Using Big Data as a Fraud Detection Tool in Medical Insurance Claims

By

Adnan Abu Alhaija

Abedelrahman Elsheikh Ali

February 2019

Presented to the

Jordan Insurance Federation

In Partial Fulfillment of

Aqaba Conference Award

Abstract

During one year, there are millions of medical insurance claims submitted to the healthcare companies, which helps make quality care more affordable. A small percentage of these claims are deceitfulness, they costs companies and the government additional amounts, leading to higher premiums and other out-of-pocket expenses.

The medical insurance fraud make up the majority of all the types of insurance frauds records, revealing why this is such an important problem.

Big data is created every day by the interactions of billions of people using computers, GPS devices, cell phones, sensors, medical devices, and data-intensive areas such as atmospheric science, genome research and astronomical studies. Today big data opened a huge opportunities to people who can use it effectively. Now realizing the great importance of big data, many researches are engaged in finding hidden information in big data.

There are available opportunities to detect insurance fraud in insurance business using data-driven tools to help assess insurance claims for fraud, for example Industry-wide databases that link accurate and comprehensive historical information to improve the assessment of any claim. Patterns of repeated claims for an item or injury can help ensure you identify suspicious claims and potential fraudsters, data visualization: Synthesizing the available information helps you make better decisions. Pictures and infographics are the often best way to simplify large volumes of complex information. This is important because the people most able to make immediate sense of the insight are those on the front line of operations and Investigation application and software used to turn the available data into checklists and provide one-touch reporting and will streamline the claims process. This reduces the scope for human error and reduces the need for inefficient manual reports.

In our research we illustrated some scenarios where fraudulent claims can happen, for example doctors, who treated whopping, say 50+ patients in a day, distance between claimant's home address and medical provider and providers prescribing certain drugs at higher rate than others, High number of treatments for type of injury.

Abnormally long treatment time off for the type of injury, Providers administering (more) tests and treatments or providing equipment that are not medically necessary and

Many other cases where the mining in these claims data as well as having different data sources will help insurer to detect such fraud.

Table of Contents

Introduction

The focus of this thesis is to present the issue of medical insurance fraud, and how insurance companies might use big data technology and tools to predict and prevent insurance fraud. While the methods presented herein will be applied specifically to a medical insurance dataset, their applicability to other industries, such as motor, credit cards and customer retention.

Since fraud is on, the increase holistic fraud prevention is required. According to the well know market research organization Gartner: "Security and fraud risk exposure is increasing as organizations are threatened at multiple points of vulnerability. Companies are reevaluating how they tackle security since a fragmented approach is consistently leaving organizations at greater risk of attack. A more holistic approach to security ensures all layers of protection function together". (1)

The first chapter of this thesis will provide background information on the matter of medical insurance. We will provide general overview about medical insurance and some figures related to Jordan market.

In the second chapter we will provide overview about fraud in general, Why Does Fraud Happen and who is responsible for the prevention and detection of fraud?

In the third chapter we will establish the relevant definitions and present the Impact of Insurance fraud supported with statistics from different markets.

In the fourth chapter we will identify the research problem illustrating some related work.

In the fifth chapter we will deep dive into the concept of big data and its applications with brief about Hadoop, Python and R Programming Environment, and big data as a fraud detection tool.

Finally in the sixth chapter we will present our model that we believe it might be applicable on our companies to fight the medical insurance fraud supported with some high level algorithms.

1. Chapter One Medical Insurance

The term ‘Medical Insurance’ relates to a type of insurance that essentially covers your medical expenses. A health insurance policy like other policies is a contract between an insurer and an individual / group in which the insurer agrees to provide specified health insurance cover at a particular “premium” subject to terms and conditions specified in the policy.

According to the Health Insurance Association of America, health insurance is defined as "coverage that provides for the payments of benefits as a result of sickness or injury. It includes insurance for losses from accident, medical expense, disability, or accidental death and dismemberment".

As at 2017, there are 23 private insurance companies providing medical insurance, In Jordan with total premium of 168.9 million JOD, 92% of the contacts were group while 8% for individual contracts. The total number of the insured persons under private companies’ policies and third party administrators are 830,000 persons as per Insurance Administration Department. In 2017, the sector paid around 159.6 Million JOD as medical claims. [(2)]

2. Chapter Two: Fraud

2.1 Fraud

Fraud encompasses a wide range of illicit practices and illegal acts involving intentional deception or misrepresentation. The Institute of Internal Auditors’ International Professional Practices Framework (IPPF) defines fraud as “any illegal act characterized by deceit, concealment, or violation of trust. These acts are not dependent upon the threat of violence or physical force. Frauds are perpetrated by parties and organizations to obtain money, property, or services; to avoid payment or loss of services; or to secure personal or business advantage.” [(3)]

Fraud affects organizations in several areas including financial, operational, and psychological. While the monetary loss owing to fraud is significant, the full impact of fraud on an organization can be staggering. The losses to reputation, goodwill, and customer relations can be devastating. As fraud can be perpetrated by any employee within an organization or by those from the outside, it is important to have an effective fraud management program in place to safeguard your organization’s assets and reputation.

2.2 Why Does Fraud Happen? (4)

Interviews with persons who committed fraud have shown that most people do not originally set out to commit fraud. Often they simply took advantage of an opportunity; many times the first fraudulent act was an accident – perhaps they mistakenly processed the same invoice twice. But when they realized that it wasn't noticed, the fraudulent acts became deliberate and more frequent. Fraud investigators talk about the 10 - 80 - 10 law which states that 10% of people will never commit fraud; 80% of people will commit fraud under the right circumstances; and 10% actively seek out opportunities for fraud. So we need to be vigilant for the 10% who are out to get us and we should try to protect the 80% from making a mistake that could ruin their lives.

Generally, fraud occurs because of a combination of opportunity, pressure and rationalization. An opportunity arises, the person feels that the act is not entirely wrong, and has pressure pushing them to commit the fraud.

Opportunity, An opportunity is likely to occur when there are weaknesses in the internal control framework or when a person abuses a position of trust. For example:

- Organizational expediency – 'it was a high profile rush project and we had to cut corners';
- Downsizing meant that there were fewer people and separation of duties no longer existed; or
- Business re-engineering brought in new application systems that changed the control framework, removing some of the key checks and balances.

Pressure, the pressures are usually financial in nature, but this is not always true. For example, unrealistic corporate targets can encourage a salesperson or production manager to commit fraud. The desire for revenge – to get back at the organization for some perceived wrong; or poor self-esteem - the need to be seen as the top salesman, at any cost; are also examples of non-financial pressures that can lead to fraud.

Rationalization, in the criminal's mind rationalization usually includes the belief that the activity is not criminal. The often feel that everyone else is doing it; or that no one will get hurt; or it's just a temporary loan, I'll pay it back, and so on.

Interestingly, studies have shown that the removal of the pressure is not sufficient to stop an ongoing fraud. Also, the first act of fraud requires more rationalization than the second act, and so on. But, as it becomes easier to justify, the acts occur more

often and the amounts involved increase in value. This means that, left alone, fraud will continue and the losses will only increase. I have heard it said that 'There is no such thing as a fraud that has reached maturity'. Fraud, ultimately, is fed by greed, and greed is never satisfied

2.3 Who is responsible for the prevention and detection of fraud?

There are two main views - one states that management has the responsibility for the prevention and for the detection of fraud. Management:

- is responsible for the day to day business operations;
- is responsible for developing and implementing controls;
- has authority over the people, systems, and records; and
- has the knowledge, and authority to make changes

Therefore, fraud prevention and detection is their problem. Audit, on the other hand:

- has expertise in the evaluation and design of controls;
- reviews and evaluates operations and controls; and
- has a requirement to exercise 'Due Diligence'

3. Chapter Three: Insurance Fraud

3.1 Insurance Fraud [5]

Insurance fraud has probably existed ever since the inception of the insurance industry itself. Insurance fraud affects not only the financial health of the insurers, but also of innocent people seeking effective insurance coverage. Fraudulent claims are a serious financial burden on insurers and result in higher overall insurance costs. The types of insurance fraud are widespread and diverse with many schemes targeting specific sectors in the industry. Vigilance is critical.

Fraud can occur at any stage of an insurance transaction by any of the following:

- Individuals applying for insurance
- Policyholders
- Third-party claimants
- Professionals who provide services to claimants

3.2 Impact of Insurance Fraud

The cost of insurance to consumers continues to grow each year due to the huge losses that occur within the insurance industry. Most carriers can only estimate what they lose to claims that are discovered to be fraudulent. Insurance fraud usually does

not become known until the fraudsters get greedy and it becomes apparent that they are involved in an insurance fraud scheme.

Insurance fraud is not limited to one group, race, or gender. An equal opportunity crime can be performed by an insurer or an insured. Insurance fraudsters are less likely to be turned over to the authorities because they are usually not equipped to handle or perform an adequate investigation due to lack of expertise in insurance. For insurance fraud to be proactively addressed, insurers must train their staff in identifying the red flags of insurance fraud schemes.

3.3 By the numbers: fraud statistics [6]

Measuring insurance fraud is an elusive target. No single national agency gathers omnibus fraud statistics. Insurance fraud data thus are relatively piecemeal, making our understanding of insurance fraud an ongoing work in progress.

Insurance companies, associations and diverse state and federal agencies each gather fraud data related to their own missions. But the kind, quality and volume of data they compile vary widely.

Overall – United States

- Conservatively, fraud steals $80 billion a year across all lines of insurance. (Coalition against Insurance Fraud estimate).
- Fraud comprises about 10 percent of property-casualty insurance losses and loss adjustment expenses each year; and
- Property-casualty fraud thus equals about $34 billion each year. (Insurance Information Institute, September 2017)
- Fraud costs for insurers
- Fraud accounts for 5-10 percent of claims costs for U.S. and Canadian insurers. Nearly one-third of insurers (32 percent) say fraud was as high as 20 percent of claims costs.
- 57% of insurers predict an increase in personal-property fraud by policyholders. Around 58 percent say the same for personal auto insurance, and 69 percent expect a rise in workers-compensation scams;
- 61 percent predict an increase in auto-insurance fraud by organized rings, and 55 percent predict an increase workers-compensation scamming;
- About 35 percent say fraud costs their companies 5-10 percent of claim volume. More than 30 percent say fraud losses cost 10-20 percent of claim volume;

- Detecting fraud before claims are paid, and upgrading analytics, were mentioned most often as the insurers' main fraud-fighting priorities;
- One-third of insurers do not feel adequately protected against fraud. (FICO, August 2013)

3.4 Public attitudes [7]

A relatively large number of consumers remain at risk of committing this crime. They believe it's acceptable to increase insurance claims to make up for deductibles. Even so, those numbers have declined in recent years.

- 24 percent say it's acceptable to pad an insurance claim to make up for the deductible — 33 percent said it's acceptable in 2002;
- 18 percent believe it's acceptable to pad a claim to make up for premiums paid in the past;
- Younger males were much more likely to condone claim padding. And 23 percent of 18-34 year-old males say it's alright to increase claims to make up for earlier premiums. This compares with 5 percent of older males and 8 percent of females of the same age;
- 86 percent of Americans think "insurance fraud leads to higher rates for everyone;" and
- 10 percent of people think that insurance fraud does not hurt anyone. (Insurance Research Council, March 2013)
- More than half (55 percent) of U.S. consumers say poor service from an insurance company is more likely to cause a person to defraud that insurer;
- More than three-quarters (76 percent) say they're more likely commit insurance fraud during an economic downturn than during normal times (up from 66 percent in 2003);
- More than two-thirds of consumers (68 percent) say they believe insurance fraud happens because people believe they can get away with it (up from 49 percent in 2003);
- Some 72 percent of consumers believe insurance companies can identify fraud (down from 83 percent in 2003). (Accenture Ltd., September 2010)

4. Chapter Four: Research Problem

4.1 Research Problem

In the medical insurance field, we have many challenges, one of them is FRAUD, and it is happen all time with a bad effect and impact to the insurance companies worldwide, in the research we studied this issue and present the most effective and useful solution using big data.

4.2 Related works

- Big Data fraud detection using multiple Medicare data sources [8]

The research conducted by Matthew Herland from Florida Atlantic University

In their study, they use three Public Use File (PUF) datasets: (1) Medicare Provider Utilization and Payment Data: Physician and Other Supplier (Part B), (2) Medicare Provider Utilization and Payment Data: Part D Prescriber (Part D), and (3) Medicare Provider Utilization and Payment Data: Referring Durable Medical Equipment, Prosthetics, Orthotics. And Supplies (DMEPOS). They chose these parts of Medicare because they cover a wide range of possible provider claims, the information is presented in similar formats, and they are publicly available. Furthermore, the Part B, Part D, and DMEPOS dataset comprise key components of the Medicare program and by incorporating all three aspects of Medicare for fraud detection, their study provides a comprehensive view of fraud in the Medicare program. Information provided in these datasets includes the average amount paid for these services and other data points related to procedures performed, drugs administered, or supplies issued. They also create a dataset combining all three of these Medicare datasets, which we refer to as the combined dataset. The last dataset examined in their study is the List of Excluded Individuals and Entities (LEIE), provided by Office of the Inspector General, which contains real-world fraudulent physicians and entities.

5. Chapter Five

5.1 What is Big Data?

The expression "big data" refers to sets of information that are so large and so complex that they cannot be interpreted, or even measured, by traditional means. For example, in healthcare, information traditionally was gathered by analyzing insurance claims data, or by mining relatively small databases from healthcare systems. It was generally acknowledged that these methods could not deliver an accurate picture of an individual patient's health, nor could it tell us the best way to deliver services to that individual. The old system gave us the best it could, given its limitations. Data were extracted from spreadsheets that could, at best, relate only the most basic pieces of information, such as age, gender and diagnosis.

Big data, by contrast, analyzes many more data points, allowing more appropriate treatment options for patients.

Big data in healthcare refers to the reams of data that are now being generated from wearable devices and mobile apps, as well as from portable monitoring devices and digital health advisors. Devices such as blood pressure monitors and continuous blood sugar monitors are designed to be worn at home and easily used by the patient. Together, these devices and the information they provide has been called the "Medical Internet of Things". These devices promise to fill in the blank spaces in the healthcare picture. [(9)]

Big data is a term that describes the large volume of data – both structured and unstructured – that drowns a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

5.2 Four V's of big data [(10)]

Every side of business today is data-led, shaped and assessed by a level of data collection and analysis that would have seemed immeasurable just a handful of years ago. While the most sophisticated computer systems in operation still struggle to approximate human thought, even everyday smartphones far outstrip our ability when it comes to analytics. And when you get to the highest level of data analysis — the titular big data being the industry that surrounds it — you run into the thorny problem of trying to place it all into context so that a human might be able to

comprehend its significance. That's where the 4 Vs of big data enter the equation: they're the high-level dimensions that data scientists use to break everything down.

Variety:

Ever-escalating levels of cross-platform and cross-channel integration ensure that more data is available on any given day than on the day before. Consequently, data scientists aren't limited to collecting data from just one source: they can collect it from numerous sources. Think about the potential of social platforms: drawing data not just from Facebook, but also from Twitter, Snapchat, Instagram, LinkedIn, Pinterest, YouTube, Twitch, Tumblr, and various others.

For big data, variety concerns the breadth of the types of data collected, going all the way from studies and sources that factor in just one data type (an Instagram post, for instance) to those that take many into account (tweets, Facebook updates, Pinterest pins, etc.). It's an important dimension because it affects the significance of the inferences made from the data.

Velocity:

Differing from regular old-fashioned data studies, today's data science doesn't seek to gather data over time then carry out a singular analysis. Its analysis is live and ever-changing, driven by constant streams of data. Velocity concerns the rate at which this data being generated, distributed, and collected. The more sensors are present on Internet of Things - enabled devices, and the more people are using the internet, the higher the velocity of data analysis will be.

This dimension is so significant because the faster data can be acquired and processed, the more valuable it will be to begin with, and the longer it will retain its value — but the system you use to analyze it must be up to the task or be left behind.

Veracity:

How much can you trust the quality and accuracy of the data you're relying on to drive valuable conclusions? It depends on various factors, including where the data comes from, how it's collected, and how it's analyzed. The veracity of your data concerns how reliable and significant it really is, and you need high-quality data. When analyzing Twitter data, for instance, the data should be extracted directly from the site (though the API or not), not through a third-party system for collecting tweets, because you can't trust the latter.

Then there's the data that's collected accurately but doesn't necessarily mean anything, such as data from poorly-designed surveys. Everyday analytics can easily stuck on vanity or arbitrary metrics that do not hold any significance and big data is just as susceptible: while it's hard for a computer to draw inaccurate conclusions, it's easy for a person to fail to define the data range strictly enough, or to have mistaken assumptions about the quality of their data.

Volume:

Very simply, volume is how much data is being generated and collected all the time. It isn't just the pace that has increased astoundingly, but also how much data there is. There are more than 2.2 billion active users on Facebook, many of them spending hours each day writing updates, liking posts, commenting on images, playing games, clicking on ads, and doing numerous other things that can be analyzed. And that's just one social media site.

Imagine the level of analysis that goes into perfecting something like Black Friday marketing — how much data must be sourced from ecommerce sites, social media conversation, forum posts, identified trends, surveys, and (of course) standard retailers, all to figure out the perfect price points for flat screen TVs across one long weekend. Now think about the kind of volume high-end enterprises and governments must use for devising predictive models. We are looking at absurd levels of data analysis, only made possible through supremely powerful computers.

5.3 Big Data Tools

In the industry there are several big data tools and projects, below some of those:

- Apache Hadoop

The long-standing champion in the field of Big Data processing, well known for its capabilities for huge-scale data processing. This open source Big Data framework can run on-prem or in the cloud and has quite low hardware requirements. The main Hadoop benefits and features are as follows:

- HDFS—Hadoop Distributed File System, oriented at working with huge-scale bandwidth
- Map Reduce—a highly configurable model for Big Data processing
- YARN—a resource scheduler for Hadoop resource management
- Hadoop Libraries—the needed glue for enabling third party modules to work with Hadoop

Most used tools in the field:

- Python

Python is an excellent tool and a perfect fit as python big data combination for data analysis for the below reasons:

- Open source
- Library Support
- Speed
- Scope
- Data Processing Support

- R Programming Environment

R is mostly used along with JuPyteR stack (Julia, Python, R) for enabling wide-scale statistical analysis and data visualization. JupyteR Notebook is one of 4 most popular Big Data visualization tools, as it allows composing literally any analytical model from more than 9,000 CRAN (Comprehensive R Archive Network) algorithms and modules, running it in a convenient environment, adjusting it on the go and inspecting the analysis results at once. The main benefits of using R are as follows:

- R can run inside the SQL server
- R runs on both Windows and Linux servers
- R supports Apache Hadoop and Spark
- R is highly portable
- R easily scales from a single test machine to vast Hadoop data lakes

5.4 Big Data as Insurance Fraud Detection

One benefit of the big data analytics can be fraud prevention. By many estimates, at least 10 percent of insurance company payments are for fraudulent claims, and the global sum of these fraudulent payments amounts to billions or possibly trillions of dollars. While insurance fraud is not a new problem, the severity of the problem is increasing and perpetrators of insurance fraud are becoming increasingly sophisticated.

What is the role for big data analytics in helping insurance companies find ways to detect fraud? Insurance companies want to stop fraud early. By developing predictive models based on both historical and real-time data on wages, medical

claims, attorney costs, demographics, weather data, call center notes, and voice recordings, companies are in a better position to identify suspected fraudulent claims in the early stages.

For example, a personal injury claim could potentially include fake medical claims or a staged accident. Companies have seen an increase in sophisticated crime rings to perpetrate auto insurance or medical fraud. These rings may have similar methods of operation that are enacted in different regions of the country or using different aliases for the claimants.

Big data analysis can quickly look for patterns in historical claims and identify similarities or bring up questions in a new claim before the process gets too far along.

Risk and fraud experts at insurance companies, along with actuarial and underwriting executives and insurance business managers, all see big data analytics as having the potential to deliver a huge benefit by helping to anticipate and decrease attempted fraud. The goal is to identify fraudulent claims at the first notice of loss — at the first point where you need an underwriter or actuary

Consider the following example. An insurance company wants to improve its ability to make real-time decisions when deciding how to process a new claim. The company's cost outlay including litigation payments related to fraudulent claims has been rising steadily. The company has extensive policies to help underwriters evaluate the legitimacy of claims, but the underwriters often did not have the data at the right time to make an informed decision.

The company implemented a big data analytics platform to provide the integration and analysis of data from multiple sources. The platform incorporates extensive use of social media data and streaming data to help provide a real-time view. Call center agents are able to have a much deeper insight into possible patterns of behavior and relationships between other claimants and service providers when a call first comes in.

For example, an agent may receive an alert about a new claim that indicates the claimant was a previous witness on a similar claim six months ago. After uncovering other unusual patterns of behavior and presenting this information to the claimant, the claim process may be halted before it really gets going.

In other situations, social media data may indicate that conditions described in a claim did not take place on the day in question. For example, a claimant indicated

that his car was totaled in a flood, but documentation from social media showed that the car had actually been in another city on the day the flood occurred.

Insurance fraud is such a huge cost for companies that executives are moving quickly to incorporate big data analytics and other advanced technology to address the problem of insurance fraud. Insurance companies not only feel the impact of these high costs, but the costs also have a negative impact on customers who are charged higher rates to account for the losses.

By using big data analytics to look for patterns of fraudulent behavior in enormous amounts of unstructured and structured claims-related data, companies are detecting fraud in real time. The return on investment for these companies can be huge. They are able to analyze complex information and accident scenarios in minutes as compared to days or months before implementing a big data platform.

5.5 Predicting and Preventing Insurance Fraud with Big Data

There are several "Big Data" options are now available for companies to achieve higher success rates in terms of predicting and even preventing insurance fraud.

5.5.1 Predictive models:

Cause-and-effect can be difficult to demonstrate when you assess risk. However, certain behaviors, attributes, and scenarios are known to have strong associations with insurance fraud.

Predictive models can integrate the known risks to produce a simple value or ranking of what to watch out for, giving managers a better idea of what to expect, thereby improving possible prevention protocols.

One step further however is the use of digital tools such as semantic processing and artificial intelligence which move the needle from probabilistic assessment to deterministic assessment. That is to say that if (A) and (B) occur, then (C) -the fraud- definitely will, based on multidimensional data insights.

5.5.2 Neural networks:

Neural networks are systems of hardware and software modelled after the human brain. These systems can learn and cope with novelty and innovation. With proper management, customization, and implementation, neural network technology can detect new and systematic insurance scams.

With this technology, you can mitigate the risk of damage from newer fraud schemes before they result in catastrophic losses. This is the case with semantic processing and artificial intelligence.

5.5.3 Data mining:

Mountains of information are buried in documents, recordings, case notes, surveys, and other sources online and offline. Accessing specific information manually that might indicate fraud is a challenge and can be inefficient and time-consuming.

Applications and automated processes which extract and synthesize relevant text, voice, or financial information can make this task much easier and in less time, putting less burden on managers and analysts so they can focus on pinpointing possible insurance fraud incidents and on the preparatory and preventive measures against them.

5.6 Fighting Insurance Fraud with Big Data

There are immediate opportunities to detect insurance fraud in insurance business, below three data-driven tools to help assess insurance claims for fraud:

1- Industry-wide databases: Linking to accurate and comprehensive historical information will improve the assessment of any claim. Patterns of repeated claims for an item or injury can help ensure you identify suspicious claims and potential fraudsters.

2- Data visualization: Synthesizing the available information helps you make better decisions. Pictures and infographics are the often-best way to simplify large volumes of complex information. This is important because the people most able to make immediate sense of the insight are those on the front line of operations.

3- Investigation apps and software: Turn the available data into checklists and provide one-touch reporting and you will streamline the claims process. This reduces the scope for human error and reduces the need for inefficient manual reports.

5.7 Traditional Methods of Medical Insurance Fraud Detection:

Up until now, efforts to detect healthcare fraud and abuse have involved laborious, feet-on-the-ground investigative work – work that occurs after payments for false claims have been made. And it can take years to gather enough evidence to make arrests and prosecute.

Even if those efforts are successful, there is the issue of recovering the money, and this, too, can take years. It involves legal receiverships and fees, liquidation of the perpetrators' assets, and with only a percentage recovered in the end.

It is very much like the old adage of "closing the barn door after the horse has escaped," rather than taking steps to prevent the horse from escaping in the first place.

5.8 The Solution – Medical Insurance Fraud Prevention with Big Data and Analytics:

Clearly, the traditional healthcare fraud detection methods are not working. The more effective way to prevent fraud and abuse is to identify it before claims are paid. And that is why healthcare payers have now embraced the same predictive analytics that other sectors of the economy know to work.

Now, it makes sense to approach fraud detection in healthcare using data mining techniques too.

Predictive analytics identifies patterns that are potentially fraudulent and then develops sets of "rules" to "flag" certain claims. For example, a provider making a claim for a procedure that is outside of his/her area of expertise would be flagged for further scrutiny, because that is one of the "rules".

Fraud detection model can be an Artificial Intelligence application, which will continually mine data, identify more and more emerging fraudulent patterns and create new "rules" for those as well. The "intelligence" in the system learns from these new rules and continually becomes more sophisticated in identifying, even more, fraud potentials. And the best models not only flag the potentials but provide the reasons for that flagging, so that investigations and assessments by management can be completed efficiently.

In short, a solid healthcare fraud auditing and detection model will provide protection to the payer in the following ways:

- Identify inconsistencies and "rule-breaking" behaviors.
- Detect and prevent potentially improper payments, by flagging them for review.
- Continually mine data to identify new fraudulent patterns and develop new "rules" for those as well.

The beauty is in the big data that can all be mined and analyzed by one tool or model, rather than a host of separate healthcare fraud detection systems that do not function in coordination, or worse, do not even "know" to check other Internet data sources. One of the most common types of fraud, for example, is the continued claims for an individual who has died. An antiquated system will not have this information, but a system that is "plugged into" big data will. [11]

6. Chapter Six: Proposed Platform

6.1 Proposed Platform:

In our research we will mine in the database to identify patterns based in the below scenarios:

1. Doctors, who treated whopping, say 50+ patients in a day.
2. Distance between claimant's home address and medical provider
3. Providers prescribing certain drugs at higher rate than others do.
4. Multiple medical opinions/providers.
5. High number of treatments for type of injury.
6. Abnormally long treatment time off for the type of injury.
7. Providers administering (more) tests and treatments or providing equipment that are not medically necessary.
8. Providers conducting medically unrelated procedures and services.
9. Providers administering far higher rates of tests than others do.
10. Providers costing far more, per patient basis, than others.
11. Changing providers for the same treatment (possibly correlated with other claim activity)
12. Providers billing for services not provided.
13. Providers administering more expensive tests and equipment (up coding).
14. Providing multiple billing for services rendered.
15. Providers unbundling or billing separately for laboratory tests performed together to get higher reimbursements.
16. Policyholders traveling long distance for treatment, which may be available nearby. (Possibly scams by bogus providers.)
17. Policyholders letting others use their healthcare cards.

And make a scoring register for both health providers and patients based on their activities, then the register day after day will be growing and being helpful in decision making regarding the providers and the patients in respect of their claims.

The good news is those discovered pattern could be also applied at claims approval stage and will be helpful to reject or approve the claim at earlier stage.

6.2 How it works?

It is a kind of Algorithm running over the whole data to bring the results based on the previous mentioned criteria.

Example1:

Case: Doctors, who treated whopping, say 50+ patients in a day.

Algorithm: Select outpatient cases AND specific day GROUPBY doctor name with COUNT aggregation → keep rows where COUNT greater than or equal to 50 (you can change the trigger based on your country experience and you can make a step to consider the doctor's activities

Example2:

Case: Distance between claimant's home address and medical provider.

Algorithm: Select medical provider address, Select patient home and work address, transform the addresses into coordinate and get the distance between the two addresses → keep rows where the distance over 100KM (the trigger could be change based on you country experience)

Example3:

Case: Abnormally long treatment time off for the type of injury.

Algorithm: Select inpatient cases AND enter date, exit date → get the duration of treatments → start "FOR" loop for every type of injury → discover the abnormal durations.

7. References:

1- Rutrell, Y., “Analytics platform helps agencies fight cyber-crime, government computer news”, Jul 12, 2012, http://gcn.com/articles/2012/07/12/sassecurity-intelligence-platfromanalytics.aspx
2- https://mit.gov.jo/EchoBusV3.0/SystemAssets/PDFs/AR/Departements/insurance/Medical-insurance-Statistics-2017.pdf
3- https://www.acl.com/pdfs/DP_Fraud_detection_INSURANCE.pdf
4- https://chapters.theiia.org/ottawa/Documents/Fraud_Detection_and_Prevetion.pdf
5- https://www.acfe.com/uploadedFiles/ACFE_Website/Content/documents/Insurance-Fraud-Handbook.pdf
6- http://www.insurancefraud.org/statistics.htm#2
7- http://www.insurancefraud.org/downloads/InsuranceResearchCouncil03-13.pdf
8- https://www.researchgate.net/publication/327432720_Big_Data_fraud_detection_using_multiple_medicare_data_sources
9- https://www.carecentrix.com/blog/big-data-impact-healthcare
10- https://forwardleading.co.uk/blog/What-Are-the-4-Vs-of-Big-Data-How-to-Apply-Them
11- https://www.romexsoft.com/blog/healthcare-fraud-detection/