ENABLING CLINICALLY BASED KNOWLEDGE DISCOVERY IN PHARMACY CLAIMS DATA: AN APPLICATION IN BIOINFORMATICS

Jason Jones*, Doug Stahl†, Michael Nichol‡ and Stanley P. Azen§

ABSTRACT

This paper describes the development, application, and evaluation of a set of methods for transforming standard pharmacy claims data into a clinically relevant database that can facilitate healthcare research. Prescription claims data represent relatively inexpensive and largely unexploited exploratory ground for understanding the relationships between prescription treatments and their healthcare and cost outcomes.

A web-based, graphical interface was developed to solicit clinical expert opinions about how claims should be combined into prescription treatments. A classification tree methodology was then applied to the database in an attempt to induce expert decisions based on a flexible set of predictor variables generated directly from the prescription claims.

Two different classification tree approaches and four versions of the predictor variable sets (PVSs) were compared with each other and with a fixed heuristic for data transformation in a sample of 11,654 expert reviewed claim pairs. The model-based classification rules significantly outperformed the simple rule when claim pairs were comprised of different drugs and performed as well as the simple rule when the drugs were the same.

The best combination of classification tree approach and PVS was used to generate a set of rules that was subsequently applied to a larger dataset and used to generate prescription treatment episodes. A sample analysis was conducted using the output database to specify inclusion/exclusion criteria, group assignment, stratification, and outcomes such as treatment discontinuation. Both visual and formal techniques were used in a way that would be commonly used in an outcomes or pharmacoeconomic research endeavor.

1. Introduction

In clinical medicine it is essential to evaluate benefits and risks associated with new treatments that are developed for a given disease. Randomized controlled trials (RCTs) are designed to establish causal relationships between treatments and their risks and benefits (Pocock, 1983; ISPOR, 1998). The primary drawback to RCTs is that the level of control required produces an artificial environment, leaving questions about how treatments will
perform in the real world (ISPOR, 1998; http://www.ajmc.com/outcomes.html). With the developments of pharmacoconomics and outcomes research, changes in healthcare delivery, and advances in computing, researchers now have the ability to evaluate the benefits and risks associated with treatments in a natural setting (Dillon, 1999; Navarro et al., 1999).

Organizations charged with delivering and managing healthcare services (Managed Care Organizations-MCOs) maintain massive data repositories (large, well-defined, central databases) containing detailed information related to services and treatments provided to patients (Navarro et al., 1999). The primary obstacle in conducting clinically oriented research with these repositories is that the data are collected for the purpose of managing financial risk (Navarro et al., 1999). These repositories can be grouped into three categories: medical service data, pharmacy claims, and ancillary data. While vast and detailed, these repositories tend to be less useful for clinically oriented research in practice than one might hope. Substantial transformation is required to mine (apply statistical methods to produce knowledge) this wealth of untapped information that could be used to answer questions related to outcomes in actual clinical practice.

1.1. Medical service data

Medical services (e.g., check-ups or surgeries) are provided to patients in physician offices, clinics, and hospitals. Each service involves procedures (e.g., x-raying an arm and putting the arm in a cast) and diagnoses (e.g., broken arm). In order to be reimbursed in a Fee for Service (FFS) environment, the provider must document what was done and the reason it was done, using standard forms and codes. This claim is then processed by the MCO and stored in a data repository such that (at the very least) the provider, patient, procedure, diagnosis, and date of the medical service can be determined. Standard coding systems exist for both procedures (CPT codes) and diagnoses (ICD codes) so that retrieving and categorizing claims is efficient (Navarro et al., 1999). Methods exist for handling cases where multiple diagnoses or procedures are involved in a single visit or when a single visit lasts for extended periods of time (e.g., a hospital stay).

1.2. Pharmacy claims data

A pharmacy claim is generated every time a member fills a prescription for a medication. The National Council for Prescription Drug Programs (NCPDP), which is recognized by the American National Standards Institute (ANSI), has developed the process and content for transferring pharmacy claims from a pharmacy to the PBM (Navarro et al., 1999). Minimally, the data transfer includes the prescriber, patient, drug identifier, prescription fill date, dose, quantity, and prescription duration.

The system was originally designed to enforce formularies (i.e., lists of approved drugs for a given health plan) and to make sure that pharmacies did not provide medications to patients who were not eligible for them (Navarro et al., 1999). Over time, the system has been enhanced to prevent adverse events such as drug-drug interactions that could be harmful to the patient. Unless a non-prescription medication is covered by the health plan, non-prescription medications do not generate pharmacy claims records.

1.3. Ancillary data

Besides the information about services that the MCO administers, information relating to the claims administration itself is also stored. There are differences in the types of ancillary data maintained by different organizations, but the following types of information are often included:
Knowledge Discovery in Pharmacy Claims Data

- Physician, hospital, pharmacy, and member names and addresses.
- Physician accreditation and contract information.
- Employer group contract information.
- Records specifying which benefits the members were eligible for and at what times.
- Information explaining how to decode procedure, diagnosis, and drug codes.

Various departments inside and outside the MCO maintain this information. For instance, many MCOs have their own accreditation and contracting departments for physicians. However, procedure, diagnosis, and drug codes are usually developed and supplied by third parties such as the American Medical Association (AMA) for procedure codes or the FDA for drug codes (ISPOR, 1998).

1.4. Medical data repositories in practice

The quality of medical data repositories varies across MCOs. Expenditures on information technology have traditionally been lower in healthcare than in other fields such as banking (Romza and Black, 1999). Like other industries, the health care industry has focused first on collecting and maintaining data directly related to business operations or finances (Navarro et al., 1999). This often leaves medical data repositories lacking both accuracy and completeness, making them difficult to use for clinically oriented research.

1.5. Data incompleteness and errors

Even in the FFS environment, providers were paid based on procedures and not diagnoses. Therefore, if patients had conditions that were not directly related to the primary, reimbursable service, it is likely that the diagnosis would not be included on the claim (Motheral, 1997). Furthermore, to maximize payments, it is possible that physicians might modify the procedure coding or diagnoses slightly.

The quality of medical services data has become worse as capitation (fixed provider reimbursement based on member count rather than services rendered) has become more common. Submitted claims often represent a biased sample to improve the provider’s chance of renegotiating improved capitation contracts. When claims are submitted, they are often inaccurate and go unchecked because they are not considered to be of critical financial importance to the MCO in the way that FFS claims were. There is evidence that MCOs have begun to place higher and broader value on their data repositories. Some MCOs have begun providing incentives for providers to submit claims for medical services and some have begun attempts to collect other clinically relevant data such as for laboratory results.

In contrast, pharmacy claims are highly regulated and well maintained. Of the 2.8 billion transactions processed by pharmacies in 1998, 95% were processed immediately and electronically (Romza and Black, 1999). This means that the pharmacy claims data tend to be the highest quality and most complete data in the healthcare industry (Navarro et al., 1999).

1.6. Clinical disconnect

While the completeness and accuracy of the MCO pharmacy claims is very good, they are not immediately useful for pharmacoeconomic or outcomes research. The purpose of storing the claims was to facilitate claim adjudication and overall financial risk management. It was only after standards were in place that people started to consider the use of these data for clinically oriented research. Because pharmacoeconomic and outcomes research is clinically oriented, not having access to clinically oriented constructs in the pharmacy claims
data presents a significant obstacle to mining these data to evaluate things such as treatment
effectiveness (Navarro et al., 1999; Romza and Black, 1999). PBMs and MCOs would benefit
having pharmacoeconomic and outcomes research studies on which to base their formulary
decisions, however they are rarely able to use their data to support their decisions (Dillon,
1999).

The purpose of this paper is to present, apply, and evaluate a set of methods for
transforming data repositories maintained by managed care organizations and removing the
barriers preventing the efficient mining of these repositories.

2. Theoretical framework for proposed methods

Many fields of business and science have accumulated large, complex, information-rich
datasets as a result of improved data collection technologies and decreased data storage
costs (Cherkassky and Mulier, 1998; Raghavan et al., 1998). As a result, the gap between
the ability to collect data and the ability to analyze, summarize, and extract knowledge
from it is rapidly widening (Fayyad et al., 1996a; Brodley et al., 1999). While database
technology has provided basic tools for efficient storage and retrieval of large datasets, the
issue of how to help humans analyze and understand them remains a difficult and unsolved
problem (Fayyad et al., 1996b). The primary emphasis in data mining is to exploit such
readily available data to derive models and estimate useful relationships. Most data mining
methods are based on concepts from machine learning, pattern recognition, and statistics.
Data mining and machine learning are both components of a more comprehensive knowledge
discovery process that is described by Fayyad et al. (1996c) as “the nontrivial process of
identifying valid, novel, potentially useful, and ultimately understandable patterns in data”.

Fayyad’s framework for Knowledge Discovery in Databases (KDD) provided a founda-
tion for the analysis of pharmacy claims data described in this study. The main components
of the KDD framework are shown in Table 1. Although the table shows the phases as sequen-
tial, it is important to note that KDD research often proceeds through multiple iterations.

<table>
<thead>
<tr>
<th>KDD Phase</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Developing an understanding of the application domain, relevant prior knowledge, and end user goals.</td>
</tr>
<tr>
<td>2</td>
<td>Creating a target data set on which discovery is to be performed</td>
</tr>
<tr>
<td>3</td>
<td>Data cleansing and preprocessing</td>
</tr>
<tr>
<td>4</td>
<td>Data reduction and projection</td>
</tr>
<tr>
<td>5</td>
<td>Selecting the data mining task (classification, regression, clustering, etc.)</td>
</tr>
<tr>
<td>6</td>
<td>Selecting the data mining algorithms and methods to be applied</td>
</tr>
<tr>
<td>7</td>
<td>Data mining</td>
</tr>
<tr>
<td>8</td>
<td>Interpreting mined patterns, possible return to previous steps for further iteration</td>
</tr>
<tr>
<td>9</td>
<td>Consolidating, documenting, and presenting discovered knowledge</td>
</tr>
</tbody>
</table>

3. Methods

Figure 1 presents a basic overview of the components and process that were used to
statistically derive rules from a sample of expert human decisions for converting pharmacy
Knowledge Discovery in Pharmacy Claims Data

claims into clinical constructs. This paper will focus on the combination of prescriptions claims into clinical treatments. Specifically, the problem was to determine how to handle claim pairs with overlapping dates. For instance, should the two claims be combined into a single treatment, or does the presence of the second claim signal the early termination of the first treatment?

The transformation process is closely tied to the KDD framework described above (Fayyad et al., 1996c). The Data Extraction process (see Figure 1) is applied to the original Typical Rx Claims Database, resulting in a select, Analytic Database. Further selection is applied to Sampling for Expert Opinion. Clinical domain expertise is applied via Expert Processing and again, coupled with pattern interpretation, in Expert Refinement, after data mining is conducted in the Modeling stage. Formal Model Assessment is accomplished by comparing the Set of Rules for Processing created in the Code Generation process to the expert classification. Clinical domain expertise is utilized again in Variable Modification to modify the transformed data in the Analytic Database. Finally, Application Processing based on the knowledge gained for transforming the data is utilized to generate the Clinical Constructs, a new database for clinical knowledge discovery via pharmacoconomic and outcomes research.

The Analytic Database contains the predictor variable set (PVS), a set of variables derived from the prescription claims and used as model inputs. In making decisions about

Fig. 1: Basic process diagram
whether or not to combine the members of the claim pair, human experts take into account not just the two claims themselves but the context in which the claims fall. In this case, the context was often a complex pattern of prescribing that could span years. The predictor variable set is summarized in Table 2 and was an attempt to represent to a computer the constructs that the human expert was creating and using to make her/his decisions.

Table 2: Basic predictor variable set (PVS) structure

<table>
<thead>
<tr>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Pair Id Information</td>
<td>Claim Id 1</td>
</tr>
<tr>
<td>2 Basic Rx Information</td>
<td>Drug Name</td>
</tr>
<tr>
<td>3 Overlap</td>
<td>IniDiff = (Claim 2 Start)−(Claim 1 Start)</td>
</tr>
<tr>
<td>4 History*</td>
<td>Pri1Cnt = # Prior Drug 1 Prescriptions</td>
</tr>
<tr>
<td>5 Future*</td>
<td>Pst1Cnt = # Post Drug 1 Prescriptions</td>
</tr>
<tr>
<td>6 Miscellaneous</td>
<td>DrugSame = Drug 1 same as Drug 2</td>
</tr>
</tbody>
</table>

* Some variables could be null.

Four PVS configurations and two classification tree algorithms — CART (Breiman et al., 1984) and QUEST (Loh and Shih, 1997) — were compared to a heuristic used in prior outcomes research (FixedX). FixedX combined the members of the claim pair into a single treatment if the claims started within X days of each other; otherwise, the first claim was terminated early. CARTi/QUESTi use the same variable (IniDiff) used in FixedX. However, rather than setting X a priori, IniDiff is used as the only variable in the model to derive the optimal split points for IniDiff. CARTs is CARTi plus a variable used to describe whether a better match for the members of the claim pair could be found (BMatch). CARTp/QUESTp use the full PVS except BMatch. CARTf/QUESTf use the full PVS.

Data were stored (components 1, 3, and 12) using standard SQL-89 compliant database management systems. Expert clinical input for classifying the claim pairs was received via a web-based application that presented textual and graphical claims histories. CART and QUEST were both implemented in SPSS AnswerTree 2.0. The C-statistic was used to compare the predictive accuracy of modeling algorithms and different PVS configurations (Hanley and McNeil, 1982; Bamber, 1975).

The best classification tree and PVS combination (CARTf) was then applied to the larger set of prescription claims and a sample outcomes analysis was conducted in the area of antidepressant therapy. Treatment discontinuation, events, and costs were evaluated for 6,577 patients starting antidepressant treatment between 1996 and 1998 with a class of antidepressants known as selective serotonin reuptake inhibitors (SSRIs: fluoxetine, paroxetine, and sertraline). Kaplan-Meier plots and Cox regression were used to analyze treatment discontinuation and treatment event. The natural log transformation was applied to the cost variables. Stratification and multivariable regression methods were used to adjust for selection bias.

4. Results: Modeling algorithm and PVS comparison

Analyses of 11,654 expert classified claim pairs were split into two groups: claim pairs involving two different drugs (multi-drug claim pairs, \( n = 6,523 \)) and those involving the same drug twice (mono-drug claim pairs, \( n = 5,131 \)). Descriptive results are presented in Figure 2.
Knowledge Discovery in Pharmacy Claims Data

Overall, 64.14% of multi-drug claim pairs and 8.77% of mono-drug claim pairs were combined by the expert. Formal hypothesis tests were carried out between select algorithm-PVS methods using the Wilcoxon Signed Rank test. For the multi-drug claim pairs, CARTf performed significantly better than QUESTf, Fixed7, and CARTi (p = 0.0020). QUESTi outperformed QUESTi (p = 0.0098). The differences between CARTf/QUESTf and CARTp/QUESTp were marginal (p = 0.0371/0.0488) especially given the multiple comparisons. For the mono-drug claim pairs, the single strong result was that CARTf/QUESTf outperformed CARTf/QUESTf (p = 0.0039/0.0098). Unexpectedly, QUESTf outperformed QUESTf and CARTf, though the difference between CARTf and QUESTf was not statistically significant. All other differences were not statistically significant.

5. An application: Sample outcomes analysis

Strong evidence was found for selection bias based on market (state of patient’s residence), product (employer versus senior plan), age, evidence of prior antidepressant treatment, and prior year overall prescription cost. Analysis was stratified by treatment start year to accommodate a drug formulary change that disadvantaged fluoxetine and occurred during the study period. All results presented below are stratified by treatment start year and adjusted for market, product, age, evidence of prior therapy, and prior year overall prescription cost.

Before the formulary change, there were no statistically significant differences in treatment discontinuation fluoxetine patients and either paroxetine (p = 0.34) or sertraline patients (p = 0.15). After the formulary change, patients starting on paroxetine were 1.46 (95% CI: 1.28, 1.66) times more likely to discontinue therapy than fluoxetine patients. Sertraline patients were 1.22 (95% CI: 1.03, 1.45) times more likely than fluoxetine patients to discontinue therapy within one year.

Prior to the formulary change, paroxetine patients were 0.74 (95% CI: 0.62, 0.89) times as likely to have a significant treatment event as fluoxetine patients. There were no statistically significant differences between fluoxetine and sertraline patients (p = 0.63). After the formulary change, there were no statistically significant differences between fluoxetine
patients and either paroxetine ($p = 0.57$) or sertraline patients ($p = 0.01$).

Paroxetine and sertraline patients had statistically significant lower overall and antidepressant utilization compared to fluoxetine patients both prior to and after the formulary change (all $p < 0.01$). There were no statistically significant differences in non-antidepressant costs across groups at the $p = 0.05$ level.

6. Discussion and conclusions

The primary obstacle to mining pharmacy claims databases to conducting clinically oriented research is the significant transformation required to convert the financially oriented databases into clinically relevant constructs. Clinical expert review of all claims would be accurate but impractical due to data volume. Computer based classification is the only practical way to deal with the voluminous data.

Building PVSs and using modeling techniques is more labor intensive in the short-term than using simple rules to determine claim pair combination. This effort is justified only when the model-based method outperforms existing methods in agreeing with expert opinion. The results presented here indicate that the appropriate modeling technique and PVS could substantially outperform existing rule sets.

The sample antidepressant analysis demonstrated how pharmacy claims could be utilized to begin to understand the implications of treatment decisions in actual clinical practice. The results can be used as a starting point for requesting further data and directing future research. Ultimately, the results could be used to justify the expense of conducting chart reviews, surveys, and/or prospective studies into particular classes of patients or treatments that may have been overlooked or excluded in the original clinical trials required for drug approval.

As an example, differences in observed discontinuation or switching patterns could be coupled with clinical understanding to justify the effort of conducting chart reviews or prospective trials in search of side effects or other adverse reactions that cannot be directly captured in administrative claims databases. (The detection of side effects depends on the nature of the side effects and the data that are available). Conversely, if no differences in treatment patterns could be found in the data, one might question the advisability of expending substantial resources on detailed research with the smaller sample sizes typical of more controlled trials.

Finally, it is important to note that Managed Care Organizations are now expanding their collection of laboratory data. As these laboratory databases grow and becomes more representative, repository based research will also benefit.

In summary, all relevant clinical questions cannot be answered with prescription claims databases. Controlled, prospective trials, and more in-depth research, would be required to sufficiently establish clinically relevant relationships. In this paper, we have presented an alternative, cost-effective approach to understanding how treatments may work outside controlled settings, thereby permitting the opportunity to generate focused and productive clinical questions. It is hoped that the methods outlined in this paper will improve the extent to which researchers exploit an underutilized resource in asking more focused and productive clinically oriented questions.

REFERENCES

Knowledge Discovery in Pharmacy Claims Data


