

Detecting Non-compliant Consumers in Spatio-Temporal Health Data: A Case Study from Medicare Australia

K.S. Ng[†], Y. Shan[†], D.W. Murray[†], A. Sutinen[†], B. Schwarz[†], D. Jeacocke[‡], J. Farrugia[‡]

[†] *Intelligence Services and Target Identification Portfolio, Medicare Australia*

[‡] *Health Advisory Branch, Medicare Australia*

Email: wayne.murray@medicareaustralia.gov.au

Abstract—This paper describes our experience with applying data mining techniques to the problem of fraud detection in spatio-temporal health data in Medicare Australia. A modular framework that brings together disparate data mining techniques is adopted. Several generally applicable techniques for extracting features from spatial and temporal data are also discussed. The system was evaluated with input from domain experts and was found to achieve high hit rates. We also discuss some lessons drawn from the experience.

Keywords—fraud detection; spatio-temporal data; health data; propositionalisation; local outlier factor; sequence prediction.

I. INTRODUCTION

Medicare Australia is a government agency that administers two fee-for-service national health funding programs for Australians. The first of these is called the Medicare Benefit Scheme (MBS) and covers medical services. The second is called the Pharmaceutical Benefit Scheme (PBS) and covers drug prescriptions. According to its 2008-9 annual report, Medicare Australia processed over 490 million MBS and PBS transactions and paid approximately A\$22 billion in benefits in the 2008-9 financial year. Medicare Australia has a role to minimise the occurrence of inappropriate billing and fraud in its administered programs. According to Medicare Australia’s National Compliance Program 2009-2010, “more than \$6.18 million in incorrect payments to providers, pharmacists and members of the public” was identified in that period [1].

Over the last ten years or so, Medicare Australia has developed, either in-house or through external engagement, several data mining systems [2]–[4] for identifying doctors and pharmacies that engage in fraudulent or inappropriate billing. In this paper, we turn our attention to the problem of identifying non-compliant consumers. In particular, we will discuss the design and implementation of a Consumer Review Assessment System (Consumer RAS). The terms consumer and patient are used interchangeably in this paper.

Consumer fraud/non-compliance can come in different forms. A major one is prescription shopping, which refers to the process of consumers targetting doctors and pharmacies to obtain prescription drugs, particularly benzodiazepines and narcotics, in excess of medical needs by deception. We

will focus on the problem of detecting such activities in this paper, but note that the methodology is more widely applicable.

The paper is organised as follows. Sect. II describes the consumer-related data that is available from Medicare Australia’s systems for this analysis. Sect. III presents some background knowledge on the prescription shopping problem. Sect. IV explains the overall analytical approach we take. Sect. V discusses several generally applicable techniques for extracting useful features out of spatial and temporal data. We then present our anomaly-detection algorithms in Sect. VI, followed by an evaluation in Sect. VII. Discussion and conclusion appear in Sects. VIII and IX.

II. AVAILABLE DATA

Consumer records are stored in three different databases: one consumer directory for general information, one database for MBS claims information, and one database for PBS claims information. There are legislative constraints on the ability of Medicare Australia to link MBS and PBS claiming data. As such, we are only allowed to use either the MBS or the PBS data in the design of Consumer RAS; we analysed the PBS data in this study because the main compliance issues of concern to us are related to the abuse of PBS prescriptions.

In the PBS database, each consumer has associated with him/her a set of transactions each taking the general form

$$(PhID, PrID, \{(Item, Cost)\}, Dos, Dop),$$

where $PhID$ is the identifier of the pharmacy at which the drugs were supplied, $PrID$ is the identifier of the prescribing doctor (henceforth called prescriber), $\{(Item, Cost)\}$ is the set of PBS items charged together with their costs to Medicare Australia, Dos is the date of supply, and Dop is the date of prescribing. We have data on individual consumers like age, sex, address, etc. We also have the address of each pharmacy and the specialty of each prescribing doctor. On top of that, we have detailed information on each PBS item like the drug’s scientific name, its manufacturer, Nordic code, treatment mode, etc, although these are available only as semi-structured English texts.

There are spatial and temporal dimensions in Medicare Australia’s consumer-related data. Spatially, we can geocode consumer, pharmacy, and prescriber addresses to work out, for example, the kind of distances people travel to obtain different services. Temporally, we can form time-series data for each consumer by grouping their PBS charges into different episodes of care and then ordering the transactions chronologically within each episode. Along the temporal dimension, we can also study the intervals between dates of supplies and the frequency of those visits.

Medicare Australia currently keeps five years’ worth of data for each consumer. This is sufficient for the purpose of identifying anomalies. Our main problem is that we have no labelled data; the default analysis mode is thus unsupervised learning. To deal with this challenge, we need to bring all the domain knowledge we have to bear on the problem, and this is elaborated in the next section.

III. SOME DOMAIN KNOWLEDGE

There are three main classes of drugs that are susceptible to abuse by prescription shoppers: opioids, benzodiazepines, and psychostimulants. Table I gives a list of drugs of potential abuse available through the PBS system. Within these drug groups, the targetted items made up 6.86% of all PBS prescriptions (in number-of-transactions terms) and cost Medicare Australia approximately \$415 million in the 2008-9 financial year. The relative percentages of the different drugs in number-of-transactions terms are also shown in Table I. We regularly revise the list of drugs to incorporate new developments.

Name	Class	Usage	Cost
Temazepam	Benzodiazepine	13.51%	\$6.3m
Diazepam	Benzodiazepine	11.99%	\$5.1m
Oxazepam	Benzodiazepine	7.46%	\$3.4m
Olanzapine	Benzodiazepine	6.54%	\$143.5m
Nitrazepam	Benzodiazepine	3.41%	\$1.7m
Quetiapine	Benzodiazepine	3.21%	\$66.2m
Alprazolam	Benzodiazepine	2.89%	\$5.0m
Clonazepam	Benzodiazepine	0.27%	\$0.7m
Codeine	Opioids	17.77%	\$16.7m
Oxycodone	Opioids	13.36%	\$57.4m
Tramadol	Opioids	4.51%	\$8.9m
Buprenorphine	Opioids	4.74%	\$25.5m
Morphine	Opioids	3.78%	\$27.2m
Fentanyl	Opioids	2.58%	\$30.2m
Methadone	Opioids	0.61%	\$2.4m
Hydromorphone	Opioids	0.04%	\$0.4m
Methylphenidate	Psychostimulant	2.34%	\$12.5m
Dexamphetamine	Psychostimulant	0.98%	\$2.0m

Table I
DRUGS OF POTENTIAL ABUSE

Prescription shopping is an important problem from the financial perspective, but we must also acknowledge the significant social cost that accompanies the problem.

The following is a compilation of variables whose values can be used as indicators of prescription shopping as identified by domain experts from within Medicare Australia

and external sources like [5], [6], and [7]. The term service provider refers to both prescribers and pharmacies in the following.

- Number of distinct service providers visited and the distribution of such service providers
- Seeing multiple service providers on the same day
- Time intervals between visits to service providers
- Percentage of pharmacy visits for drugs of concern
- Use of multiple drugs of concern, or poly drug use
- Drug supplies in excess of medical needs
- Duration of drug use and unusual dose escalations
- Distance travelled to service providers
- Behaviour predictability
- Inappropriate co-occurrences of PBS items in an episode of care.
- History of alcohol addiction and family history of substance abuse (*)
- Psychological disease and other mental problems (*)
- Prescription without consult (*)
- Lack of recent or appropriate corresponding MBS benefits to justify PBS benefits (*)

The four variables marked with an asterisk (*) are those we cannot use currently because they either require data from the MBS system or are simply not available from our data.

IV. ANALYTICAL CONSIDERATIONS

Many of the variables identified in Sect. III are somewhat predictive of inappropriate/fraudulent behaviour when they take on certain values, but individually they are usually not significant. The main challenge for us is to work out a disciplined way of taking all the different variables into account in judging whether a consumer is inappropriately utilising PBS medications or committing possible fraud, and do that in the absence of labelled data.

We saw earlier in Sect. II that Medicare Australia’s consumer data is semi-structured, multi-relational, and contains both spatial and temporal dimensions. There have been many techniques presented in the data mining literature on handling one or more (though almost never all) of these dimensions in the past, from learning from structured data [8] to spatial data mining [9], temporal data mining [10], spatio-temporal data mining [11], and multi-relational data mining [12]. There are essentially two classes of algorithms. The first are direct methods tailor-made to work on structured data in their original forms. For our purpose, we find the indirect approach, also known as the propositionalisation approach in some quarters [12, Ch. 11], attractive.

The idea of propositionalisation is to use systematic feature construction/extraction techniques to transform multi-relational, multi-dimensional data into feature vectors in \mathbb{R}^n and then plug in well-understood and efficient algorithms that exploit the geometry of \mathbb{R}^n from the standard data mining toolbox to achieve our goals. Propositionalisation is of course not a new idea; researchers in signal processing,

computer vision, and statistical machine learning more generally have been doing this in one form or another for years.

The key advantages of the propositionalisation framework are its simplicity and modularity. The framework is highly modular: different feature construction techniques and data mining algorithms can be tried to find the best combination for the task at hand. The modularity also allows quick and flexible changes every time we encounter new and additional complexities in a data mining project. This occurs often in real life. Simplicity is also important in the industrial setting because of the need for data analysts to engage and communicate with different stakeholders, most of whom have no training in computer science and/or statistics.

Propositionalisation pushes most of the complexities of data mining into the process of constructing good relevant features. This is the focus of the next section, which describes the way we propositionalise Medicare Australia’s consumer data. The techniques presented constitute a key contribution of the paper; they have wider applications beyond the specific problem discussed in this paper.

V. PROPOSITIONALISING SPATIO-TEMPORAL DATA

Most of the fraud indicators identified in Sect. III can be computed using simple SQL queries from Medicare Australia databases and represented in propositional form. We will focus on ways of propositionalising the spatial and temporal dimensions of consumer records in this section because they are not obvious. However, we do not want to give the impression that the spatial and temporal factors are somehow more important than others identified in Sect. III; in fact, some of the simplest variables like multiple visits to service providers within the same day and percentage of pharmacy visits for drugs of concern have proved valuable in helping to identify fraudulent activities.

A. Quantifying the Spatial Data Component

For a while now, compliance officers in Medicare Australia have postulated that some prescription shoppers are willing to travel large distances to different service providers to avoid detection, and that a spatial analysis of consumer data would allow us to detect such behaviour. We propose a scheme to quantify spatial behaviour in the following. We note that it is insufficient to compute the raw distance travelled by each consumer since people living in rural areas would naturally need to travel larger distances to get to their service providers. Also the relative attractiveness of different pharmacies needs to be considered. We need a model that captures how consumers pick their service providers based on the choices available to them in different locations.

The Huff Model The Huff model [13] is a simple stochastic gravity model for studying the effects of distance on consumer choices. It captures the basic intuition that, adjusting for shop attractiveness, consumers will have a higher tendency to visit a shop that is close by than one that is far

away. Formally, the Huff model specifies, for each consumer, a probability distribution over the shops the person is likely to visit as follows:

$$Huff(p|c) = \frac{1}{K_c} \frac{attr(p)}{d(c,p)^v}, \quad (1)$$

where $attr(p)$ is a measure of the attractiveness of shop p , $d(c,p)$ is the distance between consumer c and shop p , and $K_c = \sum_j attr(j)/d(c,j)^v$, j ranging over all shops, is a normalisation constant that makes sure the probabilities sum to one. Note that we use the usual $\Pr(\cdot|\cdot)$ notation for probability but write the name of the density function in place of \Pr for clarity.

The expression $attr(p)/d(c,p)^v$ in (1) is intended to capture the economic utility of p to c . The parameter v is used to account for the kind of distances people are willing to travel in different contexts; e.g. people are usually willing to travel further for the purchase of expensive products like furniture. The lower v is, the less distance matters; in particular, distance does not matter at all when $v = 0$.

A Modified Huff Model The original Huff model is proposed as a way of estimating the trading area of supermarkets [13]. More recently, it has been used to monitor fraud in public delivery programs [14]. Here, we will use the model to study how consumers choose their pharmacies since this is one case where the underlying assumptions of the Huff model clearly hold: the commodity-like service of most pharmacies makes the distance to pharmacy an important factor. Note that this analysis would not be appropriate for prescribers because people go to different doctors for a myriad reasons, often unrelated to distance.

We will use the general form of Equation (1) to specify the probability of a consumer visiting a particular pharmacy. There are different factors that determine the general attractiveness of a pharmacy. We measure the attractiveness of a pharmacy p using the number of distinct consumers $\#(p)$ that visited p in a given period; i.e. we define

$$attr(p) = \#(p).$$

We feel foot traffic is a reliable measure of attractiveness.

The distance between a consumer and a pharmacy needs more careful thought because it should not be interpreted simply as physical distance but some measure of general accessibility. For example, a pharmacy far away from a consumer’s home may still be visited frequently if it is located right beside his/her doctor, work place, or a railway station the person visits daily travelling to and from work. We define the distance between a consumer c and a pharmacy p to be

$$d(c,p) = \min\{d(s,p) : s \in doctors(c) \cup \{c\}\},$$

where $d(s,p)$ is the great-circle distance between s and p calculated using the Haversine formula (ideally, we want the road distance between s and p but $d(s,p)$ is not a bad approximation), and $doctors(c)$ is the set of doctors that

has treated c in the period. The intuition behind the use of $doctors(c)$ is that consumers are likely to visit a nearby pharmacy straight after visiting a doctor (who may be far away from the consumer) and keep going back to the same pharmacy for familiarity reasons. Unfortunately we do not have information on consumers’ work addresses.

In our case, the appropriate value of the parameter v depends on where the consumer lives. For example, consumers living in outback Australia would have a lower v value compared to someone staying right in the heart of Sydney. The v parameter is estimated from data on a postcode-by-postcode basis using maximum likelihood. The values used range from one to three.

As a sanity check, we confirmed that the vast majority of consumers do indeed exhibit spatial behaviour consistent with the Huff model.

A Deviation Measure Note that there is one Huff model per consumer, and the probabilities of different consumers under different probability distributions are clearly not comparable. What we need to do instead is to use, for each consumer c , the conditional probability distribution $Huff(\cdot | c)$ to define some notion of normal behaviour and then compute deviations from that normal behaviour. There are several possible measures; we found the following information-theoretic measure works well.

Consider the mode $p^* = \arg \max_p Huff(p | c)$ of the $Huff(\cdot | c)$ distribution. Let p be a pharmacy chosen by c . The measure

$$\log_2 Huff(p^* | c) - \log_2 Huff(p | c), \quad (2)$$

which is non-negative for all p , tells us how the consumer’s choice deviates from the most probable choice. Specifically, since we know from information theory [15] that $-\log_2 Huff(p | c)$ gives the number of bits needed to encode p , the formula (2) gives us the number of extra bits we need to encode p over p^* . Extending (2) to a list of consumer c ’s pharmacy choices in chronological order $\bar{p} = [p_1, p_2, \dots, p_n]$ yields

$$\delta(c, \bar{p}) = \frac{1}{n} \sum_{i=1}^n (\log_2 Huff(p^* | c) - \log_2 Huff(p_i | c)).$$

Clearly, $\delta(c, [p^*, p^*, \dots, p^*]) = 0$; also, the higher the δ score is, the more a consumer’s spatial behaviour deviates from the Huff model.

B. Quantifying the Temporal Data Component

In talking to domain experts, the ‘normality’ of the item sequences charged by a consumer in different episodes of care is an important consideration in their assessment of whether the consumer in question is fraudulent/non-compliant. The word normality is in quotes because it is, as used by the domain experts, an imprecise concept, related among other things to the (unusual) co-occurrence

of items, the relative frequencies of different items in a sequence, the lengths of certain drug prescriptions, the time lag between visits to care providers, and a myriad of other factors. The experts’ innate understanding of the temporal data comes from many years of clinical experience and the challenge for us is to find a way to capture and quantify that understanding.

Overall Approach To solve that problem, one can attempt to construct a probabilistic expert system but that can be a tedious and time-consuming process. Instead, we will take a completely data-driven approach, guided by the massive data collection of Medicare Australia and “the unreasonable effectiveness of data” [16]. The idea is to start with a representative set of item sequences taken from the community. From that we construct a probability distribution over items conditioned on what precedes them and use that to measure how normal a given sequence is with respect to what everybody else in the community is doing, *no matter* what they are doing.

More formally, let Σ be a finite set of symbols and denote by Σ^* the set of all finite sequences of symbols in Σ . The technical problem we are primarily concerned with is this: given a sequence $x_{1:t} = x_1 x_2 \dots x_t \in \Sigma^*$, how do we predict what the next symbol x_{t+1} is? For that, we need to know the probability

$$\Pr(x_{t+1} | x_{1:t}) \quad (3)$$

for all possible $x_{1:t}$ and x_{t+1} . We call the sequence $x_{1:t}$ we condition on in (3) the *context* in the following. Armed with (3), the predictability of a sequence $x_{1:n}$ can be quantified with a rather natural measure like

$$\begin{aligned} M(x_{1:n}) &= -\frac{1}{n} \log_2 \Pr(x_{1:n}) \\ &= -\frac{1}{n} \log_2 \prod_{i=1}^n \Pr(x_i | x_{1:i-1}) \\ &= -\frac{1}{n} \sum_{i=1}^n \log_2 \Pr(x_i | x_{1:i-1}), \end{aligned}$$

where we define $x_{i:j}$, $i > j$, to be the empty sequence ϵ . From information theory, we know $-\log_2 \Pr(x_{1:n})$ is the number of bits required to encode the sequence $x_{1:n}$ under an optimal coding scheme. The measure $M(x_{1:n})$ therefore tells us, on average, how many bits are required to encode each x_i in the sequence. This is a reasonable one-number summary of the sequence, but we note that it does not capture information like the variance of individual conditional probabilities. The use of log above also has the advantage of side-stepping potential floating-point underflow issues with low probability values. A few alternative measures to $M(\cdot)$ are discussed in [17], [18].

Dealing with Multiple Episodes A typical consumer record is made up of multiple episodes of care, each represented

by an item sequence. There are significant issues in reliably estimating an episode of care from claims data. Here, we sidestep those issues and take the approach of defining all transactions that relate to a script as an episode of care; this is quite sufficient for our purpose.

There are different possible ways to handle multiple sequences; the following generalisation of $M(\cdot)$ to (variable-length) lists of sequences seems rather intuitive:

$$M([s_1, s_2, \dots, s_m]) = -\frac{1}{\sum_i |s_i|} \left(\sum_{i=1}^m \log \Pr(s_i | s_{i-1}) \right),$$

where $s_i \in \Sigma^*$, $s_0 = \epsilon$, $|s|$ denotes the length of s , and

$$\Pr(x_{1:t} | y_{1:n}) = \Pr(x_1 | y_{1:n}) \prod_{i=2}^n \Pr(x_i | y_{1:n}, x_{1:i-1}).$$

In the above we are assuming that each item in an episode depends on what comes before it in the episode as well as the items in the last episode; in other words, the probabilities are Markovian up to the last episode. Ideally, we want to condition on all previous episodes but our assumption may be acceptable from the point of view of simplifying the associated statistical estimation problems.

Algorithm Computationally, our whole scheme comes down to our ability to estimate (3) from data. When the alphabet Σ is small, there are many efficient algorithms for doing that; see e.g. [17], [19]. For us, the main drawback of such algorithms is their space complexity, which is usually of the order $O(|\Sigma|^D)$, where D is the maximum length of the context we want to condition on. This is impractical in our case, where Σ is the set of all PBS items and $|\Sigma|$ is in the thousands. It is conceivable that some of the mentioned algorithms can be modified to deal with non-trivial alphabets, but we found the problem of estimating (3) from data can be solved with a minor modification of the random-text-generation algorithm presented in [20]. We now describe the algorithm.

There are two phases: setup and prediction. The setup phase amounts to reading and storing the entire training data in a suffix array, which is an array of integers giving the starting positions of suffixes of a string in lexicographic order. An example will help clarify the construction. Consider the following training data:

$$[23Y, 23Y, 11T, 18V, 23Y, 72U]. \quad (4)$$

Immediately after reading the sequence into a *word* suffix array, we have

$$\begin{aligned} \text{word}[0] &= [23Y, 23Y, 11T, 18V, 23Y, 72U] \\ \text{word}[1] &= [23Y, 11T, 18V, 23Y, 72U] \\ \text{word}[2] &= [11T, 18V, 23Y, 72U] \\ \text{word}[3] &= [18V, 23Y, 72U] \\ \text{word}[4] &= [23Y, 72U] \\ \text{word}[5] &= [72U], \end{aligned}$$

where each entry in the *word* array points to a suffix of (4). We now sort the *word* array in lexicographic order to group suffixes sharing the same leading items together, yielding

$$\begin{aligned} \text{word}[0] &= [11T, 18V, 23Y, 72U] \\ \text{word}[1] &= [18V, 23Y, 72U] \\ \text{word}[2] &= [23Y, 11T, 18V, 23Y, 72U] \\ \text{word}[3] &= [23Y, 23Y, 11T, 18V, 23Y, 72U] \\ \text{word}[4] &= [23Y, 72U] \\ \text{word}[5] &= [72U]. \end{aligned}$$

The sorted suffix array now allows us to efficiently find all the items that come after a certain context in the training data. For example, from the entries of *word*, we know $\{11T, 23Y, 72U\}$ are the only items that follow an item $23Y$ in the training data.

The time complexity of the setup phase is low: $\Theta(n)$ to read in the n items in the training data, and $O(dn \log n)$ to sort the n suffixes, where d is the length of the largest context we want to use, a quantity that determines the cost of comparing two suffixes.

We next describe the prediction phase. Given a sequence $x_{1:t}$, to compute $\Pr(x_i | x_{1:i-1})$, we first perform a binary search to find the first occurrence of $x_{1:i-1}$ in the *word* suffix array. We then scan through *word* to count the number of times x_i occurs straight after $x_{1:i-1}$; i.e. we assign

$$\Pr(x_i | x_{1:i-1}) = \frac{\#(x_{1:i})}{\#(x_{1:i-1})},$$

where $\#(s)$ is the number of times s occurs in the training data. To make sure the probability estimate is reasonable, we need to make sure the number $\#(x_{1:i-1})$ is sufficiently large. From the Central Limit Theorem, we know that to obtain an estimate within ϵ of the true value with, say, 95% confidence, we need $\#(x_{1:i-1}) \geq 1/\epsilon^2$.

Given a training set consisting of n items and an item sequence $x_{1:t}$, computing $M(x_{1:t})$ takes time $O(t(n + \log n))$, since the probability estimate for each item involves a binary search that takes time $O(\log n)$ and a subsequent scan through the sorted suffix array takes time $O(n)$.

Other Temporal Data The PBS item sequences are not the only consumer data component with a temporal aspect. The above analysis technique is also applicable to the sequence of pharmacies visited in each episode of care, and the time intervals between successive visits to service providers.

C. Putting It All Together

Putting together all the variables identified in Sect. III and Sect. V, we arrive at a fairly good picture of the general behaviour of a consumer. To help our domain experts, English descriptions of key characteristics like high volume of drugs of concern, lack of usual doctors/pharmacies, poly drug use, highly predictable item charges, etc. can be easily extracted

from the values of (combinations of) these variables. We emphasise the variables merely summarise consumer’s behaviour; they are not used on their own to positively identify anomalies; that is the job of the algorithms described next.

VI. ANOMALY DETECTION ALGORITHMS

There are two classes of anomalous consumer records: those we know about and can positively characterise; and those that fall in the miscellaneous unknown class, a subset of which we can characterise via deviation from statistical norms. The two classes need to be handled differently, one driven by domain knowledge and the other, data. We will look at the first class next, followed by statistical outlier detection techniques for handling the second class.

A. Positive Characterisation of Prescription Shopping

A young person with a high volume of drugs of concern that also triggers some of the following flags, all of which can be measured either directly or with the techniques discussed earlier in Sect. V, is highly likely to be engaging in prescription shopping.

- high entropy for distribution of prescribers
- high entropy for distribution of pharmacies
- highly predictable PBS item sequences
- unpredictable pharmacy sequences
- seeing multiple pharmacies in the same day
- large deviation from the Huff model

The entropy of a probability distribution measures how spread out the distribution is (the uniform distribution has the maximum entropy value). High entropy values for a consumer’s prescriber and pharmacy distributions suggest the consumer does not have usual doctors/pharmacies and is thus unlikely to be genuinely sick or known to the different practitioners that serve him. That, in combination with unpredictable pharmacy choices and occurrences of multiple pharmacies in the same day, are believed to be characteristics of a prescription shopper. Prescription shoppers also tend to want only certain drugs and exhibit highly predictable PBS item sequences. Lastly, males are believed to be more likely than females to engage in prescription shopping.

Based on the above, it is easy to construct an appropriate set of rules to identify persons of interest. We have not spelled out the exact rules and parameters we use in this paper to avoid deliberate attempts by consumers to circumvent these rules by flying *just* under the radar.

B. Outlier Detection

There are several approaches for detecting statistical outliers in data [21]. We can use indirect methods that attempt to construct a model of the data using clustering algorithms and then use deviation from that model to pick up outliers. Popular such algorithms include k-means and algorithms for learning finite mixture models. In contrast, direct methods

like [22] attempt to pick out outliers without going through the model-building step. We experimented with different algorithms and found the local outlier factor algorithm [23] to be effective on our problem. To make the paper self-contained, we now briefly review the algorithm.

Local Outlier Factor Let D be a set of objects and let $d(p, q)$ denote the distance between any two $p, q \in D$. To simplify matters, we will assume no two objects in D have exactly the same representation; the original treatment in [23] is much more careful on this point but we feel the resultant formalisation obscures the underlying idea.

Definition 1. For each p in D , let $n_k(p)$ denote the k -th closest object to p in D and let $r_k(p) := d(p, n_k(p))$. The k -neighbourhood $N_k(p)$ of p in D is then defined by

$$N_k(p) = \{q \in D \setminus \{p\} : d(p, q) \leq r_k(p)\}.$$

Intuitively, $r_k(p)$ is the radius of $N_k(p)$.

Definition 2. The reachability distance of an object $p \in D$ with respect to object $o \in D$ is defined by

$$rd_k(p, o) = \max(d(p, o), r_k(o)).$$

Clearly, we have $rd_k(p, o) = d(p, o) \forall p \notin N_k(o)$. For every $p \in N_k(o)$, the $rd_k(\cdot, \cdot)$ measure maps $d(p, o)$ to the radius $r_k(o)$ of $N_k(o)$. This is effectively a smoothing operation to reduce the variance of $d(p, o)$ for all the p ’s close to o .

Definition 3. The local outlier factor of an object $p \in D$ is defined by

$$LOF_k(p) = \frac{1}{k} \sum_{q \in N_k(p)} \frac{\sum_{o \in N_k(p)} rd_k(p, o)}{\sum_{o \in N_k(q)} rd_k(q, o)}. \quad (5)$$

The ratio in (5) measures how different p is to each q in its neighbourhood in terms of the total reachability distance to their respective neighbours. The local outlier factor of p is then the average of all such ratios. If the individual ratios are all close to one, then we know p behaves like all its neighbours. On the contrary, if the individual ratios are all high values, we know p is a lot further from its neighbours than they are from each other. Thus $LOF_k(p)$ is a measure of the degree of outlierness of p , with higher values denoting higher degree of outlierness. Further, it is a local measure because each point is only compared to its local neighbours.

There remains the question of picking a suitable k value. [23] offers some guidelines that take into account the minimum number of objects a cluster has to contain and the maximum number of objects nearby a cluster that can potentially be considered outliers. Given lower and upper bounds L and U , [23] suggests ranking all objects by their maximum LOF scores: $\max\{LOF_k(p) : L \leq k \leq U\}$. Given the severity of the decisions we have to make, we choose the more conservative approach of using the minimum values.

VII. EVALUATION

Experimental Objectives We now discuss two experiments to validate the performance of our system. The experiments are designed to answer the following questions:

- 1) Is the characterisation of prescription shoppers given in Sect. VI-A accurate?
- 2) Can Consumer RAS be used to identify prescription shoppers that do not rigidly fit the strong criteria outlined in Sect. VI-A? The false identification of genuinely ill patients that exhibit certain characteristics of prescription shopping as prescription shoppers have in the past been an issue for Medicare Australia. Can Consumer RAS avoid making such errors?

Data To conduct the experimental study, we randomly picked a populous postcode in a major capital city of Australia in which to try to identify possible fraudulent activities. (The analysis was done on a postcode-by-postcode basis because of the spatial nature of the data.) There are around 30,000 people with registered address in the chosen postcode that have transactional records with Medicare Australia in the period of study. One year’s worth of transactions are extracted for each consumer, and those with low volume are removed from consideration (up to half the people can be removed this way). We also extracted data from six other large postcodes in a major Australian capital city to use as sample data for estimating the probability of elements in PBS item and pharmacy sequences. Each of these postcodes has around 50,000 people. We stress that the data do not contain identifying information like consumer names and addresses (we only have coordinates from geocoding).

Software We use the LOF implementation in the `dprep` package in R for our analysis. The modified Huff model and the temporal feature extraction scheme described in Sect. V are implemented in-house using C++. The main algorithms are all “embarrassingly parallel” in nature and are parallelised using OpenMP. The programs were run on a Linux machine with two quad-core CPUs and took less than two hours to process a large postcode.

Experiment I The first experiment seeks to verify the accuracy of our quantitative characterisation of prescription shoppers given in Sect. VI-A. Only 12 people in the chosen postcode satisfy our criteria. These are passed on to domain experts for evaluation. We are interested in the percentage of these people that are true prescription shoppers.

Experiment II The second experiment seeks to verify whether we can use the technique discussed in Sect. VI-B to identify, with low false positive rate, prescription shoppers that do not fit the criteria identified in Sect. VI-A. To do that, we remove all consumers identified in Experiment I from the data, perform a LOF analysis on the rest, and then pick out consumers that have high volumes of drugs of concern for evaluation by the experts. Fourteen consumers were picked

this way. Some of these consumers have genuine needs for their drugs. We want to know whether prescription shoppers tend to exhibit higher LOF scores compared to patients with genuine needs? In other words, do prescription shoppers show up as statistical outliers in the data?

We excluded older consumers in this experiment because they are more likely to have genuine medical conditions and we do not want to spend time analysing such patients. For the LOF analysis, the data are normalised and we compute the minimum LOF value with $k \in [20, 50]$. People identified in Experiment I were removed from the LOF analysis because we do not want to miss people who look *normal* with respect to that suspicious group.

Evaluation Methodology Two experts (a senior medical adviser and a senior pharmaceutical adviser) from within Medicare Australia evaluated the system. The medical adviser is a registered medical practitioner and the pharmaceutical adviser is a registered pharmacist.

The evaluation procedure is as follows. For each consumer, we laid out the PBS and MBS records side-by-side on two separate projector screens. A summary of the consumer based on the features described earlier was first presented. The advisers then examined the PBS and MBS records separately and wrote down their observations for each. On the MBS side, the advisers looked for items that indicate genuine medical conditions including chronic pain issues. Among other things, they look for the presence of pathology and diagnostic items that could indicate a genuine medical condition. Consultation types and their frequencies were also examined, with frequent consultations or multiple same-day consultations with different practitioners possibly indicating prescription shopping. For PBS records, the advisers looked for the types, quantities, and combinations of drugs used. When multiple drugs of concern were taken, they also checked whether the drugs are usually taken together for medical purposes or whether this could represent prescription seeking behaviour. The safety of those combinations of drugs was also considered. During the examination of the MBS and PBS claims data, unusual issues not directly related to prescription shopping were also noted. After all the observations were recorded, the two advisers made a judgement call on whether the consumer in question was a prescription shopper.

The whole process of evaluating 26 consumers identified in the chosen postcode took three (uninterrupted) hours. The time taken to evaluate each consumer is in the range of 3 to 15 minutes. While genuinely ill patients could be rapidly identified, those identified as likely prescription shoppers took longer to assess because possible legitimate reasons for their medication use need to be eliminated.

Results for Experiment I Table II shows the result of Experiment I. The column `PShopper` records whether the experts believe the consumer is a prescription shopper, and

Conf is their confidence in that view. The column Known shows whether the consumers are known to either Medicare Australia or their doctors to have possible addiction problems as indicated by the presence of items like “Drug monitoring program” in their MBS records. The column Other Issues records whether issues other than prescription shopping (e.g. doctor malpractice) were identified in the evaluation process. The column Action indicates whether further action is being taken with respect to each consumer.

ID	PShopper	Conf	Known	Other Issues	Action
1	No	-	No	No	No
2	Yes	High	No	Yes	Yes
3	Yes	Maybe	Yes	No	Yes
4	No	-	Yes	No	No
5	No	-	No	No	No
6	Yes	Maybe	Yes	No	Yes
7	Yes	High	No	No	Yes
8	No	-	No	Yes	Yes
9	Yes	High	No	No	Yes
10	Yes	Maybe	Yes	No	Yes
11	Yes	Maybe	Yes	No	Yes
12	Yes	High	No	No	Yes

Table II
EVALUATION OF EXPERIMENT I

By and large, the evaluation confirms the effectiveness of our criteria. Of the 12 people identified, 8 are believed to be prescription shoppers (a respectable 67% hit rate), 4 with high confidence and 4 maybe’s. We are primarily interested in prescription shoppers that are not known to Medicare Australia or their doctors/pharmacies. All four high-confidence prescription shoppers are such unknown cases. It is also interesting to note that the four maybe’s are known to *some* of their doctors/pharmacies, but they are still considered risky because their data exhibit behaviours that indicate intent to hide their drug prescriptions from the *other* doctors and pharmacies.

Two of the false positives (consumers 1 and 5) could have been avoided if our system was allowed to look at the MBS records in addition to the PBS records. In the first case, the patient’s large number of pathology items suggest possible arthritic conditions. In the second case, it is clear the patient has a demyelinating (neurological) disease from the MBS records. These examples provide a measurement of the performance loss our system suffers because of the legislative constraint on linking MBS and PBS data.

The other two false positives (consumers 4 and 8) are also instructive. Consumer 4 obtained a large number of Codeine and also some Alprazolam and Olanzapine. The amount of Codeine obtained is deemed to be unusually high but not sufficient to cause harm. Consumer 8 also has a large number of Codeine prescriptions (once every 2-3 days for more than a year); in this case, the consumer is deemed not to be a prescription shopper but the experts recommend that his main doctor be investigated. The initial inclusion of Codeine as a drug of concern was a debatable choice; this evaluation suggests we need to drop Codeine from

the list of drugs of concern, in line with medical expert feedback indicating prescription shopping for this drug is not commonly encountered by general practitioners. This change will likely push the system’s hit rate closer to 80%.

Experiment I also produced a somewhat unexpected discovery: the MBS record of Consumer 2 exhibits a large number of dental items, and the number of X-ray operations recorded, if they were actually performed, are at a level capable of causing cancer. We have reasons to believe this is not an isolated case and a new line of investigation has been initiated.

Results for Experiment II Table III shows the evaluation result of Experiment II. There is a clear relationship between LOF values and prescription shopping behaviour, confirming that some prescription shoppers are indeed statistical outliers. The other encouraging result is that, to a large extent, genuine patients can be identified from their low LOF values.

ID	LOF	PShopper	Conf
13	2.19	No	-
14	2.12	Yes	High
15	1.5	Yes	High
16	1.45	No	-
17	1.42	Yes	High
18	1.35	No	-
19	1.33	No	-
20	1.29	Yes	Maybe
21	1.26	No	-
22	1.2	No	-
23	1.17	No	-
24	1.16	No	-
25	1.14	No	-
26	1.11	No	-

Table III
EVALUATION OF EXPERIMENT II

Figure 1 shows the ROC curve for different choices of LOF values as thresholds for identifying prescription shoppers. In our setting, the cost of false positives is higher than false negatives (falsely accusing a genuine patient of prescription shopping is more costly than missing some real prescription shoppers). Under a range of such cost functions, the experiment suggests a LOF value of ~ 1.4 is a suitable threshold for identifying prescription shoppers. We note however that the data comes from one postcode and this has to be verified with data from other postcodes.

The three consumers judged to be prescription shoppers (consumers 14, 15 and 17) were not picked in Experiment I because one of them is quite old and the other two have primary doctors that prescribe more than 80% of their drugs. Having primary doctors is usually a good sign (indeed most of the genuine patients in Table III have primary doctors), but in the latter two cases, the primary doctors themselves are, in the view of the experts, suspect. Consumer 20 is one of those borderline cases: the person has 12 different prescribers and 11 different pharmacies, although the primary prescriber and primary pharmacy are responsible for 54% of all the drugs-of-concern transactions. The patient’s MBS

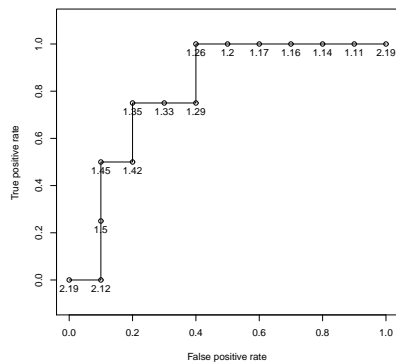


Figure 1. ROC curve for different LOF thresholds

record also show regular “Drug monitoring program” items.

VIII. DISCUSSION

General Observations The above confirms the effectiveness of the system at identifying prescription shoppers. We now remark on the usefulness of the spatial and temporal variables identified in Sect. V. The temporal variables turn out to play an important role in both the positive characterisation of prescription shopping and in the statistical detection of such behaviour via the LOF algorithm. The Huff variable, however, turns out not to be so useful, at least for detecting prescription shoppers.

None of the twelve confirmed prescription shoppers exhibit noteworthy deviation from the Huff model. There are several possibilities. The prior belief that prescription shoppers tend to travel large distances to different doctors and pharmacies to avoid detection may well not have support in the data. Another plausible explanation is that there are many pharmacies in and around the chosen postcode, and prescription shoppers need not travel far to avoid detection. In that sense, the Huff model may only come into play for more rural postcodes. The relevance of the Huff variable needs more investigation. There are a number of consumers that exhibit large deviations from the Huff model but who do not otherwise appear to be engaging in prescription shopping; these need to be investigated for other forms of possible fraud.

Further work is required to assess the potential application of this work within the Medicare Australia compliance framework. Although we have a high degree of confidence in the validity of our methodology for detecting prescription shoppers, it is presently unclear the extent to which the system could be used as a standalone and *only* method for identifying *all* prescription shoppers within a population. It is likely that the value of our approach lies in targeting higher risk prescription shoppers. The true extent of our false

negative rate with respect to the entire population, not just the targetted subset, needs to be quantified.

In ongoing work, we are evaluating the effectiveness of the system at a larger scale with many more postcodes. This process will necessarily take many months but we will have a much more complete picture by the end of it. We are also extending the system to detect overseas drug diversion, which refers to the taking or sending of PBS subsidised medicine out of Australia for reasons other than for personal use of the carrier or someone travelling with the carrier. The methodology described in this paper applies but with a different set of targetted drugs.

Limitations The main limitation for the system is not being able to see the MBS side of the story to augment what it can infer from a consumer’s PBS record. Such restrictions on linking MBS and PBS claims data are due to legislative requirements and will continue to pose difficulties for us.

A second limitation is that the system has been designed from the beginning to look at individual consumers. From a cost-benefit perspective, detection of a colluding group of consumers is clearly more useful. We may need a completely new approach for that, but we are hopeful the scheme of [24] can be adapted for use here. The idea is to identify natural groupings of people (e.g. consumers from the same address, consumers that visit the same set of doctors/pharmacies, etc) and then use Consumer RAS to work out the percentage of high-risk consumers in each group. Clearly, groups with unusually high percentages of high-risk consumers are those we are interested in. This idea is subject to further research.

Lessons Learned We now reiterate two time-honoured lessons we keep (re)learning in our data mining ventures:

- 1) It is *crucial* to get domain experts involved as early as possible in a data mining project. Actionable knowledge can only come from the right combination of domain insights and analytical techniques.
- 2) Real-world problems almost always have multiple dimensions and data miners need to be able to understand and utilise tools from diverse disciplines to tackle those problems. The main challenge is usually to find the right combination of tools; there is seldom a need to invent new algorithms.

Future Work In the months after the current system is deployed, we will slowly accumulate more and more labelled examples of non-compliant/fraudulent consumers. This will provide an opportunity for us to deploy on-line supervised learning algorithms to develop a system that can detect consumer fraud in real time. Such a system would not render Consumer RAS as described here obsolete because it would still have a role to play in identifying as yet unknown fraudulent/inappropriate trends and behaviours. In-depth social network analysis on identified communities of interest is another area of potentially fruitful work.

IX. CONCLUSION

This paper documents our experience with applying data mining techniques to the problem of detecting anomalies in semi-structured spatio-temporal health data. We conclude by summarising the main contributions of the paper.

- 1) A detailed description of a real-world application of data mining with its multitude of challenges is presented. The study resulted in a deeper understanding of prescription shopping and a system for detecting such activities. We believe the techniques discussed have wider applications in other organisations.
- 2) The propositionalisation methodology is shown to be an effective and modular framework to bring together disparate data mining techniques in the analysis of multi-dimensional spatio-temporal data.
- 3) A demonstration of how a simple economic model like the Huff model can be used to incorporate domain knowledge into the analysis of content-poor data.
- 4) A demonstration of how a simple sequence-prediction algorithm can be used to quantify, from data, notions of predictability and (in)appropriate co-occurrences of items in a sequence.

ACKNOWLEDGMENT

We thank Ken Yan for helpful clinical insights. Lannie Pomazak, Peter Thomson and Dr Jo-Anne Benson provided management support for this project; Peter also made many suggestions that improve the paper. We also thank Xiaoming Liu for his help in geocoding address data.

REFERENCES

- [1] Medicare Australia, "Medicare Australia's national complicity program 2009-2010," 2010. [Online]. Available: <https://www.medicareaustralia.gov.au/provider/business/audits/files/nation-compliance-program-2009-2010.pdf>
- [2] K. Yamanishi, J. ichi Takeuchi, G. J. Williams, and P. Milne, "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms," *Data Mining and Knowledge Discovery*, vol. 8, no. 3, pp. 275–300, 2004.
- [3] R. Pearson, D. W. Murray, and T. Mettenmeyer, "Finding anomalies in Medicare," *Electronic Journal of Health Informatics*, vol. 1, no. 1, 2006.
- [4] Y. Shan, D. Jeacocke, D. W. Murray, and A. Sutinen, "Mining medical specialist billing patterns for health service management," in *AusDM*, 2008, pp. 105–110.
- [5] A. Campbell, *The Australian Illicit Drug Guide*, 2001.
- [6] L. R. Webster and R. M. Webster, "Predicting aberrant behaviors in opioid-treated patients: Preliminary validation of the opioid risk tool," *Pain Med.*, vol. 6, pp. 432–442, 2005.
- [7] N. Katz, "Opioids: After thousands of years, still getting to know you," *Clin. J. Pain*, vol. 23, pp. 303–306, 2007.
- [8] T. Gärtner, J. W. Lloyd, and P. A. Flach, "Kernels and distances for structured data," *Machine Learning*, vol. 57, pp. 205–232, 2004.
- [9] S. Shekhar, P. Zhang, Y. Huang, and R. R. Vatsavai, "Trends in spatial data mining," in *Data Mining: Next Generation Challenges and Future Directions*. AAAI/MIT Press, 2003.
- [10] C. M. Antunes and A. L. Oliveira, "Temporal data mining: An overview," in *KDD Workshop on Temporal Data Mining*, 2001, pp. 1–13.
- [11] J. F. Roddick and M. Spiliopoulou, "A bibliography of temporal, spatial and spatio-temporal data mining research," *SIGKDD Explorations*, vol. 1, pp. 34–38, 1999.
- [12] S. Džeroski and N. Lavrač, Eds., *Relational Data Mining*. Springer, 2001.
- [13] D. L. Huff, "Defining and estimating a trading area," *Journal of Marketing*, vol. 28, pp. 34–38, 1964.
- [14] Y. Kim, "Using spatial analysis for monitoring fraud in a public delivery program," *Social Science Computer Review*, vol. 25, no. 3, pp. 287–301, 2007.
- [15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.
- [16] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, pp. 8–12, 2009.
- [17] G. Bejerano and G. Yona, "Variations on probabilistic suffix trees: statistical modelling and prediction of protein families," *Bioinformatics*, vol. 17, no. 1, pp. 23–43, 2001.
- [18] P. Sun, S. Chawla, and B. Arunasalam, "Mining for outliers in sequential databases," in *ICDM*, 2006, pp. 94–106.
- [19] D. Ron, Y. Singer, and N. Tishby, "The power of amnesia: Learning probabilistic automata with variable memory length," *Machine Learning*, vol. 25, no. 2, pp. 117–150, 1996.
- [20] J. Bentley, "Strings of pearls," in *Programming Pearls*, 2nd ed. Addison-Wesley, 2000, pp. 161–173.
- [21] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statistical Science*, vol. 17, no. 3, pp. 235–255, 2002.
- [22] E. Knorr and R. Ng, "Algorithms for mining distance-based outliers in large data bases," in *VLDB*, 1999, pp. 392–403.
- [23] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 93–104, 2000.
- [24] C. Cortes, D. Pregibon, and C. Volinsky, "Communities of interest," *Intell. Data Anal.*, vol. 6, no. 3, pp. 211–219, 2002.