RC25391 (WAT1303-073) March 18, 2013
Computer Science

IBM Research Report

Computer-Aided Auditing of Prescription Drug Claims

Vijay Iyengar, Keith Hermiz, Ramesh Natarajan
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 208
Yorktown Heights, NY 10598
USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

LIMITED DISTRIBUTION NOTICE: This report has been submitted for publication outside of IBM and will probably be copyrighted if accepted for publication. It has been issued as a Research Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distribution outside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or legally obtained copies of the article (e.g., payment of royalties). Copies may be requested from IBM T. J. Watson Research Center , P. O. Box 218, Yorktown Heights, NY 10598 USA (email: reports@us.ibm.com). Some reports are available on the internet at http://domino.watson.ibm.com/library/CyberDig.nsf/home .

Computer-Aided Auditing of Prescription Drug Claims

Vijay Iyengar, Keith Hermiz and Ramesh Natarajan
IBM Thomas J. Watson Research Center
P. O. Box 218, Yorktown Heights, NY, 10598

Abstract

We describe a methodology for identifying and ranking candidate audit targets from a database of prescription drug claims. The relevant audit targets may include various entities such as prescribers, patients and pharmacies, who exhibit certain statistical behavior indicative of potential fraud and abuse over the prescription claims during a specified period of interest. Our overall approach is consistent with related work in statistical methods for detection of fraud and abuse, but has a relative emphasis on three specific aspects: first, based on the assessment of domain experts, certain focus areas are selected and data elements pertinent to the audit analysis in each focus area are identified; second, specialized statistical models are developed to characterize the normalized baseline behavior in each focus area; and third, statistical hypothesis testing is used to identify entities that diverge significantly from their expected behavior according to the relevant baseline model. The application of this overall methodology to a prescription claims database from a large health plan is considered in detail. Audit,Fraud and Abuse

1 Introduction

The audit process for health care claims must take into account two somewhat conflicting concerns. On the one hand, health care costs must be controlled by identifying and eliminating error, fraud and waste in the claims settlement process. On the other hand, within reason, the claims review process should not inhibit or constrain legitimate medical professionals and patients from achieving the best possible health outcomes based on the most effective treatments. This intrinsic dilemma is an understated yet overriding concern for the design and implementation of a computer-aided audit methodology for health care claims.

Most computer-aided audit systems invariably rely on business rules of thumb or heuristics to discover instances of fraud and abuse, although this approach may have many limitations in the health care claims context. For instance, these heuristics are often formulated in an *ad hoc* fashion, and may not adequately incorporate relevant domain knowledge and data modeling expertise. Furthermore, a rigid application of these heuristics may be inappropriate in certain situations, and may lead to a large number of claims reviews that will undermine the utility of the computer-aided audit process. Lastly, while this approach may be quite adequate for subverting the known or obvious patterns of fraud and abuse, it may be less than adequate for unanticipated and emerging patterns, or for sophisticated "under the radar" schemes, since respectively, these either completely bypass or completely conform to the prevailing heuristics. In the light of these limitations, this class of computer-aided audit approaches may not have the required flexibility and effectiveness for the health care claims context.

Many aspects of the investigative process for detecting fraud and abuse in health care claims are human intensive, and rely on the expertise of a small number of professionals with specialized knowledge and forensic skills. However, computer-aided audit techniques are increasingly being used to supplement the human-intensive effort, and in particular, may be part of a preliminary screening process to identify a

smaller set of targets for detailed investigation and prosecution. The use of computer-aided audit techniques as an adjunct and precursor to the human-intensive process will be credible and effective, only if the audit targets are provided with high selectivity and, preferably, ranked in some order that emphasizes the severity of departure from expected audit norms. In addition, the results should be supported by a deep-dive analysis that provides background evidence for investigating the top-ranked audit targets. The need for high selectivity in identifying potential audit targets is equivalent to ensuring that the number of false positives (among the top-ranked targets) and false negatives (among the bottom-ranked targets) is small, since each false positive represents wasted time and effort in an essentially futile audit investigation, while each false negative represents a monetary loss due to an undetected instance of fraud and abuse in the audit process.

The development of a suitable computer-aided audit methodology can be an iterative process where, for instance, the results from the initial implementation may reveal certain deficiencies that can be overcome by incorporating additional data elements and algorithmic refinements. However, the difficulty with carrying through this iterative process in the health care claims context, is that the confirmation of the false positives and false negatives is expensive and time-consuming, so that the required feedback is unlikely to be provided in a timely fashion. Therefore, it is highly desirable to be able to identify potential audit targets with high selectivity in the first effort itself, without any expectation of confirmatory feedback on the results; this, in turn, is only possible by incorporating a high level of domain expertise supported by all relevant data elements in the computer-aided audit analysis, rather than deferring these considerations to the human-intensive steps in the deep-dive analysis. This aspect is quite crucial for the successful implementation of a computer-aided audit process, but is particularly challenging to achieve in the health care domain, where the claims circumstances are often obscured by complex medical diagnoses, the immense variety of procedures and treatment protocols, and by the pharmacological subtleties of the prescribed medications. The inclusion of all these relevant factors, consequently, leads to a very high-dimensional, albeit sparse, set of predictors, and a particular novelty of our approach, as described further below, is the use of specialized techniques for storing and processing the claims data in order to perform this modeling and analysis in an efficient, even tractable, manner.

The scope and extent of health care fraud is described in [22], who also advocate the use of statistical methods for detecting fraud and abuse in various scenarios such as identity theft, fictitious or deceased beneficiaries, prescription forgery, phantom or duplicate billing, bill padding, upcoding of services and equipment, and service unbundling [23].

In contrast, our approach in this paper, as applied to fraud and abuse in prescription claims data, is to specifically identify those entities who are associated with abnormal and excessive prescriptions for certain classes of medications that are of great concern and interest from an audit perspective. In particular, we do not attempt to develop statistical models for each individual scenario above [23] that might be the underlying mechanism for any of these abnormal and excessive prescriptions (e.g., forged prescriptions, doctor shopping, or prescription selling); instead, deviations from normative or baseline behavior are identified, taking into account the particular context of the interaction between the patient, prescriber and pharmacy for each prescription claim, and aggregated over the set of claims for each entity.

Although this approach and methodology has been described here for prescription claims data, it can be seen to be quite broadly applicable to many other focus areas in health care claims. However, there are some compelling reasons for considering prescription claims audit as an initial priority, even if the financial impact may be greater in some of the other focus areas in health care claims. First, as pointed out to us by the audit group at the large health care plan that advised this project, any fraudulent activity in the prescription claims data is often the proverbial tip of the iceberg, and can lead to the unravelling of a chain of supporting fraudulent billing claims for medical treatments and office visits involving the same entities. Second, the societal impact of prescription fraud is staggering and disproportionate to its direct

financial impact, since it is one of the main conduits for the illegal diversion of drugs, which is responsible for the huge challenges of substance abuse and addiction, drug shortages, and an active black market that involves and compromises the health of legitimate patients [11, 17].

In the health care claims domain, fraud detection can be carried out in both off-line or on-line modes of analysis (e.g., see [1], who consider both these modes for prescription fraud detection). In the off-line mode, the analysis can be used by audit investigators to augment the process of retrospectively reviewing claims data in order to identify target entities for further investigation. In the on-line mode, the emphasis is on early and reliable identification of a potential fraud outcome, either from an isolated claim, or from an aggregate set of claims, so that an appropriate early intervention can be initiated to restrict further losses associated with the suspected outcome.

The impact of computer-aided fraud detection methodologies for claims auditing must be evaluated carefully in terms of its repercussions in the health care domain, similar to the issues discussed in [8, 10]. For instance, the anticipated changes in the prescribing patterns of opiods induced by any enhancements in the prescription claims audit process, may lead to pro-active concerns from patients that their treatment options are compromised (e.g., patients who suffer from chronic pain symptoms that are effectively treated by this class of drugs). This issue is related to, and subsumed by, a larger set of concerns, viz., the possible impact of aggressive fraud detection systems on legitimate medical services and practices (e.g., leading to excessive delays in claims processing and reimbursement, or even to incorrect claims denials).

Our overall approach described here is consistent with previous work on statistical methods for detection of fraud and abuse in various domains such as financial trading, credit card transactions, telecommunications, network intrusion and health care [3, 4, 21, 2, 7, 1]. We direct the reader to survey papers on fraud detection, from a general perspective in [18], and from a specific health care perspective in [16]. In contrast with much of this work, however, our approach does not require any explicitly labeled instances of fraudulent claims or entities. This aspect may rule out the use of techniques for directly modeling the fraud outcomes based on supervised or semi-supervised statistical methods. However, in the health care domain, as observed earlier, the rapidly-changing nature of the fraud and abuse concerns, as well as the difficulty in obtaining timely feedback and validation on potential fraud cases, provides a good rationale for adopting the approach that is proposed here.

We now compare and contrast our approach with the closest previous work that we are aware of [1, 12], which in particular, also share the same characteristic of not requiring any explicitly labeled instances of fraudulent claims or entities in the analysis.

For instance, [1] consider fraud detection in prescription claims data using both the off-line and on-line application modes. The historical claims data was used to obtain pairwise co-occurrence frequencies in the individual prescriptions for each medication (drug) in combination with other factors such as age, gender, medical diagnosis, or other co-prescribed medications. A notion of risk (i.e., likelihood of fraud) was associated with the particular medication in a given prescription claim, for each of these co-occurrence dimensions. This risk value was specified using an exponential scale, which intuitively assigns a very high value when remaining factors in the given prescription claim have small values for the relevant co-occurrence frequencies, when compared to appropriate reference frequencies. Based on domain expertise, appropriate thresholds were defined for each risk score, and claims with risk scores above these thresholds were flagged, and the associated trigger conditions were also provided. In contrast to this approach [1], our approach is, first, to identify entities (prescribers, pharmacies, patients) who exhibit abnormal prescription behavior in the off-line audit setting, with sufficient statistical evidence over multiple claims instances to justify further audit investigation. Second, the baseline model for normal behavior (note that in this case, the behavior that is being modelled is the prescription rate for a specific medication, or specific class of medications), is capable of representing the more complex and high-dimensional

interactions that are inherent in claims data. For instance, any patient medical history, such as patient medication profile, that can be gleaned from the anonymized claims data, can also be used to refine the normalizations in the expected prescription behavior. Similarly, prescriber attributes (e.g., their practice and specialties, and their profiles and pattern of diagnoses and procedures) can also be used in a similar way. The appropriate set of conditioning attributes and their interactions, is not predefined but rather determined from data, and this attribute selection is based on rigorous statistical criteria customized to each individual segment in the attribute space that is homogeneous in the model response. The need for such complex behavioral models based on the analysis of sparse, high-dimensional data is illustrated below with several examples. Finally, one essential difference between [1] and our work, is that for the claims data set that was used here, it was not possible to link individual prescriptions to a corresponding single-valued patient diagnosis. We note that in general, prescription claims are filed by the pharmacy and contain the medication information, while medical claims are filed by the prescriber and contain the diagnosis information. These two sets of claims may even be handled by different insurance programs. The correspondence between the prescription and medical claims data requires having linked patient information between the two sets of claims; and even when this linked patient information is provided accurately, elicitation of the direct correspondence between an individual prescription claim and an individual medical claim in these two disparate claims data sets will be an imperfect exercise at best.

We have adopted much of our methodology from [12], although that work is concerned with expense auditing for corporate travel and entertainment. For instance, [12] consider expense claims submitted by employees in various focus areas (e.g., different categories such as ground transportation, restaurant tips, etc.), and the expense claims of distinct auditable entities, such as individuals or even entire departments, was evaluated in these focus areas. For each entity, any abnormal behavior was highlighted if there were significant departures from the expected behavior posited by a normalized baseline model for that focus area. This approach assumes that any abnormal behavior is only exhibited in a small fraction of the overall set of expense claims so that normalized baseline models can be reliably estimated. The identification of the entities with abnormal behavior is then based on a Likelihood Ratio (LR) score, which is computed from the actual behavior over the set of encounters for each entity, relative to the predictions of the normalized baseline model over this same set of encounters. The statistical significance of this LR score is based on evaluation of the relevant p-values using Monte Carlo methods similar to those used in scan statistics [13]. In spite of the similarities in the overall approach, the health care domain is significantly more challenging in the terms of the data and modeling complexities. For this reason, we have developed new algorithms, as described further below, which are based on learning homogeneous segments for the model response in a given focus area from sparse, high-dimensional data. For any focus drug or drug class that is of audit interest, the model response used in the baseline model is the corresponding prescription rate, and from this baseline model, entities with significant abnormal behavior are then ascertained using methods similar to [12, 13].

To summarize, we describe a methodology for the off-line application of fraud detection to augment the audit and investigative process for prescription claims data. The proposed approach relies on incorporating all available data (e.g., diagnosis codes, procedure codes, medication history, and other prescriber and patient attributes). The baseline models that we obtain are often relatively easy to interpret and, when this is the case, they can be described to audit investigators and domain experts in an intuitive and transparent format that encourages discussion and constructive feedback. From the technical perspective, the main goal in developing this methodology has been to reduce the false alarm rate for selecting candidates for the audit investigation process. Consequently, the goal is to provide a good balance between maintaining program integrity and cost containment on the one hand, and providing good health outcomes and ensuring appropriate patient care on the other.

An overview of this report is as follows. Section 2

provides the project background, and describes the prescription claims data set that was used in our analyses. Section 3 describes some of the scenarios and focus areas that we have identified as important for prescription claims fraud and abuse. Section 4 describes our methodology and algorithms, including the rule-generation algorithm that is used to obtain normalized baseline models in a scalable and efficient way, the scoring of entities to identify abnormal behavior, and the selection and ranking of these entities for further investigation. Section 5 discusses the empirical evaluation and results, and provides analyses for audit investigations in four different drug therapeutic classes. Section 6 describes the significance and impact of the work, and a discussion of some extensions that are of ongoing interest. Section 7 provides a concluding summary.

2 Background and Motivation

Our project to develop methods for computer-aided auditing of prescription claims data was done in partnership with the audit group responsible for detecting and preventing fraud and abuse within a large health care plan. The prescription claims data set used in the analysis was obtained from the fee-for-service part of the health plan, and included prescriptions that were dispensed in both in-patient and out-patient settings.

The individual records in this prescription claims data set contained the relevant recorded information, including the participating pharmacist, patient, prescriber, formulary, prescription frequency, length and dosage, and the claims and co-payment amounts. (All patient information was encrypted and anonymized in compliance with the Health Insurance Portability and Accountability Act privacy requirements).

Other supporting data tables included a list of certified prescriber profession codes, prescriber specialty codes (which were often self reported and possibly unverified), and a drug classification table (which contained the packaging, dosage, formulation and drug therapeutic class for each individual formulary). Other relevant information such as the descriptive details for the International Classification of Diseases, 9th Revision (ICD-9) codes, the Current Procedural Terminology (CPT) codes, and the Clinical Classifications Software (CCS) codes, were all obtained from reliable public sources.

In addition to the prescription claims, a set of supporting medical claims data was also acquired for all the patients in the prescription claims database. This additional data was obtained after the initial data analysis indicated that the medical claims would be useful for constructing an objective profile for the patients and prescribers, and for establishing the medical context for individual prescription claims.

In summary, for the audit analyses corresponding to a certain *analysis time window* of interest, the prescribers were profiled by their top diagnoses codes and top procedure codes from the medical claims data in a certain *history time window* (this history time window typically consists of the period up to and including the analysis time window). The patients, on the other hand, were profiled by gender, age interval, and by their medications taken in the history time window. In this profile, the medications are abstracted to the drug therapeutic class level (comprising of around 90 distinct classes) to avoid a proliferation of profile elements corresponding to equivalent or very similar medications.

For the experiments reported in this paper, a three month period in 2011 was chosen as the history window to extract the prescriber and patient profiles. The analysis window was the last month of the history window. In order to report on the generalization performance of the model, we split the data in the analysis window into equal sized training and test sets. Since we will illustrate our methods using examples of abnormal prescribers, the training/test split was done based on prescribers (i.e., all the transactions in the analysis window for any given prescriber were either entirely in the training data set or entirely in the test data set). In a production setting, the entire data will typically be used for training. The rule list generation system was given the entire data in the training set, and the relevant data summaries were then listed to provide the scale at which the analyses were done. To give an idea of the scale of the data used in the experiments described in this paper, the

training set consisted of 8.1 million individual drug prescriptions involving some 2.3 million patients, 4.5 thousand pharmacists, 99 thousand prescribers, and 19 thousand distinct formulary codes. The patient drug profile based on the three-month history window had around 11.8 million elements.

3 Prescription Drug Fraud and Abuse Scenarios

Most cases of prescription drug fraud and abuse are associated with specific drugs which invariably belong to two categories: the first consists of high-volume drugs that can be resold to pharmacies and double-billed to the health plan, while the second consists of drugs that have high street value due to their association with non-medical and recreational abuse.

Our analyses have been primarily directed towards the second category of drugs mentioned above and, in particular, the focus areas for modeling prescription rate behavior are defined at the drug therapeutic class level, specifically on four classes listed below:

- Narcotic Analgesics: This class contains two of the most widely abused prescription medications, oxycodone and hydrocodone, and also contains a variety of combination drugs which are often abused because they may have less stringent controls on dispensing and distribution.
- Ataractics-Tranquilizers: This class includes medications with benzodiazepines that are prescribed as anti-anxiety drugs but are also susceptible to addiction and abuse.
- CNS Stimulants: This class includes medications like the generic methylphenidate that are prescribed for attention deficit hyperactivity disorder (ADHD), but are also abused due to their euphoria-inducing effects.
- Amphetamine Preparations: These drugs are often abused for their performance-enhancing benefits and euphoria-inducing effects.

Our approach to defining the focus areas at this level of abstraction is a simplification. For example, from an addiction and abuse perspective, the specific active ingredients in various drugs, their potencies and equivalences, and their normal dosage and prescription frequencies are also considered to be important by domain experts. However, such a detailed analysis would require deeper pharmacological domain expertise, as well the development of quantitative models that can capture this expertise in the analyses. Nevertheless, despite simplifications, the choice of focus areas at the abstraction level of the drug therapeutic class provides a good starting point for obtaining useful results.

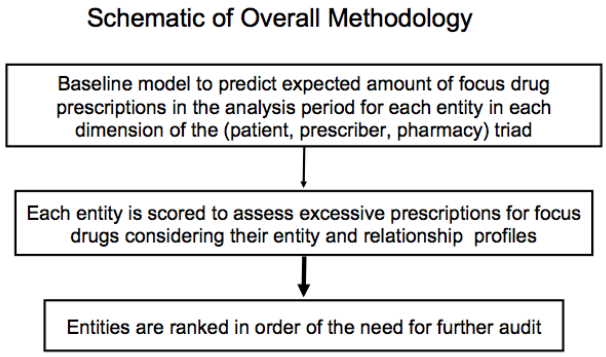

Figure 1: Schematic of methodology for identifying entities with potential abnormal claims behavior

4 Methods and Technical Solutions

A schematic overview of the analytical methodology is described in Figure 1, which consists of three steps, viz., first, the selection of a focus area and construction of a baseline model to predict the expected behavior of all entities in this focus area; second, the scoring of each entity based on its encounters in the analysis time window with respect to the baseline model; and third, the ranking and selection of scored entities as potential audit targets for fraud and abuse.

A basic assumption in this methodology is that the majority of data to be audited consists of normal patterns of behavior, so that robust estimates are ob-

tained for the baseline models without explicit labels for abnormal transactions. In addition, we note that any abnormal behavior may not always be a consequence of fraud or abuse, since incomplete data, incorrect data and lack of context may also contribute to the observed abnormal behavior.

4.1 Baseline Model Structure

The baseline model is developed separately for each focus drug class. Each distinct combination of a prescriber (e.g., physician, nurse practitioner), a patient and a pharmacy that is encountered in the analysis time window period in the prescription claims data represents an instance for learning the baseline model. For each such instance, the counts of the total number of prescriptions and the counts for prescriptions of the focus drug therapeutic class are obtained. The proportion of these two quantities is the focus drug "prescription rate" which is modeled.

The baseline model is generated by learning the relationship between patient and prescriber profiles, and the rate of focus drug prescriptions. While our methodology can also incorporate pharmacy characteristics in a straightforward way, these are typically not of primary relevance for determining the prescription rate. Therefore for brevity, the discussion of results with the pharmacy profiles is not included in this paper.

While there are many possible model structures that can be used for obtaining the baseline models, we chose a rule list model structure to segment the sparse high dimensional input space into relatively homogeneous segments with respect to the prescription rates. These models have a transparent structure, which allows for an easy inspection and validation of the model details by expert audit investigators. Another reason for the choice of rule list model structure is its ability to capture the broad segments of prescribing behavior for any focus drug class that can be determined using only claims data. The algorithm to generate the rule list model was tailored to this application and is described in the next subsection.

Clearly, predicting whether a prescription for a certain focus drug class will be given in any specific encounter between a prescriber and a patient requires detailed information about the patient profile (e.g., health status, diagnostic history and test results) and the prescriber profile (e.g., specialization and clinical expertise). However, due to various technical reasons, at the present point the prescription claims and medical claims data could only be linked for the prescribers, and hence, the analyses and results reported in this paper were only obtained with prescriber profiles. In future work, we expect to resolve the technical reasons, and be able to incorporate the relevant patient profiles and medical history in the analyses, since this may lead to further improvements in the quality of the baseline model predictions.

As mentioned earlier, the prescriber and patient profiles are represented in a sparse binary form and are based on the claims data in the history window. Our initial approach for generating prescriber profiles was to use the information on the profession codes and specialty codes elements that was provided with the claims data. However, the profession codes, which cover broad licensing categories, are too coarse to be useful. The specialty codes, which are self-reported and unverified, are often inconsistent and missing. Therefore, after this initial experimentation, rather than using either of these data elements, we instead characterized prescribers by their clinical behavior as gleaned from the claims data. For each prescriber, these profile elements included their top five diagnoses (abstracted to the first three digits/characters in the ICD-9 taxonomy), and their top five procedures (abstracted to the corresponding CCS classifications for single level procedures developed by Agency for Healthcare Research and Quality [5]). Based on our experiments, this new approach for obtaining the prescriber profiles is a much more objective reflection of the patient population that they serve and the medical conditions that they treat.

For the patient, the profile elements included gender and age intervals which were dummy-encoded to separate out children under 11 years, with the remaining population in 20-year interval bins (e.g., 11–30 years, 31–50 years, etc.). In addition, the patient profile elements also included their drug usage profiles in the history time window (abstracted at the drug therapeutic class level). Note that the focus

drug class being modeled is not included as a predictor in the patient drug profile, in order to avoid circularity, and to ensure that the resulting model can use the patient conditions in terms of usage in the other drug classes that influence the prescription rate in the focus drug class. Finally, for clarity, we note here and further emphasize in Section 6, that our models use aggregate data over identical encounters during the history time window, and therefore do not attempt to model the temporal aspects of prescription behavior in this short time period, either for patients or for prescribers,

4.2 Rule List Model Generation

The algorithm for rule list model generation is tailored to the characteristics of the sparse data that arise in this domain. All the inputs are either naturally in binary form (e.g., presence or absence of diagnoses or procedures) or have been transformed into binary form by binning (e.g., age). The structure of the rule list model is an ordered list of rules where each rule is a conjunction of terms and each term specifies either the presence or the absence of some input binary variable. As in any ordered rule list model, an instance is said to be covered by a particular rule R if it satisfies the conditions of rule R but not those of any rule preceding R in the rule list. Hence, the rule list partitions all instances into disjoint segments corresponding to each of the rules and a default segment covering instances not covered by any rule in the list. There is a predicted rate of focus drug prescriptions associated with each segment (including the default segment).

The rule list generation algorithm is sketched out in Figure 2. The algorithm starts out with an empty rule list and all the training instances to be covered. Each iteration of the outer loop potentially adds a rule R to the rule list RL. Each iteration of the inner loop potentially adds a term T to the current rule R being generated. The criterion used to select the term for possible addition to the rule and the stopping criteria for rule refinement and rule list expansion are tailored for this application. The candidate term for addition to the rule is selected in a greedy fashion using the metric from a Likelihood Ratio Test (LRT) as

```
xAlgorithm: Generate Rule List
Initialize RL to empty.
Initialize set S of instances = training set
L1: Loop Forever (add rule to rule list)
    Initialize current rule R to null
    L2: Loop Forever (add term to rule)
        For each possible additional term
            Evaluate each term
            Greedily select best term T
            Check significance of chosen term T
            If significant
              Add best term T to current rule R
            If not significant and rule R is not null
              Add rule R to rule list RL in order
              Remove instances covered by R from S
              Exit inner loop L2
            If not significant and rule R is null
              Exit outer loop L1
        end loop L2
    end loop L1
end algorithm
```

Figure 2: Schematic of rule list generation algorithm.

described next using the illustration in Figure 3. We consider all terms that have not already been used in the current rule being generated (in any order). For each candidate term T, the LRT compares two hypotheses for modeling the instance space S at that point. The null hypothesis models the entire set of instances S with a single Bernoulli model using the mean rate over S. The alternate hypothesis models the instances $R \cap T$ covered by the candidate updated rule (R with term T added) and those remaining in S using separate Bernoulli distributions with their respective mean rates. The candidate term T with the highest LRT scores is chosen greedily for consideration. However, it is added to the rule R only if a second LRT test passes a significance threshold. The second LRT test considers only the instances covered by the rule R. The null hypothesis models these instances with a single Bernoulli distribution using the mean rate over R. The alternate hypothesis models

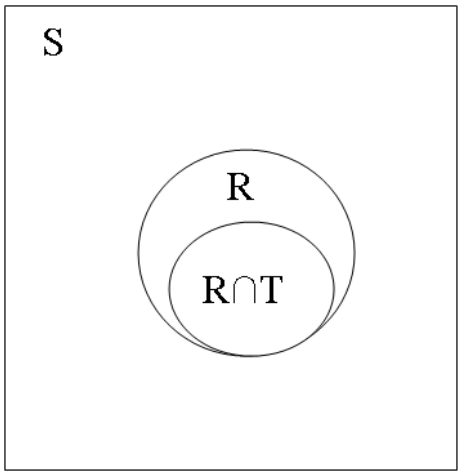

Figure 3: Greedy term selection illustration

the instances in $R \cap T$ and the remaining instances in R using separate Bernoulli distributions with their respective mean rates. The term T is added to the rule R only if the second LRT test is significant at a user specified threshold. This significance test based heuristic for adding terms (and rules) is the reason the rule list model tends to not overfit the training data as experimental results indicate.

We will use examples of rules from one of our applications to further clarify the intuition behind our heuristic. Consider an intermediate stage in the rule list generation process with the focus on tranquilizers. After three rules had been generated in sequence, there were 1,669,832 instances of the triplet (prescriber, patient, pharmacy) to be covered by the remaining rules and the corresponding rate for tranquilizer prescriptions for these instances was 2.2%. The first term chosen for consideration to build the fourth rule restricted the prescribers with "Disturbance of conduct" as one of their top five diagnoses. The rule with just this one term resulted in a segment (refer to region R in Figure 3) with 6,415 instances and a significantly higher rate of tranquilizer prescription at 20.5%. The next term to be considered by the greedy heuristic was to further refine the allowed set of prescribers by adding the constraint that the procedure "Psychological and psychiatric evaluation and therapy" was one of their top five procedures. This additional term satisfied the significance threshold for the second LRT and further refined the space (refer to region $R \cap T$ in Figure 3) to have only 2,252 instances but with an even higher tranquilizer prescription rate of 38.7%. This example illustrates how using prescribers top diagnoses and procedures allows refinement in the model beyond the broad specialty categories we also had available. Two more terms were added to further refine this segment to finally cover 2,106 instances with a tranquilizer prescription rate of 40.2%. The corresponding prescription rate for this segment in the test set was 36.2%.

Continuing with the same rule list, we will also consider the next rule generated to illustrate the role of patient attributes in the model. The tranquilizer rate in the 1,667,726 remaining instances was 2.15%. The first term chosen for the next rule was to consider patients who had filled at least one prescription for anti-arthritics (in the history time window) to start the definition of a segment with lower (than 2.15%) rate of tranquilizer prescriptions. The rule generator added 18 additional terms related to other medications (e.g., anti-depressants, anti-convulsants etc.) and related to prescribers based on their top diagnoses and procedures. The final segment defined by this rule had 321,580 instances with a significantly lower rate of 0.47% for tranquilizer prescriptions. The corresponding prescription rate for this segment in the test set was 0.5%. This also illustrates how the model allows for complex interactions to be modeled from the data; the segment in this example cannot be modeled by only considering pairwise interactions [1].

Next, we will contrast our rule generation heuristic with that used in classical algorithms like FOIL and RIPPER [19, 6]. First, our rule generator typically mixes in rules with either low or high rates in the ordered rule list being generated based on the LRT metric. Secondly, consider a hypothetical stage in the rule generation where the instance space to be covered has a total prescription count of 1000 and a focus drug count of 20, corresponding to a rate of 2%. Suppose there were two interesting choices of binary variables to build the next rule. Choice A partitions the space into two sub-spaces with (focus drug count,

total count, focus drug rate) values of $(19, 400, 4.75\%)$ and $(1, 600, 0.17\%)$. Choice B, on the other hand, partitions the space into two sub-spaces with (focus drug count, total count, focus drug rate) values of $(5, 9, 55.6\%)$ and $(15, 991, 1.51\%)$. The FOIL information gain metric would prefer choice B over choice A, whereas, our LRT based heuristic would pick choice A. This is consistent with our desire to build rules with significant evidence in the data and is consisten with the approach used for entity scoring that is described in the next subsection. In consequence, we have found empirically that subsequent phases to perturb and refine the rule list provide limited improvement to the model quality in contrast with many other rule generation algorithms.

Our heuristic leads to a rule refinement and rule list generation process that is self limiting, so that the generated rule lists do not overfit the training data when the user-defined threshold for the p-value is set quite low (e.g., 0.0001) as shown in the experimental results section. The number of segments and their sizes are not explicitly controlled with user specified parameters, but fall out as a consequence of the recursive partitioning process as the sequential list of rules is generated using the heuristic based on the significance tests described above. This is also illustrated in the results section.

The final step in the generation of the rule-list based baseline model is to determine the predicted rates of the focus drug class. For each segment induced by the rule list model the predicted focus drug class rate is simply the mean rate observed in the training set instances covered by the segment. We would anticipate some segments to cover situations where high rates of focus drug prescriptions are expected and others to cover circumstances that typically have very low rates.

4.3 Entity Scoring for Abnormalities

The rule list baseline model represents the expected behavior for focus drug prescriptions under various circumstances as represented in the rules involving patient and prescriber characteristics. The next step in the methodology is to score the target entities (prescriber, patient or pharmacy) quantifying their excessive prescriptions for the focus drug as measured by the deviation from the baseline model. It is important to note that a target entity can have prescription activity that falls into more than one segment. A simple example of this could be a physician when prescribing for a child is covered by a different segment (rule) when compared to the same physician prescribing for an adult. The scoring for an entity should aggregate the deviation from the baseline model over all the segments that the prescription activity falls into. The scoring for an entity should reflect the magnitude of the deviation and the volume of transactions with excessive prescription rates. Scoring is based on Likelihood Ratio Tests as in previous work in spatial scan statistics [13, 14].

The score for an entity E (e.g., a prescriber) is computed as follows. Consider each segment S defined by the baseline model. In this segment S, let A be the total count of prescriptions in S and F be the count for the subset corresponding to focus drug prescriptions. The expected rate of focus drug prescriptions for this segment is F/A. Consider all the data instances d for E that belong to the segment S. Let a and f be the counts for all prescriptions and focus drug prescriptions in d, respectively. Then the contribution to the score for entity E from this segment S is given by computing the log likelihood ratio based on the Bernoulli distribution as shown below. The score contributions for entity E from each segment S are aggregated by summing up after assigning a sign to each contribution based on whether the focus drug rate for the entity in that segment was higher $(+)$ or lower $(-)$ than the expected rate in the segment [14].

The entity scores can be transformed to more meaningful values by estimating the corresponding p-values. Monte Carlo methods provide a direct way for estimation [13]. The distribution of these scores under the null hypothesis as represented by the baseline model can be determined empirically by performing N randomized experiments as follows. In each experiment, a synthesized data set is created where the number of focus drug prescriptions for each instance I is determined using pseudo-random generators modeling the Bernoulli distribution with the focus drug rate expected for the segment that instance I belongs to. The maximum score achieved by any entity using

this synthesized data set is recorded. The set of these maximum scores achieved in the N Monte Carlo experiments is used to transform the entity score to the estimated p-value [13].

(1)
$$\begin{aligned} Score(E,S) = & f \log f/a + (a-f)\log(a-f)/a \\ & + (F-f)\log(F-f)/(A-a) \\ & - F\log F/A - (A-F)\log(A-F)/A \\ & + [(A-a)-(F-f)] \\ & \times \log\left[(A-a-F+f)/(A-a)\right]. \end{aligned}$$

5 Empirical Evaluation and Results

This section has the experimental results from the analyses of prescription claims over a three month time window in 2011 for all the focus drug classes discussed in Section 3. First, we assess the ability of the baseline models to explain the need for focus drug prescriptions. Then, we apply these baseline models to score and rank entities based on their abnormal behavioral patterns of excessive prescriptions for each of these focus drug classes.

5.1 Baseline Model Evaluation

The baseline model was evaluated using a 50-50 training/test split of the data. Figure 4 shows the ROC curves for the four focus drug classes. Table 1 has the key characteristics of the baseline model including the area under the ROC curve (AUC). The AUC metrics achieved (in the range 0.8–0.9) for both training and test sets indicate an acceptable baseline model that does not overfit the training data.

The number of segments in the baseline model ranges from 29 to 127 considering the four drug classes. The number of variables used as terms in the rule list range from 123 to 506. The baseline model for the narcotic analgesic class is the most complex utilizing 506 variables in the rule terms out of the 1281 available binary variables. The next subsection explores the baseline model for the narcotic analgesic class further.

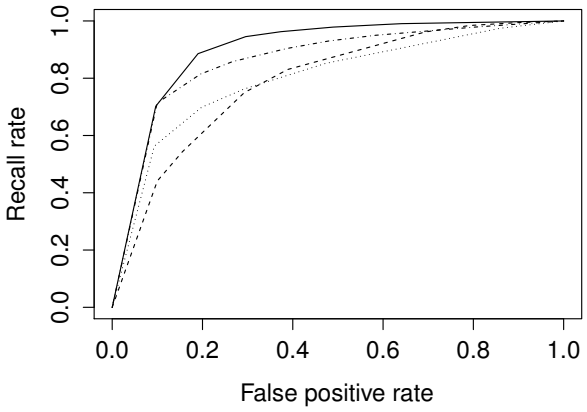

Figure 4: Baseline models:Recall versus False Positive Rate (Solid: Amphetamine Preparations, Dash-dotted: Ataractics-Tranquilizers, Dotted: CNS stimulants, and Dashed: Narcotics, Analgesics)

5.2 Baseline Model: Example Rules and Details

Some examples of segment defining rules that were generated in the baseline model for the narcotic analgesics drug class are given below.

- A rule with 29 terms covers children ages 10 and under and predicts them to have very low rates (0.16%) of prescriptions for narcotic analgesics compared to the base rate across the entire population (3.5%) when they are not seen by prescribers who perform various surgical and dental procedures. (Approximately 319,000 and 329,000 instances are covered by this rule in the training and test set, respectively.)

- A rule with 62 terms covers patients ages 11 through 70 who are taking muscle relaxants but are not on certain other medications (e.g., for diabetes) when they see certain types of prescribers (e.g., exclude gastroenterologists, exclude prescribers treating the lacrimal system) and predicts that they will have a moder-

Focus drug class	No. of segments	No. of variables	AUC	
			train	test
A	127	506	0.83	0.81
B	86	283	0.88	0.88
C	29	123	0.92	0.91
D	46	151	0.83	0.81

Table 1: Baseline models: Characteristics and Performance. The focus drug classes are: A. Narcotics, Analgesics, B. Ataractics-Tranquilizers, C. Amphetamine Preparations, and D. CNS Stimulants

ately high narcotic analgesic prescription rate (15.3%). (Approximately 88,600 and 90,900 instances are covered by this rule in the training and test set, respectively.)

- A rule with 21 terms covers older patients (age > 70) and predicts them to have low rates (0.19%) of narcotic analgesic prescriptions if they are not also taking muscle relaxants, certain antibiotics and have not been administered certain local anesthetics and when they are not seeing prescribers who typically perform various surgical procedures. (Approximately 147,000 and 140,000 instances are covered by this rule in the training and test set, respectively.)

As illustrated above, rules can have many terms to include or exclude patient conditions based on their medications and the type of prescribers they are seeing. A review of some of these rule terms with domain experts and the literature suggests that rules are extracting patterns from the data that conform to known phenomena like drug/disease or drug/drug interactions (e.g., narcotic analgesics and hypothyroidism). Such patterns are not easy to incorporate into investigations and analyses done manually.

The model induces segments whose size span several orders of magnitude, as seen in Figure 5. This figure plots a measure of segment size (total number of prescriptions covered in the training and test sets) on the x-axis (log scale) and the rate of narcotic analgesic drug prescriptions on the y-axis (log scale). The horizontal line marks the overall base rate of

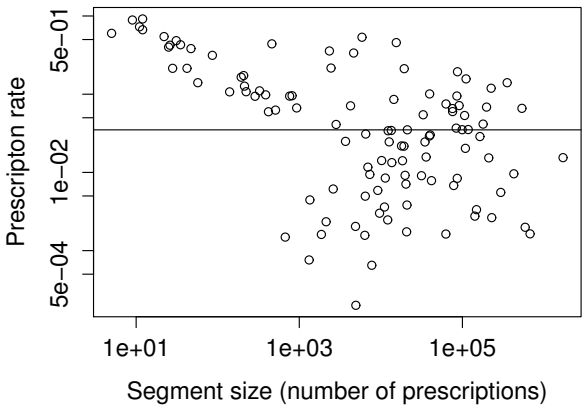

Figure 5: Narcotics prescription rates by segment size

around 3.5% for narcotic analgesics. The model has identified small and medium size segments with relatively high rates and some medium and large segments with low rates. This figure also illustrates that there is room for improvement in the baseline model by having more of the identified segments (big and small) have expected rates significantly higher or lower than the overall base rate. As mentioned earlier, without clinical data one would not expect encounter level prediction for a drug class prescription. But having patient linked diagnoses and procedure codes will help improve the baseline model significantly. For example, one of the rules indicates that high rates of narcotic analgesic prescriptions (52%) are expected when patients see prescribers performing surgical procedures on joints (with some exclusions). The model cannot refine this further without data on procedures and diagnoses linked to patients. This additional data would allow separation of encounters that involved, for example, orthopedic surgeries from those that simply were consults not leading to any surgical intervention.

Entity	Total prescrip-tions	Focus drug prescriptions		Score
		actual	expected	
1	1245	849	103	1771
2	4072	2504	1223	1273
3	746	643	85	1250
4	1257	712	143	1029
5	3010	646	78	1025
6	1730	953	253	928
7	2098	961	262	921
8	1037	676	147	885
9	2668	831	346	842
10	958	640	132	774

Table 2: Examples of entities identified by the model

5.3 Entities Identified: Examples

Our methodology, described in Section 4, utilizes the baseline model to identify entities with abnormal behavior with respect to prescriptions for any chosen focus drug class. Since the instance in our data and analyses is defined by the triplet (prescriber, pharmacy, patient), the model can be used to identify prescribers, pharmacies or patients with abnormal behaviors. We will illustrate the application of our models by focusing on prescribers. The characteristics of some of the prescribers identified by the model as being abnormally excessive in the prescribing of narcotic analgesics are given in Table 2. The table has actual counts for the focus drug prescriptions and total prescriptions in a 3 month analysis window for these prescribers. The expected number of focus drug prescriptions estimated by the model is also shown. The very high LR based scores for these entities correspond to p-values < 0.0001. It is interesting to note that the expected rate for narcotic analgesics for these entities ranges from 2.6% to 30%, considering all their encounters with patients. Their scoring and ranking for being abnormally excessive is being done taking into account these widely varying expectations on prescribing behavior for narcotic analgesics.

The validation of the entities identified by the model as being abnormal and excessive in focus drug prescriptions is ongoing. The validation process is done at various levels of rigor and human expert involvement. The first level of validation that we have completed is to determine if the model identified list includes the few known cases of fraud. For the narcotic analgesics drug class, all the known cases of fraud were correctly identified by the model as being very abnormal and excessive.

The next level of validation will be to have investigators and audit experts manually evaluate whether a sample of the specific entities identified by the model are suitable candidates for further investigation.

6 Significance, Impact and Discussion

The objective of the methodology described in this report is to consider a given focus area or scenario in the health care claims context, and obtain ranked lists that selectively identify entities with behavior that is indicative of potential fraud and abuse in this scenario. Our approach significantly extends the oft-used approach of identifying entities as audit targets based on simply ranking them according to some aggregate metrics (such as prescription counts or prescription rates). Even though these aggregate metrics may be intuitive and relevant, they ignore any normalizations that would explain and account for much of the behavior in these aggregate metrics, and which would hence modify the rankings considerably. Furthermore, our approach also goes beyond the approaches for normalizing the expected behavior of each entity based on the consideration of their peer groups at the entity level, since our baseline model is able to capture the relevant normalization from the data at finer level of granularity than the peer group, namely, at each individual and distinct encounter between the prescriber and patient. Our models and methodology, by virtue of using detailed patient and prescriber profiles based on a considerable amount of relevant context that includes medications, diagnoses and procedures, is more likely to detect "under the radar" cases where claims and supporting data have

been misreported or intentionally falsified to cover the fraudulent behavior.

Any approach that learns a baseline model of normal behavior based on training data can potentially be impacted by the abnormal instances in the data. We can get some insights into the sensitivity of our model to the abnormal instances in the training data by examining some model characteristics. Consider the rule examples and the corresponding segments illustrated in Section 5.2. The predicted mean rates for these segments are based on relatively large number of instances in the training data and unlikely to be influenced by small proportions of abnormal instances. But smaller segments can have their mean rate impacted by abnormal instances. But typically even for the smaller segments we have transactions from several entities reducing the impact of any one entity (abnormal or normal). This leaves open the question whether we could learn a rule that is based on data from a largely abnormal segment of instances, which can lead to false negatives. We hope to mitigate this possibility in the rules list validation step done by domain experts as part of our usage methodology.

In the experiments reported in the paper, we split the available data into training and test sets so that we can report on lack of overfitting by our models. In real life applications we advocate using all (or most) of the data for training. There may be reasons to keep aside a small fraction for model validation purposes. Generally, rule list models (and the closely related decision tree models) are known to be unstable, implying perturbations to the training data can impact the resulting model. However, our significance test based approach is tailored to identifying rules (and corresponding segments in instance space) that represent significant departures in behavior from the rest of the instance space. This benefits the stability characteristic of the resultant models. We ran the following experiment on a narcotics prescription data set to shed some insight on this aspect of these rule list models. We randomly split the data into two equal halves, H1 and H2. One model M1 was trained on H1 and used H2 as a test set. The other model M2 was trained on H2 and used H1 as a test set. The model performance achieved was comparable in terms of the ROC AUC (M1: training = 0.836, test = 0.822, M2: training = 0.831, test = 0.825). The rule list models, M1 and M2 were not the same, but had many similarities. For example, the first rule in both models focused on young children having very low rates of prescriptions except when they are seen by certain specialists (e.g., surgeons). Model M1 defines this rule with 9 terms and model M2 with 8 terms (dropping one specialty). The segments defined by these models differ by less than 20 instances from a total of more than 400K instances. The models differed more in the smaller segments and model M1 had a total of 60 segments compared to 66 segments in model M2. A more interesting issue is the impact of the two distinct samples of training data on the abnormal entities identified. We applied both models, M1 and M2, to the entire data and used our methodology to identify the top abnormal prescribers. The top five entities identified were exactly the same and in the same order. The next five entities were the same for both models but the order differed due to small differences in the entity abnormality scores.

The application of our models was illustrated by focusing on prescribers in Section 5.3. We have also applied our models to identifying pharmacies and patients with abnormal behavior. The larger volumes associated with prescribers and pharmacies would typically suggest more audit focus on them. In some cases, we have observed a strong transactional link between a prescriber and a pharmacy both of which have been identified as having abnormal behavior. Future work will explore systematic approaches to identify such links.

Typical application of our methodology utilizes history windows from three months to one year. The analysis window tends to range from one month to three months for large health care plans (and would typically not exceed a year). The relatively short analyses windows would not get impacted significantly with the normal progressions in medical practices except when a disruptive and abrupt change in practice happens related to the focus drug being analyzed.

We reiterate that the approach and methodology described in this paper for prescription claims data, can be extended to many other focus areas in health care claims data. Since the data for prescription

and medical claims vary across fee-for-service health plans, our approach will have to be adapted to the data elements that are available in each plan. A key aspect of our approach is to construct and use detailed profiles for all entities (patient, prescriber and pharmacy) to condition the expected normal behavior, and on statistical significance testing to detect abnormal behavior for each entity. Hence, the availability of detailed entity profiles in the data will determine the performance of our baseline models and the identification of abnormal entities.

7 Summary

To summarize our approach, a given focus area (e.g., prescription rate in a certain drug therapeutic class) is selected for audit analysis, and baseline models with the appropriate normalizations are constructed to describe the expected behavior within the focus area. These baseline models are then used, in conjunction with statistical hypothesis testing, to identify entities whose behavior diverges significantly from their expected behavior according to the baseline models. A Likelihood Ratio score over the relevant claims with respect to the baseline model is obtained for each entity, and the p-value significance of this score is evaluated to ensure that the abnormal behavior can be identified at the specified level of statistical significance. In particular, our approach is designed to be used as a preliminary computer-aided audit process in which the relevant entities with the abnormal behavior are identified with high selectivity for a subsequent human-intensive audit investigation.

8 Acknowledgements

We thank the health plan staff responsible for audits and fraud prevention and detection for their support and guidance during this project. We also thank the reviewers for their feedback and suggestions.

References

[1] K. D. Aral, H. A. Güvenir, İ. Sabuncuoğlu and A. R. Akar, *A Prescription Fraud Detection Model*, Computer Methods and Programs in Biomedicine, 106 (2012), pp. 37–46.

[2] R. A. Becker, C. Volinsky, and A. R. Wilks, *Fraud Detection in Telecommunications: History and Lessons Learned*, Technometrics, 52(1) (2010), pp. 22–33.

[3] R. J. Bolton and D. J. Hand,, *Unsupervised Profiling Methods for Fraud Detection*, Credit Scoring and Credit Control VII, Edinburgh, U.K., 2001.

[4] R. J. Bolton and D. J. Hand, *Statistical Fraud Detection: A Review (with discussion)*, Statistical Science, 17(3) (2002), pp. 235–255.

[5] Clinical Classifications Software for ICD-9-CM, `http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp#download`.

[6] W. Cohen, *Fast effective rule induction*, Proceedings of the Twelfth International Conference on Machine Learning (1995), pp. 115–123.

[7] D. J. Hand, *Fraud Detection in Telecommunications and Banking: Discussion of Becker, Volinsky and Wilks (2010) and Sudjianto et al.(2010)*, 52(1), pp. 34–38 (2010).

[8] D. E. Hoffmann and A. J. Tarzian, *Achieving the Right Balance in Oversight of Physician Opiod Prescribing for Pain: The Role of State Medical Boards*, Journal of Law, Medicine & Ethics, 31 (2003), pp. 21–40.

[9] D. W. Hosmer, T. Hosmer, S. Le Cessie and S. Lemeshow, *A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model*, Statistics in Medicine, 16 (1997), pp. 965–980.

[10] D. A. Hyman, *Health Care Fraud and Abuse: Market Change, Social Norms, and the Trust "Reposed in the Workmen"*, Journal of Legal Studies, vol. XXX (June 2001), pp. 531–567.

[11] J. A. Inciardi, H. L. Surratt, S. P. Kurtz and T. J. Cicero, *Mechanisms of prescription drug diversion among drug-involved club-and street-based populations*, Pain Medicine, 8(2), (2007), pp. 171-183.

[12] V. Iyengar, I. Boier, K. Kelley and R. Curatolo, *Analytics for Audit and Business Controls in Corporate Travel and Entertainment*, In Proc. Sixth Australasian Data Mining Conference (AusDM 2007), Gold Coast, Australia. CRPIT, 70 (2007), pp. 3–12.

[13] M. Kulldorff, *A Spatial Scan Statistic*, Commun. Statist. Theor. Meth., 26(6) (1997), pp. 1481–1496.

[14] M. Kulldorff, F. Mostashari, L. Duczmal, K. Yih, K. Kleinman, R. Platt, *Multivariate spatial scan statistics for disease surveillance*, Statistics in Medicine, 26 (2007), pp. 1824–1833.

[15] T. L. Leap, *Phantom Billing, Fake Prescriptions, and the High Cost of Medicine*, Cornell University Press, Ithaca NY, 2011.

[16] J. Li, K. Y. Huang, J. Jin, J. Shi, *A survey on statistical methods for health care fraud detection*, Health Care Management Science, 11(3) (2008), pp. 275–287.

[17] National Institute On Drug Abuse, *Prescription Drugs: Addiction and Abuse*, Research Report Series, National Institutes of Health (NIH) Publication Number 11-4881, revised 2011.

[18] C. Phua, V. Lee, K. Smith and R. Gayler, *A Comprehensive Survey of Data Mining-based Fraud Detection Research*, Artificial Intelligence Review (2005), pp. 1–14.

[19] J. R. Quinlan and R. M. Cameron Jones, *FOIL: A midterm report* Machine Learning: ECML 93, LNCS, Springer Verlag, 667 (1993).

[20] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, Wiley Series in Probability and Statistics, 2005.

[21] A. Sudjianto, S. Nair, M. Yuan, A. Zhang, D. Kern and F. Cela-Diaz, *Statistical Methods for Fighting Financial Crimes*, 52(1) (2010), pp. 5–19.

[22] P. Travaille, R. M. Muller, D. Thornton and J. van Hillegersberg, *Electronic Fraud Detection in the U.S. Medicaid Healthcare Program: Lessons Learned from other Industries*, Proceedings of the Seventeenth Americas Conference on Information Systems, Detroit, Michigan, (2011).

[23] United States General Accounting Office, *Health Care Fraud. Schemes to Defraud, Medicare, Medicaid and Private Health Insurers*, GAO/T-OSI-00-15, 2000.