Fraud Detection for Healthcare

Hoda Eldardiry
Palo Alto Research Center
hoda.eldardiry@parc.com

Juan Liu
Palo Alto Research Center
juan.liu@parc.com

Ying Zhang *
Palo Alto Research Center
yzhang@parc.com

Markus Fromherz
Xerox
markus.fromherz@xerox.com

ABSTRACT

Fraud detection in healthcare is an important yet difficult problem. We present a fraud screening solution to identify suspicious pharmacies from a large dataset of pharmacy claims. Our solution has stemmed from collaboration with medical claim investigators and proven usefulness to investigators by discovering real fraud cases. We focus on a concrete problem of probabilistic outlier detection from a feature set designed for pharmacy claims. Although the reported results are specific to pharmacy claims, this approach can be applied widely. We are currently extending the solution to fraud screening of more general medical claims and fraud detection in other verticals.

Categories and Subject Descriptors

H.1 [**Models and Principles**]: Systems and Information Theory; G.3 [**Probability and Statistics**]: Statistical software

General Terms

Algorithms, Design, Performance

Keywords

Fraud Detection, Anomaly Detection, Health Care Applications

1. INTRODUCTION

Data analytics has become increasingly important in almost every area of the economy. McKinsey's influential report on Big Data Analytics [2] lists healthcare in the United States as the most promising application domain for data analytics. Healthcare takes the top rank partly because of its financial importance – being a large segment in the US economy. Healthcare offers huge incentives for transformative technologies. From a technical perspective, the vast amount of healthcare data (insurance claims, health records, clinical data, provider information, etc.) presents unprecedented opportunities for automated data analytics solutions to dramatically improve productivity. At the same time, the diversity and complexity of healthcare data poses significant challenges to technology developers.

In this paper we present our preliminary effort on detection fraudulent activities from healthcare data. In general, fraud is a common problem in many sectors. Figure 1 lists the amount of improper payment in US government expenditure. In 2012, improper payments totaled about $120 Billion. Different color bars represent major government programs. Healthcare-related programs such as Medicaid and Medicare are apparently the most significant. The Institute of Medicine (IOM) estimates the annual loss to be due to fraud in the healthcare domain to be $75 Billion [3]. The magnitude of the fraud problem has attracted many resources from the healthcare industry, the data analytics industry, and research communities to develop fraud detection systems.

Despite the substantial incentives, the fraud detection problem is still far from being solved. Several challenges need to be addressed. Data is inherently big and complex. Data analytics solutions need to handle extreme size data sets, with terabytes of data, billions of lines of records, and millions of patients and providers. The diversity of medical data demands a coherent system to handle multiple modality data (clinical, diagnosis data, claims data, etc). Furthermore, it is often difficult to specify what is normal, let alone what is fraudulent and abnormal. In addition, fraud is often subtle and accompanied by purposeful cover-up actions. As a result, fraudulent actions manifest into a set of seemingly normal claims. It needs a lot of intelligence and effort to tackle the fraud detection problem.

Loosely speaking there are two categories of fraud detection approaches. One approach starts from domain knowledge (i.e., knowledge from subject matter experts such as doctors, nurses, and pharmacists) to design a set of fraud detection rules. Many commercial systems take this approach. This methodology works well in many occasions, but its performance is inherently limited by subject matter expert knowledge, which can be inaccurate and incomplete. Furthermore, new fraud patterns are constantly invented to circumvent the baked-in fraud detection rules. Static in nature, rule-based systems have difficulty keeping up with the

*Author's current affiliation: Google Inc., yingzhang@google.com

KDD-DMH'13, August 11, 2013, Chicago, Illinois, USA.

dynamic nature of fraud. The second category uses a data-driven method to identify normal patterns and deviations from the norm. This approach is more flexible and dynamic, but computationally intensive, as the search space for fraud is vast. We advocate a combined approach, where domain knowledge is used to guide the search, while data-driven machine learning methods are used to improve upon expert intuition to achieve better accuracy and reliability.

We are working with collaborators in Medicaid organizations and Xerox Services (which provides review and auditing services to a number of government healthcare programs and private sector health insurance companies). Our goal is to develop fraud detection capabilities to screen out suspicious activities in medical claims. The automated screening enables investigators to focus attention on a small set of suspect list, as opposed to the prohibitively large dataset. This leads to more targeted investigations, which in turn will save investigation cost and reduce loss to fraud. The output of our screening tool is a list of suspicious entities and/or activities, where suspicion is defined based on statistical rarity – an entity is suspicious if its behavior is statistically rare (i.e., improbable). There is a gap between suspicion and conviction. To convict someone of fraud, the investigator must prove fraudulent intent. This can only be done via thorough investigation. Our system is designed to help focus the investigations effort and not replace it.

We have designed a suite of claim screening capabilities to suit the need of investigators:

- *Outlier Identification.* Unsupervised learning method to screen entities (e.g., pharmacies) with behavior that is drastically different from other similar pharmacies.
- *Relational Analysis.* Analysis of relationships between entities (for instance, between doctors and pharmacies, or pharmacies and patients) to identify possible collusion.
- *Temporal Sequence Analysis.* Analysis of medical sequences (such as diagnosis, treatment, and medicine) to detect unusual patterns. An unusual sequence might be billing for an unnecessary or non-existent service, or due to identity misuse.
- *Geo-spatial analysis.* To identify improbable drug fill or procedures.

In this paper, we focus on a smaller problem of outlier identification on a sample dataset – the collection of all pharmacy claims in 2012 in a government healthcare program in one US state. While the reported work is on pharmacy claims, it is worth noting that the methodology is rather general and can be extended to fraud detection in more general medical claims and a number of other applications. For instance, currently we are extending the same set of techniques to fraud detection in debit card usage data.

The contributions of this work can be summarized as follows:

- Increasing investigators' efficiency through an automatic search and computation process that incorporates large numbers of fraud rules and claim lines.
- Reducing the false alarm rate thereby reducing unncessary investigation cost.
- Providing explainations of the results and enalbing the investigators to gain new insights and define new fraud rules.
- The ability to easily incorporate new rules as fraudulent entities get creative at figuring out new ways to commit fraud.

2. PHARMACY CLAIMS PROCESSING

Every time a patient takes a prescription to fill at a pharmacy, the pharmacy contacts the patient's insurance program and files a claim. A pharmacy claim may include information such as the following:

- Pharmacy information (e.g., National Provider Index (NPI) number, address, phone number, and tax ID).
- Patient information: name and insurance policy number, gender, age, etc.
- Information regarding the prescribing doctor: name, address, NPI, etc.
- Drug information, such as a 11-digit drug identifier known as the National Drug Code (NDC). A drug may also be identified by its generic formulation (a 6-digit Generic Sequence Number) or its ingredient (a 5-digit Hierarchical Ingredient Code List (HICL) number). Sometimes there is also information regarding special drug properties. For instance, drugs that can be abused are often associated with a narcotics label to prompt special attention.
- Drug utilization information, e.g., drug quantity, and/or days of supply the drug is supposed to last.
- Payment information: drug price, drug dispense fee, total amount billed to the insurance company, the reimbursed amount, etc.
- Other auxiliary information such as the prescription filling date. Sometimes claims may also contain diagnosis information indicating which medical condition led to the utilization of the drug.

Insurance program administrators process pharmacy claims and reimburse the pharmacies within a timeframe that is required by law. At the same time, they need to watch out for fraud to ensure the program integrity. This is a tedious and difficult process. First, one needs to know what to look for (for example, pharmacies with too much narcotics sales) then sift through the claims data to identify suspicious pharmacies and claims. Investigators then examine records from the suspicious pharmacies and pull relevant information to see if a claim suspicion is justifiable. As investigation is a costly effort, therefore finding the right pharmacies and claims to focus on is essential. In this paper we focus on the screening stage, i.e., automating the process of identifying suspicious pharmacies, to increase the investigators' effectiveness. Investigation outcomes determine whether further action is needed. It may lead to educating/training the pharmasists or be escalated into a crime investigation if fraud is noticed.

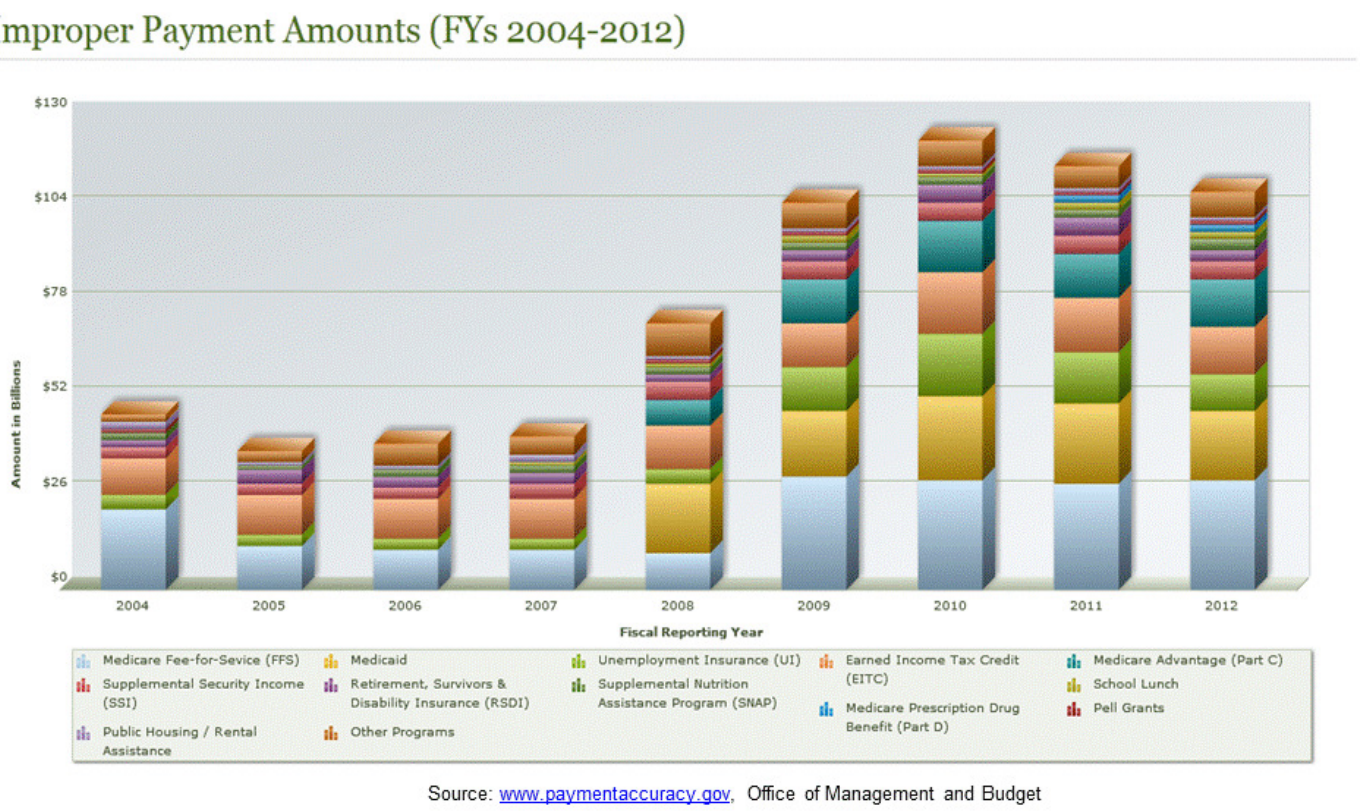


Figure 1: Improper payments in government expenditure, source www.paymentaccuracy.gov

number of claim lines	24,140,551
number of pharmacies	5,617
number of prescribing doctors	74,314
number of patients	2,514,854
number of drugs	23,275

Table 1: Raw data statistics of the pharmacy claim dataset.

For concreteness, we anchor our work on a dataset of all pharmacy claims in 2012, The raw statistics are summarized in Table 1. Sieving through millions of claims is a formidable task. The rest of the paper presents our work on fraud detection and preliminary results on this dataset.

Algorithm 1 Methodology

1: Define fraud rules R using domain knowledge
2: Compute violation values V_P^T for each pharmacy $p \in P$ and each violation type $t \in T$
3: Compute risk scores S_P for each pharmacy $p \in P$
4: Identify suspicious pharmacies X
5: Report a ranked list of suspicious pharmacies
6: Report suspicion indicators for each suspicious pharmacy

3. FRAMEWORK

Our fraud detection framework comprises five components: (1) fraud rule generation, (2) feature construction, (3) risk score computation, (4) outlier identification, and (5) reporting and visualization. We follow the methodology outlined in Algorithm 1.

We begin by working with the analysts to define rules of suspicious activities to govern our analysis. We explain how we define some rules in section 3.1. Once the rules are defined, we use them to extract and compute relevant features from the claims dataset. We discuss our feature construction process in section 3.2. Next, we compute a risk score for each pharmacy following algorithm 2 outlined in section 3.3. To identify suspicious pharmacies, we have developed an outlier identification technique presented in section 3.4. Finally, we discuss our reporting and visualization technique that we use to present our findings and recommendations to the investigators in section 3.5.

3.1 Suspicion indicators for pharmacy frauds

Working with domain experts in fraud screening and investigation, we have designed a set of pharmacy claim fraud rules (or suspicion indicators). We cannot disclose the full set of rules in details as they contain proprietary information. Here we describe a few sample indicators (Table 2) at a high-level.

Category	Features to look for
Billing Error	duplicate billing
Narcotics	class II narcotics use
Refill	automatic refill or frequent refill
Brand name drugs	unusually high share of high profit drugs
Dispense Control	signs of missing control steps
Excessive dispense	excessive qty
	excessive billed amount

Table 2: Suspicion indicators for pharmacy claims

- Billing Errors. Duplicate billing can be an unintentional mistake or a purposeful action. Regardless of the intention, duplicate billing is worth noticing, especially if duplicate bills are frequent and associated with excessive claim amounts.
- Narcotics-related indicators. Repeated use and significant sales on narcotics, especially Class II narcotics, are suspicious.
- Refill. Drug refill interval should be comparable to the day supply number on the original prescription, with a bit of random perturbation as people may visit the pharmacy several days earlier or later. An unusually high share of early refills can be suspicious. In addition, precisely regular refills are often automatic refills. Whether patients picked up the medication is not known and needs to be checked.
- Brand name drugs. A pharmacy with an usually high percentage of brand-name drugs may be dispensing generic drugs but billing for brand name drugs because the latter have a higher profit margin. The same applies to other high profit margin drugs.
- Dispense control. Drugs often come with restrictions on dispense control. Extra procedures need to be performed to check eligibility and applicability before the drug can be dispensed. We look for signs that a drug is dispensed without the required procedure.
- Excessive amount and/or quantity. Each drug dispense is compared to the norm of the same drug, and excessive amount, day supply, and quantity is flagged.

3.2 Feature construction

We process the claims data to extract the information relevant to each fraud rule. For each rule r and pharmacy p, we compute two rule-violation types: (1) the number of claims in violations of rule r by p, and (2) the total reimbursed dollar amount associated violations of r by p. To do this, we aggregate the information from all the claims for each pharmacy and normalize them generate a set $V_P^T\ \forall_{p\in P}\forall_{t\in T}$. The normalization is done based on the total number of claims submitted by each pharmacy.

3.3 Computing risk scores

Note that the feature value set V_P^T is quite detailed. It contains 2 violation types per fraud rule per pharmacy. From the perspective of fraud investigators, they would like a scalar risk score per pharmacy so as to focus on the pharmacies with the highest risk scores. How to combine the feature values from multiple violation types is a challenge that we need to address.

Algorithm 2 Compute Scores(V_P^T)

1: $W = \emptyset$
2: **for all** $t \in T$ **do**
3: $\quad m^t = \#$pharmacies with type t violation
4: $\quad w^t = \log \frac{|P|}{m^t}$
5: $\quad W = W \cup \{w^t\}$
6: **end for**
7: $S = \emptyset$
8: **for all** $p \in P$ **do**
9: $\quad F_p = \emptyset$
10: $\quad$ **for all** $t \in T$ **do**
11: $\quad\quad f_p^t = v_p^t * w^t$
12: $\quad\quad F_p = F_p \cup \{f_p^t\}$
13: $\quad$ **end for**
14: $\quad s_p = \sum_{t=1}^{|T|} f_p^t$
15: $\quad S = S \cup \{s_p\}$
16: **end for**

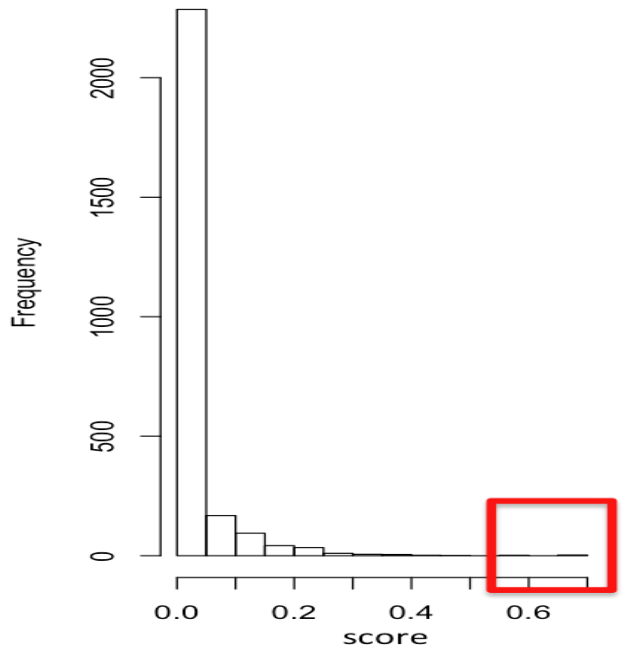
For this task, we adopt from content analysis literature a method known as TF–IDF (Term Frequency – Inverse Document Frequency) [1]. In content analysis, document content is often summarized by term frequency, i.e., how frequently a word occurs in a document. On the other hand, common words such as "a", "the", and "and" do not carry much meaning despite of their frequent use. Relying on term frequency alone for content analysis is problematic due to the bias introduced by these common words. To amend this problem, content analysis researchers designed inverse document frequency to measure the semantic importance of words.

For any given word a, the inverse document frequency is measured as the log ratio of (1) the number of all documents in a corpse, over (2) the number of documents containing the word a. In essence this is measuring how unique the word a is – if a is a common word such as "a" and "the", the log ratio is 0, meaning the word is semantically unimportant. On the other hand, if a is a word with a very specific meaning, it will only show up in a small number of documents, and hence the log ratio will be high. This log ratio is then used as weight to the term frequency. Through the inverse document frequency, the semantic importance is measured.

In our case, we want to combine the risk scores for a particular pharmacy, while differentiating between common violations and rare violations. It is analogous to the differentiation between common and rare words in content analysis. Similar to TF-IDF, the semantics of a rule violation is defined as the log ratio of the total number of pharmacies $|P|$ over the number of pharmacies with the rule violation.

Lines 1–6 in Algorithm 2 outlines the computation of the log ratio weight $w_t \forall_t$. Line 4 is the weight computation. Given the computed weights, the risk scores are combined through a weighted summation (lines 8–16). This weighting scheme ensures that rule violations are combined based on their importance. To the best of our knowledge, no previous work has used this idea for fraud detection.

Figure 2: Outlier identification



3.4 Identifying suspicious pharmacies

Figure 2 illustrates a histogram indicating a distribution of risk scores for all the pharmacies for a particular violation type. The histogram bars indicate a score frequency for various discrete score intervals. The horizontal axis indicates various discrete score intervals for all the pharmacies, and the vertical axis indicates a score frequency. For example, approximately 2300 pharmacies have a score within the interval [0, 0.05], and approximately 200 pharmacies have a score within the interval [0.05, 0.1]. Specifically, the histogram illustrates a decay (e.g., exponential) in the score frequency within the score interval [0, 0.55], such that 0 entities have a score within the interval [0.45, 0.55].

Our algorithm uses this histogram to identify pharmacies whose scores do not fit within the trend of the histogram. For example, the score interval [0, 0.55] follows a normal decay pattern and is associated with a set of "normal" pharmacies that may not be engaged in fraudulent transactions. However, the two bars shown in the red square indicate that a small number of pharmacies have an anomalous score within the intervals [0.55, 0.6] and [0.65, 0.7], respectively, which does not fit within the normal decay pattern of the histogram. Our algorithm can detect these pharmacies within the interval [0.55, 0.7] as "anomalous" or "outliers", which allows the organization to investigate these pharmacies further to determine whether they are committing fraudulent transactions intentionally.

We use an entropy based method ([5], [4]) to set the threshold for outlier identification. For a set S of scores, we compute the entropy $E = \sum_{i \in S} P(v_i) * (\log P(v_i))$ and for each element i, the surprise ratio $s_i = \frac{-\log p(v_i)}{E}$. We choose the threshold as the point i that has a surprise ratio of 2 (i.e., where the surprise is more than twice the average).

Ph	t1	t2	t3	t4	Comb	MinRank	#Indicators
1	**4**	**7**	**1**	**9**	18	1	4
2	**8**	47	**6**	**1**	17	1	3
3	**1**	**6**	**3**	28	50	1	3
4	15	**9**	**8**	**2**	20	2	3
5	11	**2**	**7**	99	11	2	2
6	**2**	12	**4**	77	72	2	2
7	27	**3**	31	16	38	3	1

Table 3: Reporting suspicious pharmacies

3.5 Reporting

We developed a reporting strategy in which we communicate our findings to the investigators in a simple yet effective way. Our goal is to provide the investigators with a list of suspicious pharmacies to investigate, and involve them in the evaluation. Although it is partial and empirical, our evaluation is feasible and has confirmed real fraud cases. For each violation type t, we apply the outlier identification technique explained in section 3.4 to the list of scores and identify the set X^t of suspicious pharmacies according to each violation type t. The final list of suspicious pharmacies contains the union of all these sets. All the pharmacies are ranked according to their risk scores for each violation type t and the combined score independently. Table 3 shows a subset of the output.

For each suspicious pharmacy, the rank according to each t is shown. For example, the rank of pharmacy 1 is equal to 1 for violation type $t3$ means that it is the most suspicious pharmacy with respect to this violation type. For each pharmacy, we also report the minimum rank for each pharmacy and the number of indicators (#violation types the pharmacy is ranked in the top 10, shown in bold text format). For example, the number of indicators for pharmacy 1 equals 4, means that this pharmacy is ranked in the top 10 most suspicious according to 4 violation types. Finally, we report the dollar amount associated with each type of violation and the total recovered amount shall the pharmacy be investigated and identified as fraudulent.

4. EVALUATION

Ideally we would also like to rigorously evaluate the performance of our anomaly detection system against an "oracle" ground truth and derive detailed metrics such as the false positive rate and miss detection rate, much like what a Receiver Operator Characteristic (ROC) curve is commonly used for in the detection/classification literature. However these metrics are not feasible in this context. Ground truth is very expensive to obtain, as it involves thorough investigation effort. For instance, a routine pharmacy audit normally cost several hundred dollars, and an in-depth investigation may cost significantly more. It is economically infeasible to investigate a claim dataset of substantial size. As a result, complete ground truth never exists.

Given this lack of ground truth to benchmark against, we take a pragmatic human investigator in the loop approach – for anomalies identified by our scheme, we involve effort from human investigators (domain experts) to validate or dis-validate. Note that this is not a precise performance evaluation approach but rather a tradeoff between rigor and

feasibility. On the other hand, involving human investigators in the loop enables active learning. Confirmation or dis-validation of detected anomalies both provide useful information, from which our system can learn to further improve.

The goal of this work is to identify a very small set of suspicious pharmacies. Due to the high cost of investigation, this set must be as small as possible and the confidence level of suspicion must be as high as possible. By this we can reduce the false positive rate, reduce cost of detecting fraud, and increase the amount of money that can be recovered. Therefore, while our ranking approach (Section 3.5) allows the investigators to compare a pharmacy to the entire population with respect to each suspicion indicator, our outlier identification method (Section 3.4) is necessary to narrow down the list of suspicious pharmacies to investigate.

We applied our technique to a set of 2,563 pharmacies and identified 7 suspicious pharmacies. The auditors investigated 5 of these pharmacies and they were all found to be fraudulent.

5. CONCLUSIONS

In this paper we present our fraud screening scheme to detect suspicious pharmacies from a large dataset of pharmacy claims. Our method is based on a probabilistic outlier identification technique and combines features from a set of fraud rules based on their semantic importance. This tool is designed to suit the needs of fraud investigators and has proven usefulness by identifying real fraud cases. We plan to continue our collaboration with fraud investigators to develop new fraud detection capabilities. In addition, we will build an active learning framework to engage investigators and learn from investigation results. It is challenging to design an active learning system to maximize learning benefit while avoiding over-burdening the investigators. We hope to address this problem in our future research.

6. REFERENCES

[1] A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 1972.

[2] Big data: the next frontier for innovation, competition, and productivity. *McKinsey Global Institute Report*, 2012.

[3] The price of excess: identifying waste in healthcare spending. *Price Waterhouse Coopers (PWC) Health Research Institute Report*, 2012.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. New York, NY: John Wiley and Sons, Inc., 1991.

[5] P. E. Hart, D. G. Stock, and R. O. Duda. *Pattern Classification, 2nd Edition*. John Wiley and Sons, Inc, 2000.