx	xxx	xxx		xxx	xxx	xxx	xxx	
1						xxx			xxx
1	xxx	xxx				xxx			xxx
2	xxx	xxx	xxx			xxx			xxx
1		xxx			xxx	xxx			xxx
1	xxx	xxx			xxx	xxx			xxx
1								xxx	xxx
1	xxx	xxx						xxx	xxx
1				xxx				xxx	xxx
1	xxx	xxx		xxx				xxx	xxx
1	xxx	xxx	xxx		xxx			xxx	xxx
4						xxx		xxx	xxx
3		xxx				xxx		xxx	xxx
1	xxx	xxx				xxx		xxx	xxx
1		xxx		xxx		xxx		xxx	xxx
1		xxx	xxx	xxx		xxx		xxx	xxx
1	xxx	xxx			xxx	xxx		xxx	xxx
1		xxx	xxx		xxx	xxx		xxx	xxx
1	xxx	xxx	xxx		xxx	xxx		xxx	xxx
1		xxx	xxx		xxx		xxx	xxx	xxx
1						xxx	xxx	xxx	xxx

Table 5.17 Respiratory Comorbidities
Composite Reference Standard (CRS) Positives Versus Negatives

	Prcnt CRS POS with Attribute	Prcnt CRS NEG with Attribute	Difference in Proportions	95% Confidence Interval of Difference		Relative Risk
				Low Bound	High Bound	
Associated Respiratory Conditions						
Number of Subjects	138	854				
ICD 461_Acute_sinusitis	27.5	20.0	7.5	0.2	15.9	1.4
ICD 466_Acute_bronchitis	26.1	8.7	17.4	10.5	25.5	3.0
ICD 473_Chronic_sinusitis	5.1	2.2	2.8	0.0	7.9	2.3
ICD 477_Allergic_rhinitis	21.0	8.2	12.8	6.5	20.5	2.6
ICD 490_Bronchitis_NOS	11.6	5.0	6.6	1.9	13.1	2.3
ICD 491_Chronic_bronchitis	4.3	1.1	3.3	0.8	8.1	4.1
ICD 496_Chron_airways_obstructio	5.8	2.6	3.2	0.1	8.5	2.3

Table 5.18 Classification Frequencies of Subjects with Associated Respiratory Conditions

Associated Respiratory Conditions	Number w/ Attribute	Prcnt True Pos	Prcnt False Pos	Prcnt True Neg	Prcnt False Neg
ICD 461_Acute_sinusitis	209	12.4	3.8	78.0	5.7
ICD 466_Acute_bronchitis	110	21.8	8.2	59.1	10.9
ICD 473_Chronic_sinusitis	26	19.2	0.0	73.1	7.7
ICD 477_Allergic_rhinitis	99	21.2	4.0	66.7	8.1
ICD 490_Bronchitis_NOS	59	22.0	5.1	67.8	5.1
ICD 491_Chronic_bronchitis	15	20.0	20.0	40.0	20.0
ICD 496_Chron_airways_obstruction	30	23.3	23.3	50.0	3.3

Table 5.19 Demographic, Encounter and EHR Documentation Characteristics Composite Reference Standard (CRS) Positives versus Negatives

	Prcnt CRS POS with Attribute	**Prcnt CRS NEG with Attribute**	**Difference in Proportions**	**95% Confidence Interval of Difference**	
				Low Bound	**High Bound**
Demographic/ Encounter/ Documentation					
Number of Subjects	138	854			
Age Greater Than 47	57.2	57.1	0.1	-8.8	8.9
Female	64.5	53.5	11.0	0.9	18
ICD 278 Obesity and other alimentation	7.2	5.2	2.1	-1.5	8
*ICD9 codes for pain**	14.5	10.0	4.5	-1.4	10.8
6+ Ambulatory Provider visits/year	58.0	40.7	17.3	8.0	25.5
Emergency department visit	31.2	18.9	12.3	4.5	20.7
Populated Problem List	79.0	66.3	12.7	4.8	19.8
Populated Medication List	97.8	90.6	7.2	2.5	9.4
* Pain, migraine, pain and symptoms associated with female organs					

Table 5.20 Classification Frequencies of Subjects by Demographic, Encounter and EHR Documentation Characteristics

Demographic/ Encounter/ Documentation	**Number w/ Attribute**	**Prcnt True Pos**	**Prcnt False Pos**	**Prcnt True Neg**	**Prcnt False Neg**
Age Greater Than 47	567	8.1	3.4	82.8	5.8
Female	546	8.1	3.3	80.4	8.2
ICD 278 Obesity and other alimentation	54	14.8	5.6	75.9	3.7
ICD9 codes for pain	105	13.3	0.9	79.7	5.7
6+ Ambulatory Provider Visits in One Year	428	12.6	3.7	77.6	6.1
Emergency department visit	204	13.7	3.4	75.3	7.3
Populated Problem List	675	9.2	3.7	80.1	7.0
Populated Medication List	909	8.3	3.2	82.0	6.6
TOTAL DISTRIBUTION	992	7.6	2.9	83.2	6.4

Table 5.21 Significant Comorbidity and Symptom Associations with Asthma

	Prcnt CRS POS with Attribute	**Prcnt CRS NEG with Attribute**	**Difference in Proportions 95% Confidence Interval**		**Ratio of Proportions* 95% Confidence Interval**	
			Low Bound	**High Bound**	**Low Bound**	**High Bound**
Significant Comorbidity & Symptom Associations						
ICD 530 GERD or Proton pump inhibitors	43.5	23.9	11.1	28.4	1.4	2.3
ICD 787 Symptoms digestive system	20.3	12.8	4.0	18.0	1.1	2.3
ICD 782 Symptoms of skin	26.1	10.7	8.4	23.5	1.7	3.4
Allergy Specialist Visit	6.5	0.6	2.8	11.3	4.0	31.2
ICD 327 Organic sleep disorder	13.8	6.9	1.6	13.7	1.2	3.2
ICD 250 Diabetes mellitus	*17.4*	*14.2*	*-2.8*	*10.7*	*0.8*	*1.8*
ICD 401 Essential hypertension	*42.8*	*35.5*	*-1.80*	*15.7*	*1.0*	*1.5*
CARDIOLOGY Visit	22.5	14.3	1.5	16.1	1.1	2.2
ICD 785 Symptoms cardio system	15.2	6.8	3.0	15.5	1.4	3.5
Arthritis (ICD 715, 716, 719, 726, 727, 728)**	48.6	33.0	6.7	24.3	1.2	1.8
ICD9 729 Fibromyositis, Neuralgia, CFS	23.9	15.9	1.1	16.1	1.1	2.1
Narcotic Analgesics	48.6	33.5	6.4	24.1	1.2	1.7
ICD 793 Nonspecific abnormal findings	12.3	7.4	0.4	12.0	1.0	2.7
Antipsychotic meds	18.1	10.8	1.1	14.6	1.1	2.5
Antipsychotic & antidepressant meds	8.7	4.0	0.8	10.7	1.2	4.0

*** Ratio of Proportions also known as Relative Risk**

** National Arthritis Workgroup definition

Table 5.22 Document Types Used in Text Mining
Having Frequency > 5% Among Evidence Sample

CLINICAL NOTES: Jan 1, 2004 - Dec 31, 2008.

Named Entities: ASTHMA, ASTHMATIC, ASTHMATICUS, ASTHAM, ASTHM, ASTHMAS, ASTHMATICS, ASTHAMA, ATHMA, ASHTMA

Clinical Document Type	Evidence Sample			No Evidence Sample		
	n = 1,075			n = 4,607		
	Document Type Count	% Cases w/ 1 or More Documents of Type	% Cases with 1 or More Documents of Type w/ Entities	Document Type Count	% Cases w/ 1 or More Documents of Type	% Cases with 1 or More Documents of Type w/ Entities
Progress Note	1075	100.0	90.7	4584	99.5	25.1
Lab Annotation	836	77.7	0.4			
XR Chest 2 Views Frontal Lat	606	56.3	3.9	1384	30.0	0.3
Emergency Department Report	599	55.7	33.1	1807	39.2	2.1
Letters	583	54.2	3.2	2184	47.4	0.1
Urgent Care Note	561	52.1	39.0	2004	43.5	9.7
History and Physical Report	502	46.6	20.7	1701	36.9	1.8
Surgical Pathology Report	501	46.6	0.0			
Emergency Dept Visit Note	487	45.3	9.9	1284	27.8	0.0
Radiology Annotation	463	43.0	0.5	1506	32.7	0.0
Operative Report	448	41.6	0.3			
Physician Order	438	40.7	0.9	1550	33.6	0.0
Discharge Summary	345	32.0	10.4	903	19.6	0.0
Endoscopy Procedure Report	342	31.8	2.6	1287	27.9	0.0
Consultation Report	310	28.8	8.1	975	21.1	1.2
X/Ray Report	275	25.5	0.6	687	14.9	0.0
Outside Medical Information	243	22.6	0.0			
Formal Letter	240	22.3	2.4	735	15.9	0.1
Echo Report	231	21.4	0.8			
Laboratory Report	196	18.2	0.0			
Pulmonary Function Study Rep	192	17.8	1.2			
Oximetry Report	166	15.4	1.1	330	7.1	0.0
XR Chest Frontal 1 View	152	14.1	0.2	377	8.1	0.0
Bone Mineral Density (DEXA)	139	12.9	0.7	498	10.8	0.0
Polysomnography Report	110	10.2	3.3	231	5.0	0.1
MRI Brain WO W Cnt	103	9.5	0.1			
Outpatient Clinic Report	102	9.4	2.7	376	8.1	0.0
Comprehensive Eye Exam Rep	100	9.3	3.6	434	9.4	0.0
Addendum Report	97	9.0	0.3			
CT Angio Chest	88	8.1	0.2	138	2.9	0.0
Endoscopic Report	86	8.0	1.0			
History/Physical - Pre-Op Rep	86	8.0	3.4	270	5.8	0.0
Cardiac Catheterization Report	77	7.1	0.3	212	4.6	0.0
XR Chest 1 View Portable	75	6.9	0.0			
NM Myocard SPECT Ex Rest	66	6.1	0.0			
Progress Notes - Ortho Surg	62	5.7	0.9	177	3.8	0.0

Table 5.23 WEKA Bayesian Network Model Annotated for Cumulative Counts Coded Evidence Asthma Conditioned on Named Entity Counts Asthma

Named Entity Count	Coded Evidence case count	Likelihood Coded Evidence	Conditional Probability Coded Evidence	No Coded Evidence case count	Likelihood No Coded Evidence	Conditional Probability No Coded Evidence	Likelihood NE count
0	30	0.003	0.005	5984	0.573	0.995	0.576
1	73	0.007	0.034	2053	0.196	0.966	0.203
Cum 0-1	***103***	***0.010***	***0.013***	***8037***	***0.769***	***0.987***	***0.779***
2	81	0.008	0.098	746	0.071	0.902	0.079
3	79	0.008	0.228	268	0.026	0.772	0.033
4	80	0.008	0.415	113	0.011	0.585	0.018
5	70	0.007	0.603	46	0.004	0.397	0.011
6	69	0.007	0.711	28	0.003	0.289	0.009
Cum 7+	***678***	***0.065***	***0.931***	***50***	***0.005***	***0.069***	***0.070***
7 thru 24	515	0.049	0.912	50	0.005	0.088	0.054
25 +	163	0.016	1.000	0	0.000	0.000	0.016
Totals	***1160***	***0.111***		***9288***	***0.889***		***1.000***

Table 5.24 Clinical Document Review Search Terms

ASTHMA-RELATED TERMS
WHEEZE OR WHEEZES OR WHEEZING OR DYSPNEA
OR COUGH OR COUGHS OR COUGHED OR COUGHING
OR BRONCHITIS OR ALLERGIC OR AIRWAY OR BRONCHIAL
OR BREATHING OR BREATHLESS OR BREATH OR BREATHLESSNESS
OR ALBUTEROL OR SALMETEROL OR MONTELUKAST
OR FEV OR SPIROMETRY OR EXPIRATORY
OR INHALER OR BRONCHODILATOR OR BRONCHODILATER
OR BRONCHOSPASM OR BRONCHOPROVOCATION
MEDICAL ASSESSMENT TERMS
P OR ASSESSMENT OR PLAN OR PROBLEM OR PROBLEMS OR PRESENTING
OR PRESENTS OR PRESENTED OR IMPRESSION OR HISTORY OR DIAGNOSIS
OR DIAGNOSES OR DX OR CHIEF OR COMPLAINT OR TRIAGE OR STATUS
OR SUGGESTS OR SUGGESTED OR SUGGESTING
OR TROUBLE OR TROUBLED OR REPORTS OR REPORTED
OR INDICATION OR INDICATIONS OR SUSPECT
OR SYMPTOM OR SYMPTOMS OR FINDING OR FINDINGS
OR PROBABLE OR PULMONARY OR RESPIRATORY

Appendix A

Text Mining Methods to Predict Negative Asthma Status

Approximately 11% prevalence of asthma was expected in the study population. Most of the positive cases were identified by coded evidence for asthma, leaving approximately 90% of the remaining validation set to be assessed for uncoded cases of asthma. Text mining methods were developed to identify the most likely cases for manual review. The Intermountain Enterprise Data Warehouse (EDW) stores all clinical documents for hospital and ambulatory visits. These are parsed by an Oracle text indexing program,[44] and each word is indexed. The process of searching for particular words or phrases, called ‘named entities’, is a common method described in the text mining literature.[33, 45] The frequency of each test subject’s clinical documents having named entities for asthma over a five-year period were derived for a test sample (n=10,448). A probability-based classification rule was modeled on the frequencies and coded evidence for asthma using machine-learning methods. A simple Bayes rule classifier demonstrated a 1.3% error rate in the prediction of negative asthma on 78% of the test cases, having zero or one named entity for asthma over five years. The result was validated on a convenience sample of 2% (n=160), which demonstrated an error rate of 2.4%.

The named entities to represent asthma were based on the UMLS Unified Medical Language System® (UMLS)® Metathesaurus® terms for asthma. The terms may have been more expansive but were covered by the words: ‘ASTHMA’, ‘ASTHMAS’, ‘ASTHMATICUS’, and ‘ASTHMATIC’. ‘ALLERGIC BRONCHITIS’ was the only

UMLS term not used because it was not used in the study population. Six commonly occurring misspellings were added: ‘ASTHAM’, ‘ASTHM’, ‘ASTHMAS’, ‘ASTHMATICS’, ‘ASTHAMA’, ‘ATHMA’, and ‘ASHTMA’. These terms were the ones used at least 0.01 % (.0001) as frequently as the primary term ‘ASTHMA.’ These were the ten named entities used to search the clinical text.

All clinical text documents were searched for the named entities over a five-year period (2004-2008) for a random sample of study subjects (n = 10,488). These were divided into a subset with coded evidence of asthma (n = 1,075), referred to as the ‘evidence’ sample, and a random subset of half those with no coded evidence (n = 4,607), the ‘no evidence’ sample. Coded evidence was defined as at least two ICD-9-CM ambulatory or one inpatient/ED code or an active Problem List code over the five-year period. There were 1,130 document types used. Intermountain Healthcare did not have standardized document type names, and several types could be functionally similar. The document types with at least one named entity for either sample (n = 161) were reviewed for appropriateness of document type and relative density of named entity mentions among the no evidence to evidence samples. Appropriateness was determined by whether this was a document type generally authored by providers of healthcare, who customarily document either disease status or elicit/request and record patient disease history in order to perform their clinical services. Five types were removed: ‘Message Log Notes’, ‘Nursing Notes’, ‘Lab Req – not a part of the Medical Record’, ‘Mental Status Exam’ and ‘Emergency Department Triage Note.’ Message Log Notes may be authored by nonclinical staff. The Emergency Dept Triage Note was a preliminary, admitting record and was authenticated as a longer, final Emergency Dept Visit Note.

The others were not considered typical sources of diagnostic or patient history observations and contained only 1-3 named entity mentions in total.

Progress Notes were the most frequently used document type among study subjects. All of the evidence sample and 99.5% of the no evidence sample had at least one Progress Note. Progress Notes usually documented an ambulatory clinic visit. The average number of Progress Notes per case during this period was 22 for the evidence sample and 15 for the no evidence sample. The average number of Progress Notes with named entities per case was 10 for the evidence sample and 0 for the no evidence sample. Multiple Progress Notes per case gave the most opportunity for the expected mentions among negative cases: hypothetical statements about the disease such as "at risk for", historical diagnosis statements, references about someone else and the disease such as in "family history of", or a negation of the disease diagnosis.[37, 38] Among no evidence cases, 25.1% had at least one named entity mention in a Progress Note (Table 5.22). This rate was then used as a heuristic measure of the expected density of mentions among no evidence cases for a document type. Samples of cases with document types with a higher density than 25% among no evidence cases were reviewed. Nine document types contained templated text or other reasons for a higher density of mentions of asthma among noncases and were removed. There were finally 147 document types used for the subject-level counts of documents and named entities. The document types occurring for more than 5% of the evidence sample are shown in Table 5.22, with the frequencies of cases having at least one instance of the document type and among those, the frequencies of cases having at least one named entity for asthma in that document type.

Next the individual asthma named entity counts were analyzed using attribute selection and classification methods in WEKA.[46] Two separate random samples (n = ~10,000) of cases from the study population were generated. The two samples were used for both training and testing of the other. Document counts and counts of documents with a named entity over the five year period, 2004-2008, were computed for each case over the 147 document types. Only one count per document type per day was allowed because there were occasional duplicate or redundant document instances in the EDW. One count of the total number of documents and one count for documents containing at least one named entity for asthma were stored in the study database. The count of documents with named entities divided by the count of documents was computed for each case as a density function and populated for training cases in the study database. The class attribute was coded evidence of disease versus no coded evidence. The four attributes were analyzed in WEKA using the Bayesian network classifier. The Bayesian network classifier was selected for its ability to learn the best structure and expose the probability tables.[46] Attribute selection methods in WEKA consistently agreed upon the named entity count as the best attribute, and the wrapped Bayesian network selection method preferred it as a single attribute. The attributes in the model were reduced to the named entity counts and the class (outcome). The final model was essentially a simple Bayes rule classifier.

The WEKA Bayesian network model was recreated in Excel to show the most predictive discretizations of the named entity counts. The model is shown in Table 5.23. The classifier demonstrated a 1.3% error rate in the prediction of negative asthma status on 78% of the cases, having zero or one named entity for asthma. The result was validated on a convenience sample of 2% randomly selected cases with zero (n=100) or one (n=60) named entities. Of the zero named entities sample, there was one case documented as ‘azthma’. Of the one named entity sample, there were four cases with statements of possible asthma. Asthma status was not further ascertained in the validation. The error rate was 2.4%.

Accordingly, for the study composite reference standard, cases requiring resolution by review of the clinical documentation were determined negative if they had a named entity count of zero or one. Cases with 2+ named entities were manually reviewed if they were not already determined positive by the coded evidence in the composite reference standard. The simplified version of the negative asthma status prediction model is shown in Table 5.8.

Appendix B

Clinical Document Review Procedures

The Intermountain Enterprise Data Warehouse (EDW) stores all clinical documents for hospital and ambulatory visits. These are parsed by an Oracle text indexing program,[44] which stores each word in an index. The process of searching the indexed words for 'named entities' is a common method described in the text mining literature.[33, 45] Three groups of named entities were defined for this study for different purposes. The first was used for both clinical document review and development of the negative asthma status prediction model. The other two were defined for clinical document review only.

The first group was asthma terms. The named entities used to represent asthma were based on the Unified Medical Language System® (UMLS)® Metathesaurus®. The terms were more expansive but were covered by the words: 'ASTHMA', 'ASTHMAS', 'ASTHMATICUS', and 'ASTHMATIC'. 'ALLERGIC BRONCHITIS' was the only UMLS term not used because was not used in the study population. Six commonly occurring misspellings were added: 'ASTHAM', 'ASTHM', 'ASTHMAS', 'ASTHMATICS', 'ASTHAMA', 'ATHMA', and 'ASHTMA'.

The second group consisted of terms that describe symptoms, medications, and comorbidities associated with asthma. This group was generated from the literature on asthma symptoms, diagnosis and treatment. The asthma-related terms are shown in Table 5.24. The third group consisted of terms used by clinicians in their documentation to state assessments, impressions, diagnoses, plans and included almost all direct medical care documents. These terms were found by iterative queries against a set of known clinical documents while adding to the search terms from documents not found. The medical assessment terms are shown in Table 5.24.

Reports were developed to query and display variable length snippets of text around named entity search terms as well as the full text with the search terms highlighted for each clinical document selected. This enabled fast, direct access to the documents and the specific statements in the documents containing the named entities. The reports were easily modified and ran in a public domain SQL query application. This enabled the nurse reviewers to perform the reviews on their desktop computers with no protected health information removed or stored from the EDW repository.

A 13% random sample of the 784 validation set subjects having zero or one clinical document with an asthma term and not resolved by the coded evidence were reviewed. One experienced nurse volunteer used the asthma-related search terms to attempt to find clinical documents for 60 subjects with zero asthma terms. Two subjects had documentation suggestive of asthma but no definitive statement or further evidence. For those with an unconfirmed suggestion, a search was done using the medical assessment terms to display diagnostic statements. In this manner, the suspicious subject’s full documentation was efficiently displayed in reverse chronologic order. It gave focus to the assessment statements as one line listings and the ability to toggle to the complete document with highlighted search terms. Forty-two subjects with one clinical document with an asthma term were similarly reviewed. For these, the first step was to search and review the asthma terms in the one clinical document. If these statements implied a negative asthma status, the subject’s review was negative. If asthma was suggested, the review process continued with the asthma-related terms as described above. One missed asthma case was ascertained from an asthma diagnosis statement and a history of asthma medications.

All validation subjects with more than one clinical document having asthma terms and not resolved by the coded evidence (n=101) were reviewed. In the first stage, the asthma terms were searched and reviewed as described above. If these statements implied a negative asthma status, the subject's review was negative (n=60). If asthma was suggested (n=41), the relevant document snippets were copied to a temporary spreadsheet, carefully excluding any protected health information as snippets were transferred. The asthma-related search terms followed by the medical assessment search terms were used as needed to include as definitive an asthma picture as possible. These de-identified case abstracts were shared with one asthma medical expert. He labeled each one as definite or probable asthma versus negative or unlikely asthma versus asthma by history only. The criteria used to decide a probable asthma status are shown in Table 5.5. Five borderline case abstracts were reviewed by a second asthma medical expert. One status was changed.

References

1. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol. 2009 Aug;62(8):797-806.

2. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess. 2007 Dec;11(50):iii, ix-51.

3. Alonzo TA, Pepe MS. Assessing the accuracy of a new diagnostic test when a gold standard does not exist. UW Biostatistics Working Paper Series, University of Washington: The Berkeley Electronic Press; 1998.

4. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. Stat Med. 1999 Nov 30;18(22):2987-3003.

5. Staquet M, Rozencweig M, Lee YJ, Muggia FM. Methodology for the assessment of new dichotomous diagnostic tests. Journal of Chronic Diseases. 1981;34(12):599-610.

6. Mathews WC, Cachay ER, Caperna J, Sitapati A, Cosman B, Abramson I. Estimating the accuracy of anal cytology in the presence of an imperfect reference standard. PLoS ONE.5(8):e12284.

7. Expert panel report 3 (epr3): Guidelines for the diagnosis and management of asthma. National Heart Lung and Blood Institute, National Institutes of Health; 2007.

8. Izquierdo JL, Martin A, de Lucas P, Rodriguez-Gonzalez-Moro JM, Almonacid C, Paravisini A. Misdiagnosis of patients receiving inhaled therapies in primary care. Int J Chron Obstruct Pulmon Dis. 2010;5:241-9.

9. Kaplan AG, Balter MS, Bell AD, Kim H, McIvor RA. Diagnosis of asthma in adults. CMAJ. 2009 November 10, 2009;181(10):E210-20.

10. Pacheco JA, Avila PC, Thompson JA, et al. A highly specific algorithm for identifying asthma cases and controls for genome-wide association studies. AMIA Annu Symp Proc. 2009;2009:497-501.

11. Buffels J, Degryse J, Liistro G. Diagnostic certainty, co-morbidity and medication in a primary care population with presumed airway obstruction: The didasco2 study. Primary Care Respiratory Journal. 2009;18(1):34-40.

12. Levy ML. Guideline-defined asthma control: A challenge for primary care. European Respiratory Journal. 2008;31(2):229-31.

13. Lucas AEM, Smeenk F, Smeele IJ, van Schayck CP. Overtreatment with inhaled corticosteroids and diagnostic problems in primary care patients, an exploratory study. Family Practice. 2008;25(2):86-91.

14. Marklund B, Tunsater A, Bengtsson C. How often is the diagnosis bronchial asthma correct? Fam Pract. 1999 Apr;16(2):112-6.

15. Tinkelman DG, Price DB, Nordyke RJ, Halbert RJ. Misdiagnosis of copd and asthma in primary care patients 40 years of age and over. Journal of Asthma. 2006;43(1):75 - 80.

16. Chin ES. Pediatrics, reactive airway disease. 2010 Apr 6, 2010 [cited Nov 1, 2010]; Available from: http://emedicine.medscape.com/article/800119-overview

17. Fahy JV, O'Byrne PM. "Reactive airways disease". A lazy term of uncertain meaning that should be abandoned. Am J Respir Crit Care Med. 2001 Mar;163(4):822-3.

18. To T, Dell S, Dick PT, et al. Case verification of children with asthma in ontario. Pediatr Allergy Immunol. 2006 Feb;17(1):69-76.

19. Prosser RJ, Carleton BC, Smith MA. Identifying persons with treated asthma using administrative data via latent class modelling. Health Serv Res. 2008 Apr;43(2):733-54.

20. Schatz M, Nakahiro R, Jones CH, Roth RM, Joshua A, Petitti D. Asthma population management: Development and validation of a practical 3-level risk stratification scheme. Am J Manag Care. 2004 Jan;10(1):25-32.

21. Icd-9-cm official guidelines for coding and reporting The Centers for Medicare and Medicaid Services (CMS) and the National Center for Health Statistics (NCHS); 2009.

22. Blais L, Lemiere C, Menzies D, Berbiche D. Validity of asthma diagnoses recorded in the medical services database of quebec. Pharmacoepidemiol Drug Saf. 2006 Apr;15(4):245-52.

23. Lix L, Yogendran M, Burchill C, et al. Defining and validating chronic diseases: An administative data approach. Winnipeg: Manitoba Centre for Health Policy; 2006 July 2006.

24. Meystre SM, Haug PJ. Randomized controlled trial of an automated problem list with improved sensitivity. Int J Med Inform. 2008 Sep;77(9):602-12.

25. Szeto HC, Coleman RK, Gholami P, Hoffman BB, Goldstein MK. Accuracy of computerized outpatient diagnoses in a veterans affairs general medicine clinic. Am J Manag Care. 2002 Jan;8(1):37-43.

26. Tang PC, Ralston M, Arrigotti MF, Qureshi L, Graham J. Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: Implications for performance measures. J Am Med Inform Assoc. 2007 Jan-Feb;14(1):10-5.

27. Weber V, Bloom F, Pierdon S, Wood C. Employing the electronic health record to improve diabetes care: A multifaceted intervention in an integrated delivery system. J Gen Intern Med. 2008 Apr;23(4):379-82.

28. Wright A, Ricciardi TN, Zwick M. Application of information-theoretic data mining techniques in a national ambulatory practice outcomes research network. AMIA Annu Symp Proc. 2005:829-33.

29. Masys D. Extracting phenotypes from ehrs: The emerge consortium experience. 2009 Information Technology Roundtable Meeting. Washington, D.C.: The Clinical Research Forum; 2009.

30. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. J Clin Epidemiol. 2004 Feb;57(2):131-41.

31. Twiggs JE, Fifield J, Apter AJ, Jackson EA, Cushman RA. Stratifying medical and pharmaceutical administrative claims as a method to identify pediatric asthma patients in a medicaid managed care organization. J Clin Epidemiol. 2002 Sep;55(9):938-44.

32. Vollmer WM, O'Connor EA, Heumann M, et al. Searching multiple clinical information systems for longer time periods found more prevalent cases of asthma. J Clin Epidemiol. 2004 Apr;57(4):392-7.

33. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform. 2005 Mar;6(1):57-71.

34. National text mining centre: Text mining. Joint Information Systems Committee (JISC); 2008.

35. Ruch P, Baud R, Geissbuhler A. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. Artif Intell Med. 2003 Sep-Oct;29(1-2):169-84.

36. Taft M, Krishnan R, Hornick M, et al. Oracle data mining concepts. 2005 [cited Nov 1, 2010]; Available from: http://download.oracle.com/docs/html/B14339_01/title.htm

37. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001 Oct;34(5):301-10.

38. Harkema H, Dowling JN, Thornblade T, Chapman WW. Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform. 2009 Oct;42(5):839-51.

39. Newcombe RG. Two-sided confidence intervals for the single proportion: Comparison of seven methods. Stat Med. 1998 Apr 30;17(8):857-72.

40. Newcombe RG. Interval estimation for the difference between independent proportions: Comparison of eleven methods. Stat Med. 1998 Apr 30;17(8):873-90.

41. Miettinen O, Nurminen M. Comparative analysis of two rates. Stat Med. 1985 Apr-Jun;4(2):213-26.

42. Li JT, Pearlman DS, Nicklas RA, et al. Algorithm for the diagnosis and management of asthma: A practice parameter update: Joint task force on practice parameters, representing the american academy of allergy, asthma and immunology, the american college of allergy, asthma and immunology, and the joint council of allergy, asthma and immunology. Ann Allergy Asthma Immunol. 1998 Nov;81(5 Pt 1):415-20.

43. Lurie IZ, Dunlop DD, Manheim LM. Trends in out-of-pocket medical care expenditures for medicare-age adults with arthritis between 1998 and 2004. Arthritis Rheum. 2008 Aug;58(8):2236-40.

44. Ford R. Oracle text. 2007 [cited Aug, 2009]; Available from: http://www.oracle.com/technology/products/text/pdf/11goracletexttwp.pdf

45. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: A review of recent research. Yearb Med Inform. 2008:128-44.

46. Witten IH, Frank E. Data mining: Practical machine learning tools and techniques with java implementations: Morgan Kaufmann Publishers; 2000.

CHAPTER 6

SUMMARY

In this dissertation, the purposes and design of the framework were presented. The framework was evaluated on two diseases. The framework-generated rules to identify diabetes were similar to the rules generated by domain experts but also added useful knowledge to refine them. Rules to identify asthma added to the sensitivity of a validated algorithm based on ICD-9-CM codes. The framework required no domain knowledge to find these patterns. The machine-learned patterns quantified specific EHR data uniquely associated with a disease. This knowledge is best used in conjunction with the specialized knowledge of domain experts to refine and strengthen the identification of cases.

Observations that have not been identified by experts in existing disease-identification algorithms were exposed. In the diabetes identification rules, capillary blood glucose testing by glucometers was performed more often, was abnormal more often, and served as a better observation type to *identify* patients with diabetes in the EHR than glucose measured in venous blood in a clinical laboratory. Laboratory-based glucose values are the standard for diagnosis of diabetes, have been proven more accurate than glucometer results,[1] and consequently were specified as one of the criteria to identify diabetes in the expert-based algorithms reviewed.

There are many potential predictive combinations in the EHR data. Combinations of attributes are rarely seen in expert-generated algorithms. There were many combinations with the bronchodilator, albuterol, in the asthma rules. This was because albuterol was a frequent and associated medication, but it was not strong enough alone to form an accurate rule. Albuterol is used for other bronchopulmonary problems. Albuterol and an inhaled glucocorticoid are found together in expert-generated rules to identify asthma. However, albuterol and female gender was a novel combination, even though it was known that more women than men are seen for adult asthma.

The rules may refine the observations within categories. For example, instead of a list of all antihyperglycemic drugs, the rules identified specific chemical/ therapeutic classes as separate rules. Some were stronger than others. Insulin was very strong. Metformin was not strong enough to form a single-attribute rule. It was known that Metformin can have false positives, yet there are algorithms to identify diabetes that include all antihyperglycemic medications. Since treatments constantly change, the framework process was useful in identifying exactly which medications were in common usage during the time period of interest. In fact, it identified them so well it picked up a decreased frequency from one year to the next when a particular diabetes drug was the subject of an FDA warning for potential adverse events.

There were three assumptions underlying the design for the cohort amplification framework, and all three were validated in this work:

- Patterns learned from exemplars drawn from known cases may be used to identify other unknown cases. Some bias must be anticipated since the known cases were identified in some manner, and others were not. In this dissertation, the proof of concept was demonstrated by other asthma cases identified from known cases from the coded Problem List in the electronic health record.

- Knowledge discovery methods over a common set of EHR data could expose patterns for multiple diseases and conditions. A highly abstracted set of candidate EHR data, modeled on the derived data types used in some existing disease identification algorithms[2-4] and drawn from the standard data categories required for certification of an EHR, was shown to generate rules consistent with disease-specific processes of care.
- The transaction records of medical care process and documentation could provide patterns from the data in the EHR that may be used to identify particular disease cases. Patterns or rules learned by the computer from the candidate EHR data used were shown to identify unseen cases of asthma in the EHR.

The framework classification rules can be analyzed for other purposes. They do not have to be applied in prediction. Interesting trends were seen in the rules when comparison groups were created, such as two different clinical specialties. There were aggregate differences in the choice of medications. It could be seen that practice patterns were different.

Finally, the rules may expose interesting associations that were not expected. The main objective of this work was to apply database knowledge discovery methods to generate classification rules. Comorbidities were found among the asthma cases identified by the framework rules that were not described in the clinical literature. Their physiologic connection had been described. The associations were with other inflammatory diseases of the joints and other connective tissue.

Why Associative Classification?

Associative classification (AC) has not been widely used in health care data. Association rule mining (ARM) was reported more often in the health care literature. In Chapter 2, some examples of health care applications of AC and ARM were presented. It was pointed out that the use of these approaches is growing, as are machine learning methods in general, in health care data. Association mining was developed for retail sales problems and has been informally called market basket analysis. It is widely used for business purposes and has continued research, development and improvement as reported mainly in the computer science literature. ARM was developed to find patterns among broad and inter-related transaction data, directly from databases without domain expertise.

At the onset of this research, various classification methods were investigated. Associative classification accommodated the broad but sparse data best. Many other methods could not process data with so many missing values. In the framework, missing values were considered unpopulated EHR observations and participated in patterns by their absence. Naïve Bayes accepted the sparse data but was not able to predict as accurately as AC. This assessment was based on raw results from WEKA prior to the improvement of accuracy developed in the framework. Rule-learners, such as Ripper,[5] found similar rules but were not transparent nor popularly used and studied like AC. ARM, as implemented in WEKA, provided the interim metrics of the Apriori algorithm.

This enabled access to the ARM frequent item sets from which downstream pruning, testing and knowledge discovery were accomplished.

Contributions to Associative Classification Methods

Ranking and pruning methods were introduced to improve the generality of the rules as generated using classic ranking and pruning methods. These methods not only improved the generality but also the accuracy of the rules. There was a serendipitous benefit in these methods, which used a final frequency sort, to reduce the final rule set to the highest concept level of attributes participating in multiple concept levels.

The framework used balanced sample sizes. This enabled interestingness metrics that discriminated the data characteristics of the disease exemplars no matter what their prevalence in the EHR. These were normalized metrics used to focus the rules to identify one class only. By design, one set of exemplars represented a disease or condition of interest. The other represented controls without the condition. Since the goal was to generate rules to identify cases with the condition, it was practical to leverage one-class rules. The evaluation of the merit of the rules was based on sensitivity and specificity. Sensitivity and specificity are not affected by prevalence.

Significance to Biomedical Informatics

The development and validation of the framework are significant in the field of biomedical informatics because they demonstrate a successful application of machine learning in the electronic health record. Development of the framework included modification of associative classification methods to address the unique data content in the EHR. Data mining methods have not been applied to their potential in the EHR. This work suggested that standard methods developed for other industries might need to be optimized for the health care data environment. Further, the methods support high-throughput phenotypic cohort identification for genomic research. The framework does not generate definitive phenotyping algorithms but exposes and quantifies generalized EHR data toward that purpose.

Opportunities for Further Research

The framework should be evaluated in other diseases and EHR settings. Diabetes and asthma are well characterized as medical diagnoses, with standard diagnostic and treatment patterns. The current research was conducted in a health care delivery environment with mature EHR systems and programmatic efforts to train and support providers to use the EHR. Care documentation for diabetes and asthma was also focused on in recent years for institutional clinical quality improvement goals. Evaluation of classifiers for medical conditions where standards of care are not well defined would further test the contributions of the framework. It would also be useful to test the reliability of rules for diabetes and asthma in other health care settings. Are the rules overfit to the care setting or more general across settings? Is there value in seeking rules that generalize across settings or is it reasonable to use rules suited to each setting?

The framework exposes the frequency and strength of associations among EHR data elements to build the classifiers, so the differences in raw data among settings can be explored.   It can be applied to the problem of testing differences between settings, provider groups or time periods. A study is underway to test differences in EHR observations between two large provider organizations, using the framework to identify observations associated with type 2 diabetes mellitus as a step in attribute selection for further statistical analysis.

The framework approach can be extended to attributes from other databases, including population and behavior data that may or may not be linked to EHR data. In addition, the broad and highly abstract view of EHR data used in this research can be focused for particular medical conditions, including data that are more detailed. Association rule mining was called a "brute force" method of knowledge discovery because it discovers all associations in the attributes presented, then applies various metrics and operator-provided thresholds to measure and select the *interesting* ones. Therefore, it provides an opportunity to explore very general problem domains. Interesting associations may be refined and constrained to particular focus areas using these methods. The approach can also be used with other knowledge discovery methods that model relations that are more complex but usually require a more focused problem domain.

This research demonstrated success of a high throughput, generalized approach to learn classifiers for two medical conditions from the EHR data of exemplars of those conditions. Its potential application to identify, or amplify, health related cohorts was posed as an improvement in the efficiency of conducting biomedical research. These methods can also be applied to improve identification of cases for chronic disease registries. Registries support care management activities as well as health services research toward improved health care delivery and patient outcomes. The cohort amplification rules are well suited to this task in theory. Specificity and sensitivity are acceptance thresholds set in the rules generation process. This enables a range of potential accuracy. In the population of registries, cases may be identified on a scale of uncertainty. Future research includes the development and evaluation of the framework for this purpose.

Limitations of this Research

Existence/nonexistence of EHR observations over an arbitrary time period was the data type of all candidate data used in the framework. This gave a broad but shallow pattern search. This was intended as a first look at patterns for exemplar conditions, given data that are generally populated in the EHR. The intent was to apply the framework to better understand the data content as well as to find useful patterns in this superficial view. With a better understanding of actual patterns in stored data for particular conditions, a deeper, condition-specific data set might be designed for further machine learning approaches. Useful patterns for disease case identification were found in the two conditions studied in this work, diabetes and asthma. In other conditions, the shallow data patterns may not.

Administrative codes (ICD-9-CM and CPT) were used in the framework. There are known problems with their reliability and validity as addressed in Chapter 1. At this time, the administrative codes in the EHR are the most comprehensive disease documentation available. The ICD-9-CM codes for the target conditions were removed from the candidate data because they would have dominated the rules. They were used to associate comorbidities. In the short term, while ICD-9-CM codes are widely used for cohort identification, the framework rules may augment or validate these algorithms with additional rules. As better data become available, for example by use of NLP to extract coded clinical data, the candidate data for the framework can be modified.

The query to gather EHR data can be streamlined to better automate the current process. The current process generated attributes that had to be manually removed. Most of the superfluous attributes were administrative in nature, redundant or simply useless to the problem at hand. An example of the latter is common medical specialties visited frequently by all patients. This is to be expected when mining patterns directly from a transaction database. The least number of candidate attributes is best since computer memory consumption is large for Apriori rule mining.

A minimum of one thousand training records from disease and control exemplars was necessary to find reliable patterns. This number of exemplars may be difficult to find. If patterns were proven reliable across institutions, training data could be shared for less prevalent training cases.

References

1. Rush E, Crook N, Simmons D. Point-of-care testing as a tool for screening for diabetes and pre-diabetes. Diabet Med. 2008 Sep;25(9):1070-5.

2. Hedis and quality measurement. 2011 [cited Feb 1, 2011]; Available from: http://www.ncqa.org/tabid/59/default.aspx

3. Quality measures. 2010 Sept 24, 2010 [cited Feb 1, 2011]; Available from: https://www.cms.gov/QualityMeasures/

4. Library of phenotype algorithms. 2011 [cited Feb 1, 2011]; Available from: https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Library_of_Phenotype_Algorithms

5. Witten IH, Frank E. Data mining: Practical machine learning tools and techniques with java implementations: Morgan Kaufmann Publishers; 2000.