U.S. Department of Health & Human Services

CMS CENTERS for MEDICARE & MEDICAID SERVICES

Electronic Fraud Detection in the U.S. Medicaid Healthcare Program

MSc Thesis

Peter Travaille, 26th of January, 2011

University of Twente
Enschede - The Netherlands

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

UC San Diego

'Fraud control is a miserable business. Failure to detect fraud is bad news; finding fraud is bad news, too.'

- Prof. Dr. Malcolm K. Sparrow
(License to Steal, page viii)

Study	Industrial Engineering & Management
Track	Information Technology & Management
Faculty	School of Management & Governance
University	University of Twente, Enschede
	University of California, San Diego
Name	Peter Travaille
Email	ptravaille@sdsc.edu
Student #	s0073032
Examination Committee	Prof. Dr. Roland Müller (University of Twente)
	Prof. Dr. Jos van Hilligersberg (University of Twente)
	Dallas Thornton, MBA (University of California San Diego)

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Executive Summary

Background

Health care fraud in the United States is a severe problem that costs the government billions of dollars per year. The costs as a result of the fraud waste and abuse is estimated to be 1/3rd of the total health care costs in the United States; $700 billion dollar (Kelly, 2009). The health care programs Medicare and Medicaid have to deal with fraudulent practitioners, organized criminal organizations and honest providers who make unintended mistakes while billing for their legitimate services. The U.S. government has trouble fighting the fraud and abuse in the federal programs; concerning Medicaid an extra difficulty is that the states are governing the program. A consequence is diverged legislation and different eligibility criteria and that impedes national fraud detection and control.

Objective

What lessons can be learned from related industries to improve the fraud detection system in the Medicaid health care program? The support of information technology is essential and electronic fraud detection techniques and systems are developed in several industries. In the insurance, telecommunications and financial industry, particularly the credit card industry, fraud detection is a vital aspect to do business. In general, insurance fraud and abuse occur and the existing difference in information possession between the insurer and the beneficiary (asymmetric information) create opportunities to commit fraud. In the Medicaid program fraud detection is the responsibility of the state and federal government. Hence, it is also their responsibility to reduce these opportunities to commit fraud to an absolute minimum; applying electronic fraud detection techniques is one of the opportunities to reduce the information gap.

Medicaid Integrity Program

The Medicaid Integrity Program is a nationwide attempt of the U.S. federal government to fight fraud and abuse in the Medicaid program. The national Medicaid database is hosted at the San Diego Supercomputer Center at the University of California San Diego. A systematic literature review supported by an internship at the national Medicaid database forms the foundation of this research regarding the support of information technology in the Medicaid fraud detection.

Findings

In this research, the telecommunications, credit card, and computer intrusion industries are analyzed to provide an overview of the applied electronic fraud detection techniques. Rule based methods, summarizing statistics in the form of profiling, and unsupervised and supervised data mining are covering the spectrum of electronic fraud detection methods. From a data perspective is it hard to understand and recognize the difference between fraud, abuse, and mistakes, therefore sophisticated techniques are required. Furthermore, the involvement of humans is a critical aspect of fraud detection and an understanding of the data is an essential key to effective detection. Systems cannot understand the complexity and dynamic nature of fraud detection; therefore it is essential to keep humans involved. Due to the huge amount of data, humans need fraud detection systems to support the data analysis process to scan for fraud and abuse. To create solid fraud detection systems humans and computers have to collaborate and work closely together.

Given the fact that the input of beneficiaries is reduced to a minimum, electronic fraud detection is even more important in the Medicaid program than most other related industries. Supervised data mining such

as classification has been proven to successfully support the fraud detection in the analyzed industries. The classification of fraudulent and legitimate transactions (are) based upon transaction achieved in the past and therefore it can detect the sophisticated fraud schemes existing. However, the Medicaid program does not possess labeled data to indicate which claims are fraudulent or abusive. Labeled data is an absolute requirement to apply supervised data mining and multiple stakeholders and fragmented responsibility are also hampering the process of labeling fraudulent and abusive data. Therefore, the most important data analysis opportunity in the form of supervised classification is severely restricted. Supervised classification is the most sophisticated data analysis technique because the model can be trained and adjusted and therefore best suited to detect sophisticated fraud schemes. In the credit card industry supervised classification, neural networks in the 1990s, and currently support vector machines and random forests, form the basis for sophisticated and effective fraud detection. Certainly the argument that new fraud schemes are not detectable since the training data does not contain the newest fraud schemes is legitimate. The addition of unsupervised data mine techniques such as anomaly detection would enable the fraud detection system to detect new sorts of fraud. Furthermore, profiling and rule based methods have been proven to be successful in the related industries and therefore would be an effective addition to the fraud detection system.

Contribution

The result of a systematic literature review includes an overview of the various Medicaid fraud and abuse schemes. An overview is presented in the table below.

Fraud Type	Fraud Scheme	Short Explanation	Strategy
I	**Identity Theft**	Stealing identification information from providers and beneficiaries and using that information to submit fraudulent bills to Medicaid.	Fraud
II	**Fictitious Practitioners**	Using false documents and identification information to submit fraudulent bills to Medicaid.	Fraud
III	**Phantom Billing**	Submitted claims for services not provided.	Fraud
IV	**Duplicate Billing**	Submitting similar claims more than once.	Fraud/ Abuse
V	**Bill Padding**	Providing unnecessary services and the submitting these claims to Medicaid.	Fraud/ Abuse
VI	**Upcoding**	Billing for a service with a higher reimbursement rate than the service which was actually provided.	Fraud/ Abuse
VII	**Unbundling**	Submitting several claims for various services that should only be billed as one master claim that includes ancillary services.	Fraud/ Abuse

Table 1: Overview of Fraud Schemes

Subsequently an attempt to provide a framework of the applied electronic fraud detection techniques in related fraud detection industries is shown in the following figure.

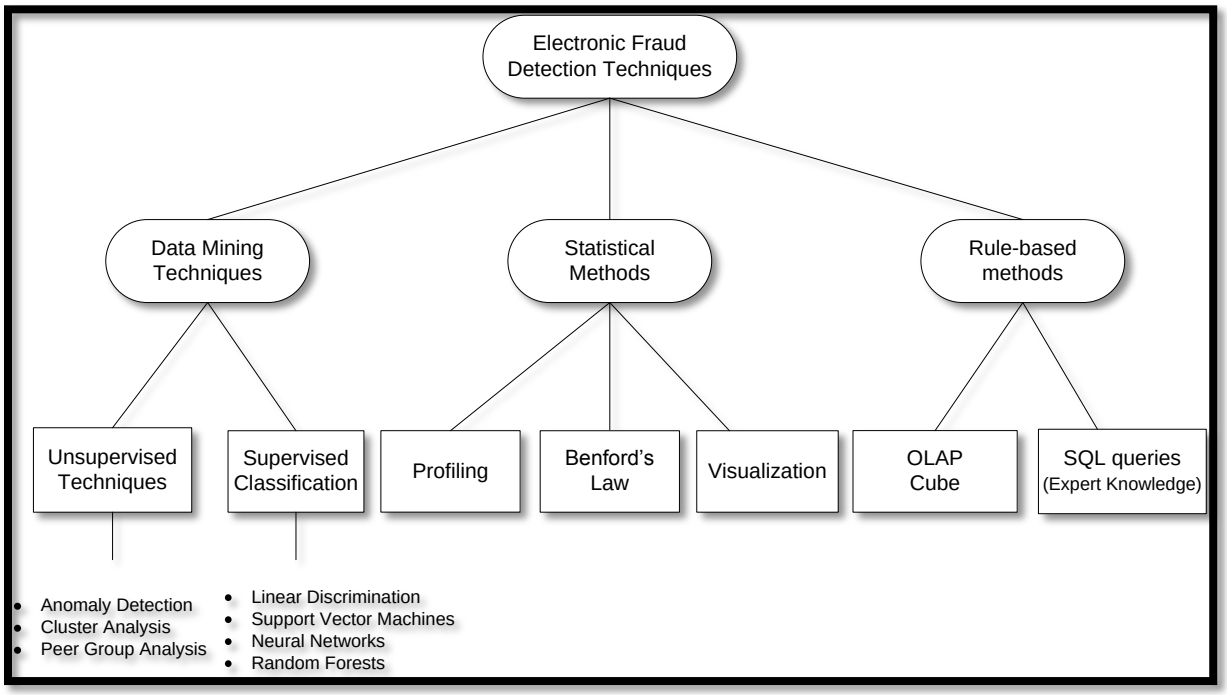


Figure 1: Framework of Electronic Fraud Detection Techniques

Consequently, the feasibility of the several fraud detection techniques is discussed for every fraud and abuse scheme. This theoretical approach is the first step to provide an overview of the current Medicaid situation and is based on a literature review. However, further research is required to empirically test the fraud detection techniques and pilots need to be established for case studies with real data sets. In this study an attempt is made to outline the importance and complexity of the fraud and abuse problems in Medicaid and solutions regarding electronic fraud detection techniques.

Implications

This research contains several implications for the Medicaid fraud detection:

- Awareness of the magnitude of the Medicaid problem;
- A preparatory step for further in depth research about how involved stakeholders can collaborate in a more effective manner;
- Awareness of the complexity of fraud detection in the Medicaid program;
- Keep humans involved in the fraud detection process supported by information technology;
- Sophisticated and complex supervised classification techniques have been proven to be successful in related industries regarding electronic fraud detection support;
- The need for a modular and flexible fraud detection system consisting of several fraud detection techniques;
- A national effort to label fraudulent data to enable supervised classification in the Medicaid Integrity Program.

Preface

Several months of research has resulted in this thesis, the last milestone in my university career as a student Industrial Engineering & Management. I owe an enormous debt of gratitude to a number of people for their support during the road towards the completion of my graduation project.

First of all, I would like to thank my supervisors of the University of Twente; Roland Müller and Jos van Hillegersberg for this opportunity and their enthusiasm from the start onwards. I appreciate the sharp suggestions, remarks, relevant literature, and support during this project. The clear comments absolutely improved my research quality and the sharp deadlines gave me confidence and motivation. You were always available via email and Skype to give me guidance and support to make this project a success.

Furthermore I would like to thank my supervisor Dallas Thornton, Division Director Cyberinfrastructure Services at the San Diego Supercomputer Center, for offering me the opportunity to live and work in San Diego, California. Due to your faith in the outcome and the significant latitude I felt strong-minded to successfully continue the project in San Diego.

Moreover I would like to thank all my fellow colleagues at the Supercomputer Center for their ongoing support and input and the great atmosphere. The discussions helped me tremendously to understand the complex world of the U.S. health care insurance programs and the corresponding problems. Furthermore, I am particularly grateful to Shailendra Revankar, my direct supervisor in San Diego, for his precise support and comments and of course to all the fun we had outside the Supercomputer Center. Starting up the implementation project from the beginning and building a strong team, was the foundation of the beginning of my professional career.

Of great importance is the support of all my friends for their ongoing support whether I was in Hengelo (OV), San Diego, CA or New York. Especially I would like to express my appreciation to Kat Mara for her role as editor of my thesis and a great buddy.

Finally, I would like to express my last word of thanks to my family, for their unconditional faith and support. The strong relationships with family and friends form the basis for me to chase and fulfill my dreams.

Hopefully you will enjoy reading this report as much as I have throughout my graduation project.

Contents

1. Introduction

'Health care fraud; what you see is never the problem; it is what you not see'

– Professor Dr. Malcolm Sparrow (License to Steal, page 119)

Based on health care fraud research (e.g. Sparrow (2002)) a significant amount of the health care budget is being lost to fraud or fraudulent behavior in the United States health care programs Medicaid and Medicare. In this research the focus will be on the Medicaid program which is governed by the states and overseen by the federal government. In the program it seems relatively simple to hide or modify essential information for the control system of the state government (GAO, 2000; Sparrow, 2002 and Hyman, 2002). This might contribute to opportunistic behavior of health care providers in the form of fraud and abuse of the program. The incentive of the providers is to maximize their profit and if the risk of being caught is small enough, fraud might be an enticing option (Derrig, 2003, Todd & Benbasat, 1999, Sparrow, 2002). After a literature study and interviews it seems that the fraud detection units of the Medicaid program do not have the proper resources to Medicaid fraud. Medicaid is available throughout the United States including some special regions (e.g. Puerto Rico) and due to the variety of legislation and organization amongst the fifty states it is hard to apply an effective general fraud detection system. The magnitude of the program is represented in the yearly budget of $321 billion dollars (Fiscal Year 2008) and the 49.8 million people who are enrolled (Chapterhouse, 2008).

The foundation of this research will be a systematic literature study to analyze the existing electronic fraud detection techniques applied in related industries. The claiming process is outlined in chapter 2 to provide insight into the process and the fraud schemes detected in the past. In chapter 3 an outline is provided to illustrate how currently information technology is used to support the fraud detection process and which opportunities can be extracted from the related industries. Subsequently, the lessons learned from the related industries, the advantages, disadvantages, applicability and constraints of the fraud detection techniques in the Medicaid program are discussed in chapter 4. An evaluation proposal is presented to further investigate the applicability and effectiveness of the fraud detection techniques. Chapter 5 provides an insight in the federal program to combat the Medicaid fraud and abuse with a national Medicaid database. Finally the lessons learned are formulated in the conclusions and discussion and recommendations to empirically test the techniques are provided.

1.1 Research Focus

The focus of this research will be on the opportunities of electronic fraud detection techniques. A literature review of the academic literature is conducted regarding the availability of data driven electronic fraud detection techniques and current practices in the computer intrusion field, telecommunications and credit card industry. Subsequently the opportunities and requirements of applying data analysis in the Medicaid health insurance program are discussed to improve the fraud detection process.

The importance of available datasets accessible for analysis and monitoring behavior of providers and recipients is highlighted in the literature. Furthermore, a case study is conducted at the national Medicaid database in San Diego where all the Medicaid of the last ten years is hosted. This national project is a federal attempt to prevent fraud and abuse of the Medicaid insurance program. In the current literature the

focus is mainly on applying statistics, machine learning, data mining and knowledge discovery to uncover patterns which suggest that academics have a tendency to focus on just one small part of the problem (Hand, 2010). The workflow management system is a centralized information system that supports the whole process from fraud detection until the recovery of illegitimate payments.

The magnitude of the Medicaid program is comprehensive and the quantity of health insurance claims in the national Medicaid database surpasses three billion claims per year (including adjustments of submitted claims).

1.2 U.S. Health Care Fraud in Medicaid and Medicare

Medicaid and Medicare are two government programs that provide medical and health-related services to specific groups of people in the United States. Although the two programs are very different, they are both managed by the Centers for Medicare and Medicaid Services (CMS), a division of the U.S. Department of Health and Human Services. Medicare is a federal program which has consistent rules across the fifty states and covers almost everyone 65 years of age or older. Medicaid is a state administered program in which each state provides a unique health care program for individuals and families with low incomes and resources (CMS, 2010). The eligibility differs per state and each state sets its own unique guidelines regarding eligibility and services. It applies to people living below the federal poverty line and to special categories of people for example pregnant women and children (CMS, 2010). This program became law in 1965 as a jointly funded cooperative venture between the federal and state governments to assist states in the provision of more adequate medical care to eligible needy persons. The Medicaid Program provides medical benefits to groups of low income people, some who may have no medical insurance or inadequate medical insurance. Within the broad national guidelines provided by the federal government, each of the States (CMS, 2010):

(1) Establishes its own eligibility standards
(2) Determines the type, amount, duration, and scope of services
(3) Sets the rate of payment for services
(4) Administers its own program

Thus, Medicaid programs vary considerably from state to state and within each state over time. Medicaid operates as a vendor payment program, within which states pay providers directly. With a few specific exceptions, each state has broad discretion in determining the reimbursement methodology and resulting rate for services within federally-imposed upper limits and specific restrictions. The portion of each state's Medicaid program which is paid by the federal government, known as the federal Medical Assistance Percentage (FMAP), is determined annually by a formula that compares the state's average per capita income level with the national income average. By law, the FMAP cannot be lower than 50 percent and not exceeding 83 percent. The wealthier states have a smaller share of their costs reimbursed. In 1994, the FMAPs varied from 50 percent (paid to 11 States) to 78.9 percent (to Mississippi); with the average federal share among all States being 57.5 percent (Center for Regulatory Effectiveness, 2010).

Currently the United States faces a serious fraud problem concerning the social health care. Although the differences between the programs Medicaid and Medicare are significant, the detected fraud schemes appear in both programs (Sparrow, 2002). The fundamental reason is that the industry’s standard detection and control systems are not aimed at criminal fraud at all (Hyman, 2001). The software “edits” and “audits” are built into modern, highly automated claims processing systems which have all been

designed with honest providers in mind and serve the purpose of catching errors, verifying eligibility, making sure procedure codes match up with the diagnosis, and checking that the price charged is in within bounds (Sparrow, 2002). The system is rather designed to reimburse honest providers and not to prevent fraud, waste and abuse. Furthermore the quantity of the Medicaid claims is an important factor to take into account and it is governed by the states which generate diversity in rules and the resulting sets of claiming data. Billions of claims are submitted and the processing accuracy is used as an measurement control to avoid high processing costs. However claim verification is an important aspect which is much harder to measure but just as important. A system of controls that would routinely check that services paid were actually administered and that those services were medically necessary is currently not in place. This lack of supervision gives providers and other people with fraudulent intentions the opportunity to get away with fraudulent behavior (Sparrow, 2002).

The states process the submitted claims and currently they do not have the resources to accurately verify submitted Medicaid claims. Billions of claims enter the system per year and the current detection system is not prepared to verify all of these claims. Based upon several studies it can be concluded that a significant amount of the programs expenses are due to fraud, waste and abuse (Sparrow, 2002; Hyman, 2001).

1.3 Research Questions

Section 1.2 highlighted the current worrisome situation of the social healthcare program Medicaid in the United States; providing this research with a focus on electronic fraud detection systems. Researching the use of information technology in fraud detection applications will provide an overview of the widely used fraud detection techniques. The applicability of the fraud detection techniques in the Medicaid healthcare program is discussed per discovered fraud type. Furthermore the research is complemented with a case study at the national data center of the U.S. Medicaid program. An in depth analysis of a workflow management system is presented which is currently implemented to support the post payment overutilization review

The broad research focus on electronic fraud detection systems in fraud detection domains throughout several industries combined with the focus on the complex health care program Medicaid leads to the formulation of the main research question:

How do electronic fraud detection systems facilitate security in other industries and how can the fraud detection of the U.S. health care program Medicaid be improved?

Based on the literature study on Medicaid (as described in chapter 2) and fraud detection techniques (as described in chapter 3), this research identifies three research sub question to support the analysis.

1.4 Research Sub Questions

The sub questions are based on the literature study to support the analysis. to investigate several fraud detection techniques from related fraud detection industries and to verify the findings in the Medicaid application.

SQ1: What are structural factors of the U.S. health care program Medicaid and what is the current situation regarding fraud and fraud detection?

Structural factors describe the relatively stable aspects of an insurance program; based upon the actors and systems within the program with the appearing consequences.

SQ2: What are the important aspects and requirements for an effective electronic fraud detection system?

Electronic fraud detection techniques comprehend the analysis of data that can be classified as one of the data analysis function: data mining techniques, statistical methods, and rule-based methods. The foundation of these techniques and the corresponding requirements are researched, outlined and discussed.

SQ3: What important fraud detection techniques from other industries can be applied in the fraud detection of the Medicaid program?

Analysis of the applicability of the fraud detection techniques leads to an overview of the fraud detection opportunities and limitations in the Medicaid program.

The three sub questions allow for a complete analysis of the U.S. Medicaid health care program and related fraud detection domains. The results of the analysis are used to theoretically apply the fraud detection techniques on the Medicaid program to evaluate the opportunities and limitations to combat health care fraud and abuse.

1.5 Research Method

A systematic literature review will be conducted to collect the data and gather understanding of the Medicaid health care program and electronic fraud detection techniques. Inspired by case studies in the literature, where successfully electronic fraud detection has been applied, the opportunities and limitations of electronic fraud detection in Medicaid are outlined and discussed. These case studies include the credit card, telecommunication and computer intrusion industries where fraud detection is an important aspect of the business. Furthermore by conducting a systematic literature review the Medicaid program with its corresponding fraud and abuse problems is outlined. What lessons can we learn from the other industries concerning the opportunities and limitations of electronic fraud detection? The systematic literature review is supported by some interviews at the San Diego datacenter to gain further insight in the Medicaid opportunities and limitations regarding electronic fraud detection.

Based upon the findings of the systematic literature review appropriate fraud detection techniques are proposed in the Medicaid setting to fight the current fraud and abuse schemes. Subsequently an expert proposal will be presented to indicate an approach to empirically test the outcomes of this systematic literature review. Empirical testing of the fraud detection techniques is not included in this research.

1.6 Outline of the Thesis

Chapter 2 introduces the U.S. health care program Medicaid and the current set up of the Medicaid insurance system. The consequences of the set up are outlined in the form of the current fraud problems the U.S. is facing and an outline of the detected fraud schemes in the past is provided. In chapter 3 a systematic literature study is outlined which forms the foundation for the analysis of the Medicaid health

care program and electronic fraud detection techniques applied in related industries. Fraud detection models known in the scientific literature and case studies applied in the credit card, telecommunications, financial and insurance industries are reviewed an overview of the existing fraud schemes is presented. In chapter 4 the opportunities and limitations of applying the fraud detection techniques in the Medicaid program are discussed and the opportunities and limitations corresponding with the data driven approach are outlined. In chapter 5 the Medicaid Integrity Program is outlined to provide an insight of the current status of the national Medicaid electronic fraud detection program. Chapter 6 concludes this study with by answering the research questions and the corresponding conclusions. Finally the recommendations for further research are presented.

2. Medicaid and Health Care Fraud

'The worst possible situation to be in is to believe that we have a huge problem but no be able to prove it to anyone. That is precisely the condition within the health care industry.'

– Professor Dr. Malcolm Sparrow

In this chapter published work and overview reports relevant to fraud detection in the Medicaid system are reviewed and the current situation of the Medicaid program is discussed. The following sections contain the general information and stakeholders of the Medicaid program (§2.1), furthermore the phenomenon of asymmetric information is discussed (§2.2), next is the claim submission process (§2.3) followed by the fraud and abuse schemes detected in the past (§2.4 and §2.5).

2.1 Medicaid

In 2008 according to the Centers for Medicare and Medicaid (CMS, 2010) 49 million people were enrolled in Medicaid and the budget was $321 billion (Chapterhouse, 2008). The following table provides an overview of the stakeholders involved and their goal and responsibility.

Stakeholder	Explanation organization	Responsibility
Center for Medicare and Medicaid (CMS)	Federal agency that oversees and administers the U.S. Medicare and Medicaid program.	Supporting states to combat health care fraud, abuse and waste in the programs Medicaid and Medicare.
Department of Justice (DOJ)	To enforce the law and defend the interests of the United States.	Investigating health care fraud and bringing fraudsters to justice.
Department of Health and Human Services (HHS)	United States government's principal agency for protecting the health of U.S. citizens.	Coordinating federal, state and local law enforcement activities with respect to health care fraud and abuse.
Office of Inspector General (OIG)	To protect the integrity of HHS programs, as well as the health and welfare of the beneficiaries of those programs.	Reporting management problems and making recommendations to correct them. The OIG's duties are carried out through a nationwide network of audits, investigations, inspections and other mission-related functions performed by OIG components.
Medicaid Fraud Control Units (MFCU)	Fraud detection teams organized per state are required by statute to employ investigators, auditors, and attorneys.	To investigate and prosecute fraud by Medicaid providers as well as patient abuse and neglect.
U.S General Accounting Office (GAO)	The GAO investigates how the federal government spends taxpayer dollars and governs Medicaid Fraud Control Units.	To help improve the performance and ensure the accountability of the federal government for the benefit of the American people.
Federal Bureau of Investigation (FBI)	National intelligence agency designed to protect the United States.	Investigating national health care fraud to protect the U.S. citizens from fraud.

Table 2: Overview Stakeholders Medicaid

The Federal Bureau of Investigation (FBI) plays a major role in assisting the Department of Justice (DOJ) in investigating and developing health care fraud cases. Within the Department of Health and Human Services (HHS), the Office of Inspector General (OIG) is responsible for investigating fraud cases and bringing enforcements actions involving administrative sanctions. The General Accounting Office (GAO) investigates at a federal level how the tax dollars are spent; in the case of Medicaid how money disappears due to health care fraud and abuse. The GAO is supported on a state level by the Medicaid Fraud Control Units (MFCU) which work closely with local investigators, auditors and attorneys. Individual states have their own Medicaid fraud control units and local prosecutors who can bring such cases to justice (Hyman, 2002).

The Centers for Medicare & Medicaid Services, as part of the Department of Health and Human Services, oversees and administers the U.S. Medicare and Medicaid programs. Section 6034 of the Deficit Reduction Act (DRA) of 2005 established the Medicaid Integrity Program (MIP) in Section 1936 of the Social Security Act. MIP is a five-year comprehensive plan to combat fraud, waste, and abuse beginning in 2006. The MIP goals which include:

- Increasing provider audits
- Increasing state collections by identifying inappropriate payments
- Helping the states improve their payment systems by sharing successful data algorithms, best practices, and lessons learned

Medicaid Integrity Program is a federal attempt to reduce the fraud and abuse in the Medicaid program since the states have difficulties to control the situation. The problem the government has to manage is the asymmetric information which is explained in the next section.

2.2 Asymmetric Information and the Principal Agent Problem

To provide additional background information about relevant aspects of the insurance problem an introduction of the asymmetric information, agency theory and principal agent is provided. An insurance company simply does not have all of the information about the claim compared to the customers. The asymmetric information and the relationship with fraud and abuse are outlined in this section.

Agency theory in its simplest form is a relationship between two people: a principal and an agent who takes actions and makes decisions on behalf of the principal (Douma and Schreuder, 2002). Some examples are:

- The relationship between a landlord (the principal) and a tenant farmer (the agent)
- The owner of the firm (the principal) and the manager of the firm (the agent).
- An insurance company (the principal) and the customers (agents)

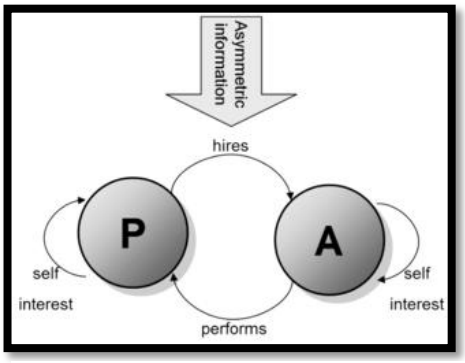


Figure 2: Principal-Agent relationship (Eisenhardt, 1989)

The principal-agent approach is a thorough and insightful treatment of the case of a single principal who engages an agent to perform a task. As shown in figure 3 both parties have to deal with self interest and the other party. The key aspect is the amount of information that is available so that both of the approaches can be stretched out in terms of cases. The agent makes decisions and takes actions on behalf of the principal and the problem arises if those decisions and actions are not beneficial and lucrative for both parties. However in reality it is quite hard for the principal to obtain the information necessary to control or monitor the behavior of the agents, this is because the principal has to deal with asymmetric information. Either the agent's information differs from that of the principal's, or it is too costly for the principal to monitor the behavior of the agents. Gathering extra information to verify the situation and reduce the information difference costs money and resources. While buying a second hand car a car inspection by a car expert would provide information about the status of the car and it could reveal hidden damages. However this would costs a bit extra but it provides relevant information to verify if the car is in the right condition. In this case the principal (buyer) does not know exactly what the agent (seller) has done with the care. Given the self-interest of the seller, the seller is not guaranteed to behave as expected or agreed upon. According to Eisenhardt (1989) this is where the agency problem arises because (I) the principal and the agent have different goals and (II) the principal cannot determine if the agent has behaved appropriately. The principal can assign resources to gather more information about the behavior, however this involves extra costs and it is practically impossible to grasp.

Risk and Adverse Selection

Oliver Williamson's work on economic governance has been rewarded with the Nobel Memorial Prize in Economics, including his two books *Markets and Hierarchies* (1975) and *The Economic Institutions of Capitalism* (1985). In Williamson's view people are not only boundedly rational, but rather sometimes human beings display opportunistic behavior. He describes this as 'self interest seeking with guile'; or in other words, trying to exploit the situation to your own advantage. This is in line with Eisenhard's assumption that people are self interested, bounded rationality, and averse to risk (Eisenhard, 1989).

Life Insurance for Smokers

Adverse selection refers to a transaction process where the party with the least or worst information makes relatively the worst decision. Adverse selection can appear whenever a group or individual has the freedom to decide whether to buy or not to buy (Akerlof, 1970). An example of this potentially adverse phenomenon is a life insurance policy for smokers and non-smokers. If the insurance company offers the same price for both parties it is more attractive for smokers to purchase this particular life insurance. "There is a potential adverse selection in the fact that healthy term insurance policy holders may decide to terminate their coverage when they become older and premiums mount" (Akerlof, 1970). Smoking is bad for your health and therefore non-smokers are assumed to live longer and the life insurance will be more expensive. Smokers will be likely to buy this insurance with the result that the policyholder group shifts to more smokers and that the average mortality will be higher than the general average mortality rate. This will result in higher costs for the insurance company so the prices will rise, fewer non-smokers will

be likely to buy this insurance and the percentage of non-smokers in the policyholder group is reduced. The costs will keep on rising and the situation is out of balance.

Concerning the U.S. health care insurance a similar phenomenon has occurred. Since a medical insurance is not obligatory and therefore is the market mechanism determinative. Theoretically healthy people do not need a health insurance and therefore they do not contribute in the form of paying for the health insurance. The Americans health care system costs twice as much as residents of other developed countries on healthcare, but get lower quality, less efficiency and have the least equitable system, according to a report of the Commonwealth Fund report (Davis *et al.* 2010). The United States ranked last when compared to six other countries; Britain, Canada, Germany, Netherlands, Australia and New Zealand. Every other system covers all its citizens, the report noted the U.S. system leaves 46 million Americans or 15 percent of the population without health insurance.

To provide the principal with relevant information and reduce the asymmetric information between the two parties data analysis is a widely applied in several industries. Automated processing and the importance to keep humans involved in the processing are outlined in the next section.

2.3 The Claim Submitting Process

To illustrate the claiming process an example of an accident will be given with the corresponding cost of treatments provided to the beneficiary. A certain beneficiary got enrolled in the Medicaid program after an eligibility screening performed by Medicaid. The person's circumstances and income will be verified and if the criteria of the state are met the person can enroll in the Medicaid program. At a certain moment the beneficiary has a car accident and he ends up in the hospital where he got the necessary treatment. An overview of the treatment and the corresponding costs are given in the following figure.

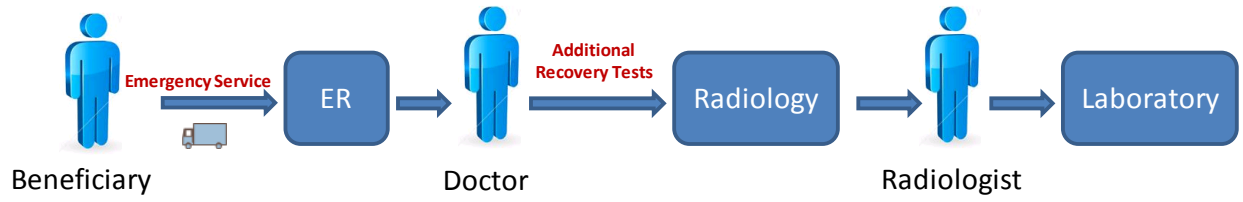


Figure 3: Example of medical services

As depicted in figure 4, the providers of the services provided to the beneficiary (e.g. ambulance, laboratory tests) will submit the bill individually to Medicaid because most of the providers work independently. The participation of the providers determines the price of the medical service and Medicaid reimburses a fixed amount per service. It is generally assumed that all of the providers have an agreement with Medicaid and that they participate in the program. However that is not always the case as shown in figure 5; the laboratory does not participate in the Medicaid program. Since Medicaid only reimburses $25 for the blood test and the total laboratory bill is $350, the beneficiary has to pay $325 out of his own pocket for this service. The laboratory does not have an agreement with Medicaid and therefore they can charge their own prices determined by the company itself.

The Bill:		Participating?	Medicaid reimburses
• Ambulance	$1000	Yes	$ 1000
• Hospital (ER):	$ 450	Yes	$ 450
• Doctor:	$ 150	Yes	$ 150
• Radiology:	$ 600	Yes	$ 600
• Radiologist:	$ 150	Yes	$ 150
• Laboratory:	$ 350 +	No	$ 25 +
Total	$ 2700		$ 2375

Figure 4: The overall medical bill

The reimbursement rates stated by Medicaid are not taken into account when billing for these services, consequently this often means that beneficiaries have to pay the rest of the costs. For the payment of the services of participating providers the beneficiaries do not have to take any actions. As shown below the provider sends the claims directly to Medicaid where the claim is reviewed and processed.

When a provider participates in the Medicaid program it includes that the provider agrees to the reimbursement rates set by Medicaid; the provider can simply send the bill directly to Medicaid. A different scenario takes place when a provider is not enrolled in the Medicaid program (see the laboratory cost in figure 3); then the provider sends the beneficiary the bill which he has to pay and then submit a request for reimbursement to Medicaid. In both scenarios an Explanation of Benefits (EOB) is sent to the beneficiary; an EOB contains an overview of the provided services. This is an automatically generated detailed overview with the provided services and corresponding codes and amounts.

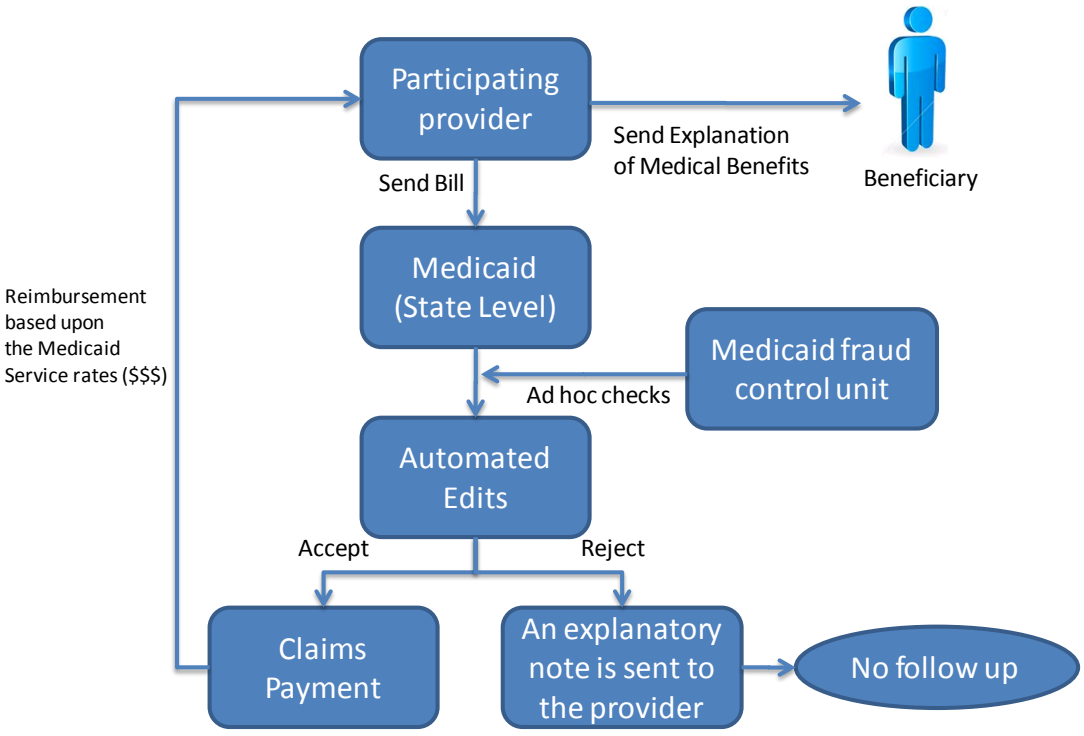


Figure 5: Provider submitting a claim to Medicaid

Every state is responsible for organizing and governing Medicaid themselves, it is their responsibility to process the submitted claims and verify if the claim is legitimate. The states process the claims with the support of a claim processing system which is different in every state. Medicaid receives the bill and the claims processing system performs several prepayment checks and edits to verify if the claim is legitimate. Sparrow (2002) provides some examples of the automated audits:

- Is the data entered correctly? Have the fields been properly filled in? (e.g. field date exists only out of numbers and no words, does every field contain information)
- Procedure codes matches diagnosis?
- Is the pricing in range with the set boundaries for the service or procedure?
- Duplicate claims: has the claim been submitted and paid already?

The edits and audits are designed to verify the information with honest providers in mind, the system lacks of effective fraud-referral mechanisms (Sparrow, 2002). These systems do not verify if the service was provided as claimed, or that the diagnosis is genuine, or that the patient is aware of the claimed services. The system assumes that the information is true and genuine and Sparrow (2002) points this out with a striking example. If a claim is rejected by the system, there is no follow up to investigate why the provider submitted this claim. Instead of vetting these claims, the system sends an explanation note with the reason why the claim has been rejected and where it went wrong. So instead of investigating the possible fraud, the fraud perpetrators get wiser and learn about the billing rules.

The explanations of medical benefits (EOB) do not provide protection against fraud either. Sparrow (2002) points out the reasons:

1. Often EOB's are not sent at all
2. Little or no financial incentive to pay attention to them
3. Recipients do not understand the complex computer generated forms and there is no reason to try
4. Fraudulent suppliers find innovative ways to withhold recipients of opening the envelopes (examples of suppliers paying $5 per not opened envelope)
5. Many fraud schemes deliberately target vulnerable populations (Kelly, 2009)

No follow up after the claim is rejected is a major problem if a fraudster tries to defraud the program. The fraudster receives the information with the explanation why the program rejected the claim. The lesson how not to submit the claim can be mastered and in the next attempt taken into account. Especially when the program requested for more information a significant number of providers did not reply al all (Sparrow, 2002). This seems highly suspicious because why would a provider not hand in extra documentation to prove the service was provided? In the end it is his income and if he provided the service it is not a lot of effort to submit additional documentation. However a fraudulent provider has a good reason not to reply; the fraud attempt failed and there are no consequences when he does not reply since there is no follow up. However if an honest provider made a mistake with the submitted claim he would send the additional documents to get his provided services reimbursed. Further investigation of rejected claims is a major source of possible fraudulent attempts (Sparrow, 2002) and the Medicaid Fraud Detection Teams can benefit of this information source.

2.4 Definition of Fraud and Abuse

Fraud, defined as 'criminal deception or the use of false representations to gain an advantage' in the Oxford Dictionary, is as old as humanity itself and occurs in different degrees of severity. The terms fraud and abuse used in the literature encompass a wide spectrum of conduct, ranging from intentional misrepresentation of services provided to inadequate documentation of provided care (Hyman, 2001). The terms fraud and abuse are frequently used interchangeably in the literature. Definitions more specific for the health care industry is given by the Center for Medicare and Medicaid Services (2010):

Fraud: To purposely bill for services that were never given or to bill for a service that has a higher reimbursement than the service produced. The intentional deception or misrepresentation that an individual knows (or should know) to be false, and makes, knowing the deception could result in some unauthorized benefit to himself or some other person(s).

Abuse: Payment for items or services that are billed by mistake by providers, but should not be paid for by Medicaid. This is not the same as fraud. Basically abuse is applied to provider practices that result in unnecessary costs to Medicaid; a range of the following improper behaviors or billing practices including; billing for a non-covered service; misusing codes on the claim (i.e., the service coded on the claim) does not comply with national or local coding guidelines or is not billed as shown. More examples are billing for services provided by unqualified individuals, or providing services not medically necessary.

Legislative Attempts to Fight Fraud and Abuse

The detection of fraud and abuse in the health care system is a comprehensive and challenging task, however to prove it in court might be even harder. If real fraud were relatively easy to recognize then there is no doubt that red flags would go up every time a physiotherapist orders 12.5 miles of one inch adhesive tape for a single patient (sufficient to wrap the patient from head to toe in adhesive tape six times a day for six months (Hyman, 2001)). However there are several levels of fraud which are significantly harder to judge if it is a violation of the law. Several laws are developed to fight fraud; the federal anti-kickback statute (U.S. Code; § 1320a-7b) prohibits individuals or entities from knowingly and willfully offering, paying, soliciting or receiving remuneration to induce referrals of items or services covered by Medicare, Medicaid or any other federally funded program. The self-referral provision (U.S. Code; § 1395) which encounters the limitation on certain physician referrals if there is a financial relationship. The Civil False Claim Act (U.S. Code; §3729 False claims) is a Federal law which allows people who are not affiliated with the government to file actions against federal contractors they accuse of committing claims fraud against the government. These laws are the three most significant health care specific fraud control provisions currently in effect. According to Hyman (2001) and Sparrow (2002) addressing health care fraud is exceedingly complicated and the fraud control measurements were primarily designed to protect the fiscal integrity of programs like Medicaid and Medicare. The Medicaid budget was $321 billion in 2008, and all of the stakeholders try to benefit from the opportunities in the lack of legislation. The legal aspects of fraud detection will not be taken into account concerning the scope of the research.

2.5 Medical Insurance Fraud and Abuse

Several fraud schemes to defraud Medicaid and Medicare are presented in this section to demonstrate the fraud and abuse problems in the Medicaid program. This does not imply that these schemes are the only fraudulent schemes that currently exist in the U.S. health care system; however, these are all of the

published fraudulent schemes uncovered in the past. As stated before publishing about fraud is not an effective way to combat fraud since information is shared with the fraudsters as well. Probably unknown and new fraud schemes have been developed which are not discovered or published and are therefore this overview should not be considered as complete.

2.5.1 Fraud Strategies

Sparrow (2002) points out two particular categories of fraud perpetrators of which the fraud control system should be aware of. These categories are both considered to be the extremes of the fraud spectrum, the "hit-and-run" scheme and the "steal a little, all the time" scheme. The short term heavy hit is a strategy to get high amounts of money in a short time and disappear afterward before anyone realizes what happens. Hit-and-run operations bill quickly and furiously because they know that their time is limited. At the opposite extreme lies the white-collar criminal who steals a little all the time to get rich over the long term. Legitimate health care providers who provide genuine services use their bulk of legitimate claims to hide their stealing. "When they steal, they use familiar methods as billing for services not provided, billing for more expensive services or products than those actually provided and falsifying diagnoses to support more expensive claims" (Sparrow, 2002). Although there are many other strategies, and many variations of these strategies, these two polar strategies are not adequately controlled by the existing systems.

Medicaid	Telecommunications	Credit Card
Hit and run	Subscription fraud	Application fraud
Steal a little all the time	Superimposed fraud	Behavioral fraud

Table 3: Types of fraud throughout the fraud detection industries

Sparrow's (2002) analysis is in line with fraud strategies present in other industries such as the telecommunications and credit card fraud detection. In the telecommunications industry Cahill *et al.* (2002) describe subscription fraud (false identification and no intention to pay) and superimposed fraud (slow and hidden). Similar fraud schemes exist in the credit card fraud detection like application fraud and behavioral fraud (Bolton and Hand, 2001). The big difference between the two extremes is the self revealing aspect (Sparrow, 2002); whereas the hit and run techniques are in the telecommunications and credit card industry self revealing because customers are losing money rapidly. Steal a little all the time might continue because the customers do not notice the fraud committed with their accounts because it is covered by all the legitimate transactions. Those strategies might even be more dangerous since it is continuously draining the system. If the customer is not paying the bill because the account is defrauded the company will block the account immediately to prevent bigger profit losses. Furthermore credit card companies have the ability to redraw the money from the merchants until they prove the opposite. In the Medicaid program states and the federal government do not posses real time data to react quickly and they are not able to redraw the money from the providers once it is paid. Furthermore is the self revealing aspect completely missing in the Medicaid program since beneficiaries do not have the right incentives to report fraud and abuse.

A more specific overview is provided in the following sections to describe the fraud schemes in detail. This will provide some insight about the existing and known Medicaid fraud schemes detected in the past.

Fraud Type	Fraud Scheme	Short Explanation	Strategy	§
I	**Identity Theft**	Stealing identification information from providers and beneficiaries and using that information to submit fraudulent bills to Medicaid.	Fraud	2.5.2
II	**Fictitious Practitioners**	Using false documents and identification information to submit fraudulent bills to Medicaid.	Fraud	2.5.3
III	**Phantom Billing**	Submitted claims for services not provided.	Fraud	2.5.4
IV	**Duplicate Billing**	Submitting similar claims more than once.	Fraud/ Abuse	2.5.5
V	**Bill Padding**	Providing unnecessary services and the submitting these claims to Medicaid.	Fraud/ Abuse	"
VI	**Upcoding**	Billing for a service with a higher reimbursement rate than the service which was actually provided.	Fraud/ Abuse	"
VII	**Unbundling**	Submitting several claims for various services that should only be billed as one master claim that includes ancillary services.	Fraud/ Abuse	"

Table 4: Medical Fraud Schemes

The intention is the factor what determines the difference between abuse and mistakes and it is impossible to prove the intention of a provider. It is possible that a billing mistake has been made or that the billing software was billing incorrectly for certain services. However providers can easily hide behind these excuses when the intention was to defraud the program. It is basically impossible to prove the intentional what determines the difference between abuse and mistakes. However billing for services not provided, fictitious health care centers and identity theft are examples of fraud that once discovered cannot be denied. Real practitioners committing white collar crimes can hide their activities with their legitimate services. Sparrow (2002) warned for these schemes particularly since the traceability is minimized and hard to prove. The Californian Medicaid (Medi-Cal, 2010) states that fraud is generally defined as the billing of the Medi-Cal program for services, drugs, or supplies that are:

- Unnecessary
- Not performed
- More costly than those actually performed

In the end, all types of fraud and abuse must be detected to protect the financial interests of the health care program to protect the tax dollars of American citizens.

2.5.2 Fraud Type I: Identity Theft

Identity theft of beneficiaries and providers to bill for services not provided or using a provider ID to bill Medicaid are two examples of this category. To submit a Medicaid claim, the identification of the patients is required; if a patient is enrolled in the Medicaid program they have a Medicaid ID number. With that information Medicaid can be billed for the service provided as shown in figure 5; and an explanation of benefits is sent to the beneficiary. The problem is that when Medicaid beneficiaries do not check their Explanation of Benefits to verify if they received the services, therefore it is hard to discover identity

theft. Since beneficiaries do not have an (financial) incentive to verify the Explanation of Benefits, "the government is paying anyway" (Sparrow, 2002). Therefore information of Medicaid beneficiaries can be easily abused. An obtained list of beneficiaries and using those Medicaid ID numbers to bill for fictitious services is a rather common type of health care fraud. In this study this type of fraud is defined as phantom billing (§ 2.5.4) where Medicaid is billed for services, tests and products never provided. There are examples of providers who got blackmailed by criminals with a result that the provider Medicaid ID is used for criminal purposes. Often when these schemes are discovered the money is transferred to foreign accounts and is not traceable anymore.

2.5.3 Fraud Type II: Fictitious Practitioners

Fictitious health care providers that solely exist on paper are a big threat to the Medicaid program. Phantom corporations use a fictitious name to submit false claims without the intention to deliver any service. As shown in figure 7 (GAO, 2000); a commercial mail receiving agency post office box and a (stolen or illegally obtained) list of insurance information of Medicaid beneficiaries could be enough to significantly drain the system.

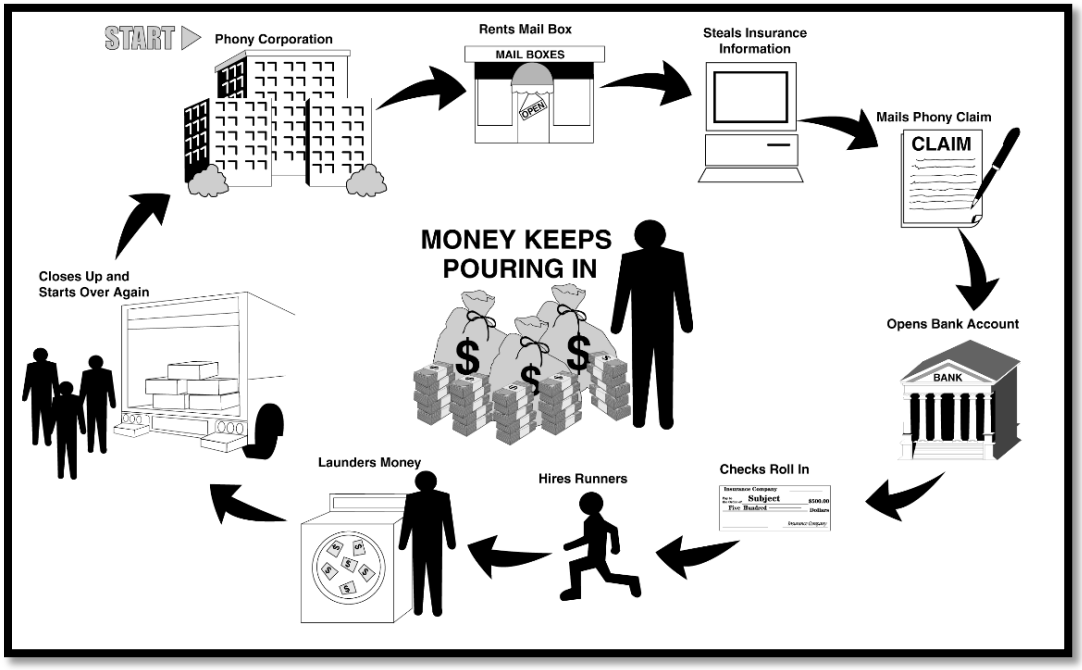


Figure 6: Fictitious Identity Source (GAO, 2000)

- In June 1994 the newspaper "Miami Herald" reported that a fictitious company named Med EO Diagnostics used the names of dozen of dead patients and rented a mailbox to collect $333,939 from Medicaid in two months. The owner of the company was caught when he tried to withdraw $200,000 at once. However, often the fraudsters vanish with the money before anyone finds out (Sparrow, 2002).

- In August 2009 an owner of a local home health care agency was convicted of defrauding Medicaid and sentence to four years in prison for stealing more than $2.2 million dollars (FBI, 2010). By using her own and a fictitious identity the owner obtained millions of U.S. tax money

by submitting false claims from February 2005 until May 2008. To increase revenues the owner was in the process of securing a third identity.

2.5.4 Fraud Type III: Phantom Billing

Billing for medical services, tests or products not provided (Sparrow, 2002); this is categorized as 'Phantom Billing'. Besides using a fictitious identity providers bill for goods and services not rendered using their own identity as well. Due to the problem of asymmetric information it is hard for any insurer to verify if a certain service took place or not. The goods and services are not delivered or provided to the beneficiaries or supplies that have never been purchased. An example is the asset theft of durable medical equipment (DME); prescribed goods which are never purchased or delivered.

- In December 1997, an ophthalmologist in California settled for $375,000 because the suit alleged him from billing continuously for a rarely used pre-cataract service. He admitted that he never provided that type of service and that he bought a list with patient information on the black market; nevertheless he routinely billed every cataract patient for this type of service (Sparrow, 2002).

2.5.5 Fraud Types IV – VII: Billing Errors / Creative Billing

A consequence of the current set up of the Medicaid fraud detection system and of (former) gray areas in the law several creative billing techniques to defraud the system have been developed. All of the abuse schemes could be a result of a billing error and as a consequence it is quite hard to prove the fraudulent intention. However the money needs to be refunded in both scenarios and all these examples are collected and explained in this miscellaneous section:

- "Duplicate Billing" is submitting similar bills with the exact same information and then paid multiple times by Medicaid. If the system does not verify if a certain claim already has been submitted the opportunity exists for providers to deliberately submit a claim multiple times. This category does include the mistakes made by providers who accidently submitted a claim twice, however their intention is not to defraud the system. Since it is not possible to define the intention when a duplicate claim has been discovered. Multiple duplicate claims submitted on a regular basis may be an indication of fraudulent intentions, the system should prevent processing these claims to solve this issue.

- "Bill Padding" is billing for services not needed but performed by the provider in order to be reimbursed by Medicaid. In September 1998, a dentist in Michigan pleaded guilty to submitting false claims to Medicaid. He was charged with abusing patients by pulling perfectly healthy teeth to create Medicaid eligibility for partial lower dentures (Morris, 1999). Furthermore examples are known of specialists who perform unneeded procedures while the patient is unconscious (Sparrow, 2002). A medical expert is needed to determine if a certain procedure is required or needed by a patient. Patients trust that doctors and health care provider will do what is in their best interest; however some providers are abusing and exploiting this situation to benefit financially. An example of an organized bill padding scheme is shown in the following figure; the 'Rent-A-Patient' scheme (GAO, 2002). Recruiters look for vulnerable people like homeless people or drug addicts and give them $20 to corporate; beneficiaries are paid to receive unnecessary tests, services and medicines.

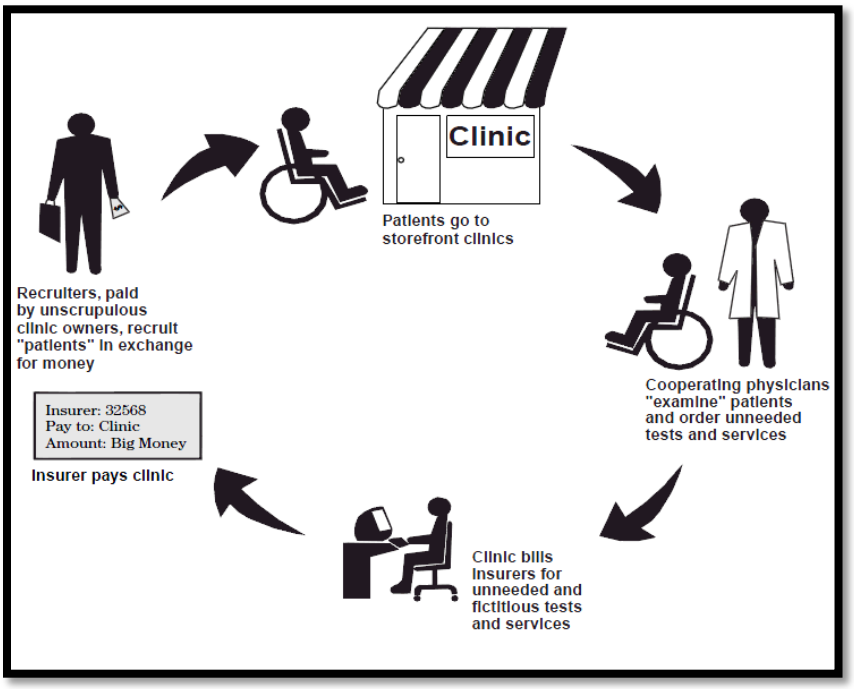


Figure 7: 'Rent-a-Patient Scheme'

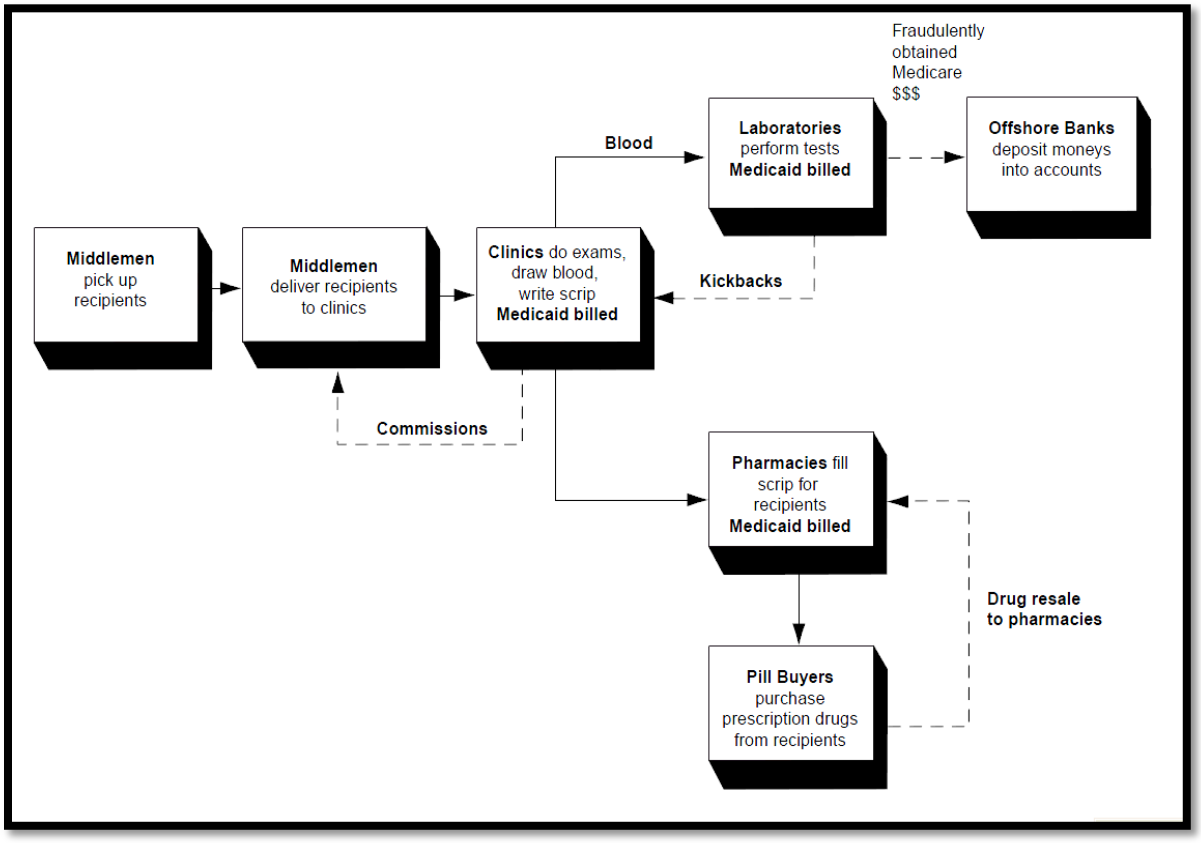


Figure 8: 'The Pill Mill scheme'

- Depicted in figure 8 is the 'Pill Mill Scheme' (GAO, 2000), a variation on the 'Rent-A-Patient' scheme. This scheme is characterized by the collusion of two or more entities, often a network of providers consisting of; clinics, brokers, laboratories, middle men and pharmacies. In a structured and organized way Medicaid is defrauded for millions of dollars and all of the stakeholders in the scheme benefit considerably as shown in the flowchart above. The medicines received by the recipients are re-sold to the pharmacies and all participating parties benefit from the scheme.

- "Upcoding" means inflating bills by using diagnosis billing codes that indicate the patient experienced medical complications and/or needed more expensive treatments (Sparrow, 2002). Several examples are; coding for a more comprehensive surgery with complications than was actually the case (Psaty *et al*, 1999), billing for a brand named drug when generic drugs were provided (Sparrow, 2002), or billing for complex services when only simple services were provided (Psaty *et al*, 1999).

- "Unbundling" is the phenomenon of billing for several services instead of using one inclusive code for a defined procedure (Sparrow, 2002). Submitting multiple bills, in order to obtain a higher reimbursement for tests and services that were performed within a specified time period and which should have been submitted as s a single bill.

Example of Miscoding/Upcoding by a Hospital

Psaty (1999) extend the HHS OIG (2002) approach to investigate the potential costs of upcoding for heart failure by comparing Medicare bills to information collected from both patient interviews and medical records. The most important criterion was the diagnosis of the physician plus confirmatory evidence like diagnostic tests (e.g. x-rays). For more than one-third of the cases examined they could not find even a modest level of supporting evidence for the diagnosis of heart failure, meaning that 37.5% of all heart failure cases reflect incorrect diagnosis. In the table the three diagnosis-related groups (DRGs) for heart failure are shown. If a patient has been diagnosed for the heart failure (DRG-127) the reimbursement in 2007 was approximately $7,000, however shown in the table below DRG-121 has a higher reimbursement of $4,000 and Psaty *et al.* (1999) conclude in their research that the miscoding of hospital discharge diagnoses for heart failure is rather common.

DRG	Diagnosis	Reimbursement
127	Heart failure and shock	$ 7,098.98
122	Accute myocardinal infarction without complications and discharched alive	$ 7,987.58
121	Accute myocardinal infarction with complications and discharched alive	$ 11,332.64

Table 5: DRG Codes, Diagnosis and estimated reimbursement for 1997

Siverman and Skinner (2001) investigated whether for-profit hospitals are correlated with upcoding of patients with certain illnesses and they found that there is in fact a positive correlation. The incentives of hospitals (profit driven or not) is to maximize their Medicaid reimbursement and to assure a certain quality of care. If "a simple mistake" can lead without any effort to an increase in revenues the existing control system lacks accuracy. The key is to recognize the difference between a simple mistake and consistently upcoding the diagnosis with the intention to increase revenues.

2.6 Important Factors of the Current Fraud in Medicaid

According to the United States General Accounting Office and health insurance industry sources, between three percent and ten percent of any state's Medicaid budget is lost due to fraud and abuse. The National Health Care Anti-Fraud Association (NHCAA) estimated that ten percent of all healthcare claims contain some element of fraud (Frieden, 1991). An important factor to take into account when evaluating this statistic is the time frame in which this estimation was made: in the beginning of the nineties. The industry has been using this estimation of three to ten percent but Psaty et al (1999) and Hyman (2001) and Sparrow (2002) expressed their concerns about this figure. According to Sparrow (2002) due to the dynamic nature of fraud and a lack of proper measurement estimation vary between the 25 and 50%, a significantly higher estimation.

An extended research into the situation in Medicaid and Medicare is conducted by Sparrow (2002). Since it is considered as the first research that pointed out the big picture of U.S. health care fraud which has been extensively cited in the industry; an elaboration of the main malfunctions of the system are outlined in the following sections.

2.6.1 No Routine Systematic Measurement

An example of an effective systematic measurement is given by the Internal Revenue Service (IRS). In 1994 the IRS decided to systematically measure the fraud problem by reviewing 1,000 claims in the beginning of that year. The claims were verified by criminal investigators; they were required to validate the information provided by the taxpayer by checking with third parties. They would verify employment status with the employer and check with the neighbors if there was doubt about the children's residency. The investigators were requested to make their own best judgment about the situation and to classify any misrepresentation as intention or unintentional. The results suggested that roughly 4 billion dollars was incorrectly paid. During that year the IRS's regular detection system found only 160 million worth of refund fraud. If the IRS never started the systematically profound investigation they would underestimate the magnitude of the fraud problem by a factor of twenty. The detection systems showed the IRS only 5% of the real problem.

For health care fraud outlined in section 2.5 the Sparrow (2002) proposes to perform effective measurement in the form of a random audit with three criteria:

1. Selection of statistically valid random sample of claims
2. A thorough audit of every single claim
3. The audit should include external validation of the information within the claim rigorous enough to identify fraudulent claims

"To control fraud you must be able to see clearly" (Sparrow, 2002). Without systematic and proper measurement no one can tell whether an increased detection rate is a good or a bad sign. It could be that the fraud detection system has been improved; however it might just as easily imply that the overall amount of fraud increased. Therefore the three measurement improvements are proposed to create a baseline from where every year a proper analysis can be executed and compared with other years to structurally monitor the fraud and abuse in Medicaid.

2.6.2 The Dynamic Nature of the Fraud Detection Game

The fraud control game is a dynamic interaction between fraudsters and fraud detection teams. Fraudulent and legitimate behavior constantly changes for several reasons (Cahill *et al*, 2002). The former want to stay ahead of the fraud detection systems which means if a certain fraud scheme is discovered they will change their strategy to find other ways to defraud the system. Anderson *et al* (2006) provide a striking example involving bank account fraud detection. In France the magnetic strip on the debit card was replaced with an electronic chip to prevent fraud and abuse. Fraud initially fell but increased quite rapidly due to cross-country fraud. Fraudsters shifted their focus on using French cards in Germany and the United Kingdom because in those countries the chip was not implemented. Fraud detection is constantly evolving and a static set of filters and edits have only a limited short term utility. Maintaining an effective fraud control system requires continuous assessment of emerging fraud trends and constant revision of control tools (Sparrow, 2002). Legitimate behavior changes naturally overtime as well, therefore an effective detection technique (e.g. profiling) must change accordingly to capture the differences overtime in order to keep false alarms to a minimum.

2.6.3 The System is not Prepared for Fraud

According to Sparrow (2002) the fundamental fact is that the industry's standard detection and control systems are not aimed at detecting criminal fraud. The software audits and edits built into modern, highly automated claims processing systems were designed with honest providers in mind and serve the purpose of catching errors, verifying eligibility and checking if the price is within bounds. Automated prepayment edits and audits generally serve only to ensure that the claim is presented and processed correctly. As described in section 2.3 provides several examples of these automated edits and audits. Three important factors that are causing the health care fraud stated by Sparrow (2002):

I. Informing Providers Instead of Investigating

When a claim is declined by the Medicaid system a notification message is sent to the provider with the reason why the claim was rejected (Figure 4; section 2.3). There is no follow up action to verify the reason why this claim is submitted in the first place. As described (§2.6.2) fraudsters will learn from the feedback and try to defraud the system with a more informed and refined technique. Humans make mistakes and during submitting a claim mistakes are made too, but the claims submitted with the intention to defraud the system need to be detected and the providers need to be vetted.

II. No Random Routine External Validation of Claims Information

The standard economic model of crime and punishment indicates that health care providers who engage in fraud and abuse do so after rationally assessing the benefits and cost (Becker, 1968). The strength of the deterrence effect depends on the probability of being caught, the probability of being convicted once caught, and the seriousness of the punishment once convicted. All three are notoriously low in the U.S. health care programs (Sparrow, 2002). The Medicaid Fraud control units do not have the resources to randomly perform routine checks to (externally) validate randomly picked claims to create awareness by providers that every claim might be investigated.

III. Automated Processing

Sophisticated fraud schemes are applied by perpetrators who assume the appearance of transaction level fraud control filters who therefore design their fraud scheme so that each transaction comfortably fits the legitimate profile and is processed without any problems. The amount of these apparently legitimatly

submitted claims can be easily increased by using a software program (similar to email spam software) to generate multiple claims using a certain randomness or variation to avoid detection by the detection system (Hyman, 2001).

According to Sparrow (2002), due to the amount of annually submitted claims the basic failure of control in the Medicaid system is the failure to distinguish processing accuracy and claim verification. Processing accuracy is the amount of processed claims in a certain time frame and claim verification is the control mechanism to verify the legitimacy of the claim; is the claim fraudulent or abusive. Therefore humans are necessary to keep in the loop (Becker et al, 2010) and the primary focus should be payment accuracy; humans analyze and the system supports (Sparrow, 2002). From a cost perspective the truthfulness of a claim is more important to verify than the actual processing costs. The costs as a result of the fraud waste and abuse is estimated to be 1/3rd of the total health care costs in the United States; $700 billion dollar (Kelly, 2009). To illustrate the rising costs in the health care the increase in costs is shown in the growth in Medicaid long-term care expenditures.

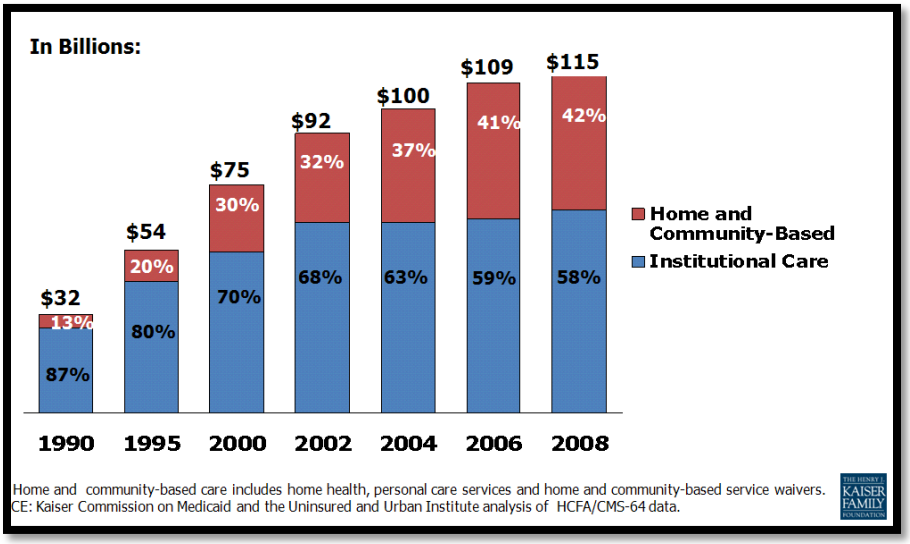


Figure 9: Growth in Medicaid Long-Term Care Expenditures, 1990 – 2008 (Kelly, 2009)

After 18 years the expenditures are almost 4 times as high and the demand of the long-term care has not grown that rapidly. Therefore the conclusion can be drawn that fraudsters consciously choose this discipline to defraud the Medicaid program. Another example of a vulnerable sector is the hospice care where people receive medical care for the last half a year of their life. Examples of beneficiaries who are enrolled for 5 years in a hospice care center are no exception (Kelly, 2009).

2.7 Conclusion

An important factor for fraud control is the big picture; this means collecting intelligence from different sources and systematically measuring and monitoring the shifting patterns of fraudulent behavior. Therefore, it is necessary to establish a situation in which the Medicaid fraud control units have the ability and tools to effectively verify the claims for fraudulent and abusive behavior. Measurement plays an important role in indicating the impact of the fraud problem and the progress of monitoring fraud and abuse detection in the U.S. Medicaid program.

The focus of the processing system must be on the verification of the claims instead of the processing accuracy. Concerning claim verification the costs of investigation should be taken into account because the benefits must at least equal the costs of investigation. To decrease the asymmetric information between Medicaid and the providers the benefits of the investments must outweigh the costs otherwise the costs will rise even more. Therefore a suspicion score might be an interesting tool to give the claimant an indication of the likelihood fraud occurred; suspicion scores will be discussed in detail in the next chapter.

The results of the conducted literature study is an overview of the existing fraud and abuse schemes in the U.S. health care programs and presented in the table 4. Fraud detection is a dynamic process and new fraud schemes will be developed since fraudsters will constantly try to avoid the safety mechanisms of the program. It is important for the fraud detection to keep up with the fraudsters in order to minimize the financial losses due to fraud and abuse and maintain the viability of the Medicaid program.

3. Fraud Detection Systems

'Fraud detection is more difficult than finding a needle in a haystack; there is only one needle that does not look like hay and the hay nor the needle do not change overtime'

- Michael Cahill (Author "Detecting fraud in the real world")

"Fraud management requires a holistic approach, blending tactical and strategic solutions with the state-of-the-art technology solutions and best practice in fraud strategy and operations" (Hand, 2010). This is in general a broader aspect which academics need to bear in mind, which perhaps have a natural tendency to focus on just one small part of the problem (Hand. 2010). This chapter discusses the importance of fraud detection systems and how other relevant industries are dealing with fraud and abuse. The first tow sections will elaborate on the structured literature review and previous work on fraud detection in the health care industry is presented. The systematic literature review forms the foundation of this researcha and subsequently the outcome of the literature review will be discussed in the next sections. The analysis starts with an outline of the various aspects of fraud detection (§3.3) and afterward the need to keep humans involved is highlighted (§3.4) and the importance of data is discussed (§3.5). Consequently a taxonomy of fraud and fraud detection techniques from relevant industries is presented (§3.6) and how the most relevant industries fight fraud is further outlined. Finally credit card fraud detection (§3.7) telecommunications fraud detection (§3.8) and the computer intrusion industry (§3.9) are discussed and the findings are summarized and presented in an overview (§3.10).

3.1 Systematic Literature Review

In this section a literature review is conducted and the method and details are outlined in this section. The foundation of this research is a systematic literature review of scientific literature. Libraries, academic journals and the World Wide Web are the sources of the literature review used to gather all of the relevant articles, journals, presentations, and white papers concerning electronic fraud detection. Access to the libraries of the University of Twente and the University of California San Diego were great sources for the background information on data mining and the use of statistical methods to detect fraud. In order to systematically review the scientific journals with a focus on the top 25 information systems journals (Schwartz and Russo, 2004), the following databases have been reviewed:

- Scientific Databases:
  - Web of Science
  - Scopus
  - PiCarta
- World Wide Web:
  - Google Scholar
  - Google

Using the scientific databases to search the top information systems journals the ranking of the search results was often determined by the number of citations. The most important or influential papers are identified by their journal ranking and citation analysis. To outline the systematic literature review the top down search methodology driven by keywords has been depicted in the following figure. On the upper

left side of the figure the keywords utilized are stated and the number of possible relevant articles is shown on the bottom of every step in the process.

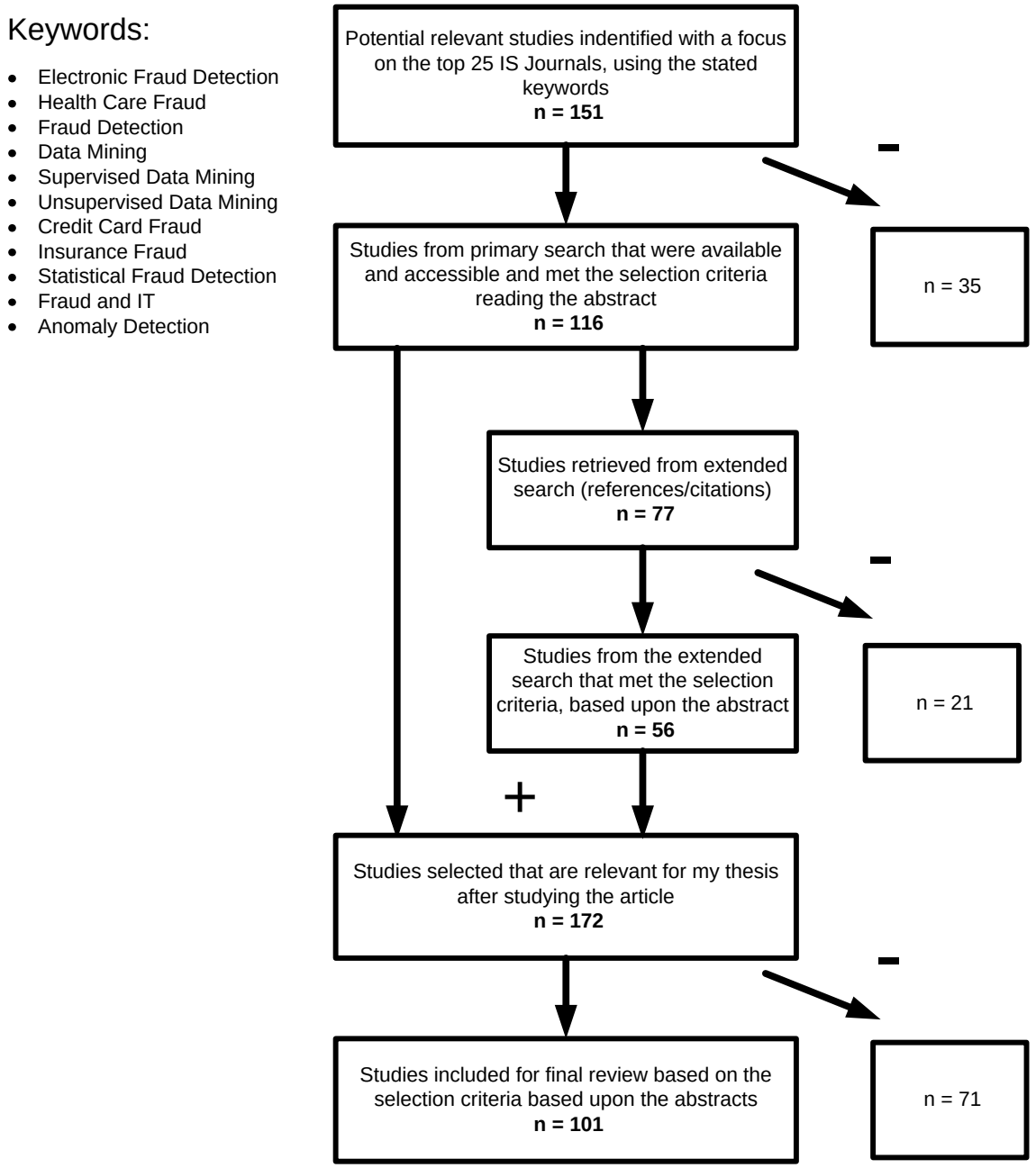


Figure 10: Systematic literature review

Figure 10 shows the several steps in the top down search driven by the keywords, as well as the bottom up search approach using forward and backward citation analysis. The selection criteria used to determine which articles would be included in this research were; relevance regarding fraud detection, year of publishing, relevance to health care insurance, and written in English. Exclusion criteria were; articles older than 15 years, and technical complexity. Articles with a technical and complex explanation of the algorithms of the data mining techniques were excluded because this report is focusing on the results of

these techniques regarding fraud detection. Finally, when access via the scientific databases was restricted Google Scholar and Google were utilized in an attempt to access the articles via the World Wide Web. If access proved to be restricted the articles were excluded from the systematic literature review.

3.2 Previous Work on Fraud Detection in the Health Care Industry

What state of the art articles are written about fraud detection in the health care industry and what research has already been done? To point out the gap in the scientific literature and to underpin the contribution of this research former relevant work in this field is outlined.

Major and Riedinger (1992) presented an important electronic fraud detection system in the health care industry in its time. However the research is quite outdated and since the time of writing the U.S. health care industry has become bigger and the data has increased exponentially. The foundation of their fraud detection system is 27 heuristics to support fraud analysts in the U.S. health care industry; however it is questionable if the system is robust enough to perform in the current situation. Furlan en Bajec (2008) propose a holistic approach to fraud management in the health insurance, they highlight six important aspects of fraud management: deterrence, prevention, detection, investigation, sanction, and monitoring. Regarding fraud detection they use unsupervised and semi-supervised techniques due to a lack of labeled data. The importance of visualization to recognize patterns and outliers is noteworthy, but the focus of the research is mainly on the next step after detecting fraud.

From a statistical perspective Sparrow (2002) points out the dangerous situation the Medicaid and Medicare programs are in and the corresponding costs for the U.S. government. However except for several established edits and prepayment checks electronic fraud detection is not discussed in detail. The information technology component is missing in this research; although it is a clear and proper analysis of the problematic situation facing the U.S. health care programs. The work of Hyman (2002) is in line with the analysis of Sparrow (2002) pointing out the difficulties of fraud detection and the problems associated with proving that fraud and abuse was committed. Hyman (2002) approaches the Medicaid and Medicare situations from a legal point of view and analyzes who is responsible for the fraud and abuse in the programs.

According to the conducted systematic literature review, the detection of fraud in the health care system by way of electronic fraud detection techniques is under researched. The worrisome situations of the U.S. health care programs have been pointed out in several papers. Furthermore fraud detection using data mining and statistical methods is extensively researched; a lot of information is available, especially fraud detection in the credit card and telecommunications industries. However, the application of electronic fraud detection techniques in order to combat health care fraud has not received enough attention. A possible reason might be the privacy information encompassed in the health care datasets, however given this opportunity in the scientific literature this research was initiated and conducted.

Finally, fraud detection is a not a discipline itself and is more of a combination of legal, information technology, and financial disciplines. Furthermore, fraud research is severely hampered by the limitation that results cannot be published and shared because this information will be used by fraudsters to learn how to avoid the detection systems. Applying electronic fraud detection techniques to data sets to gather empirical evidence is needed to verify the applicability and feasibility of the various techniques.

3.3 Prevention and Detection of Fraud

Fraud is a dishonest attempt to convince an innocent party that a legitimate transaction is occurring when in fact it is not. Detection is the ability to discover that a crime or violation, depending on the severity, has occurred. A detection system tries to identify patterns and trends in order to discover suspicious behavior (Sudjianto *et al.*, 2010). The development of new fraud detection models is made more difficult by the fact that the exchange of ideas in the arena of fraud detection is severely limited. Describing fraud detection models in great detail in the public domain would be imprudent and detrimental for the process of combating fraud; as it educates criminals and equips them with the information necessary to decrease their chances of being discovered. For that reason data sets are often not made available and results are often censored, making them difficult to access (Leonard, 1993).

In the sections 3.5 - 3.9 the tools and techniques available for statistical fraud detection are outlined and relevant industries as well as corresponding fraud detection techniques are discussed. There is a significant difference between prevention and detection of fraud. Fraud prevention describes measures to stop fraud from occurring (Bolton and Hand, 2002) while fraud detection involves identifying fraud as quickly as possible once it has been perpetrated (Bolton and Hand, 2010; Sparrow, 2002).

Fraud prevention (Bolton and Hand, 2002; Sparrow, 2002) tools include personal identification numbers (PIN) for bank cards, passwords on computers, watermarks, internet security systems for credit card transactions and subscribers identity module (SIM) cards for mobile phones. Preventive measures need to find a compromise between expense and inconvenience for the client on the one hand, and effectiveness on the other. *Fraud detection* (Bolton and Hand, 2010; Sparrow, 2002; Kou *et al*, 2004) involves identifying fraud as quickly as possible once it has been perpetrated. Fraud detection starts where fraud prevention has failed; although in practice fraud detection is a continuous process since one would always be unaware that fraud prevention has failed.

Sparrow (2002) suggests that the effectiveness of the fraud detection process depends on:

- The speed at which the crime is detected
- The amount of false alarms generated
- The range of crimes that can be detected

Typically a system will generate a suspicion score to indicate how likely it is that a certain transaction is criminal. Cases exceeding a certain suspicion score threshold will be further investigated.

Given these criteria the ideal system would detect the broadest range of fraud as soon as possible while generating minimal false alarms. In theory fraud can be reduced to as low a level as one likes; however it must be worth the effort and costs (Bolton and Hand, 2002). In practice some compromise has to be reached, it is often a commercial compromise between the costs of analyzing fraudulent cases and the savings made by detecting it. In the insurance industry the principal agent problem (Akerlof, 1970) has been extensively researched (§ 2.2). The presence of asymmetric information, where the principal (the insurer) does not have the same information about the activities of the agents (the clients), represents the need for data analysis and further investigation to gain more insight about possible fraudulent and abusive activities. However data analysis and further investigation cost money and resources and it must be worth the costs otherwise the process is not profitable. Therefore a suspicion score is a useful tool to indicate the probability that a certain alert is considered worthy to further investigate. Given the fact that it is

expensive to undertake a detailed investigation, efforts should be focused on the cases most likely to be fraudulent. As a result the recovery rate could be used as an indication of the result of the prevalence of fraud within the system. However, the fact remains that an increase in recovered funds could be due to either an improved fraud detection system or an increase in overall committed fraud and abuse.

A Comparison with the U.S. Customs Service

The best way to appreciate the fraud problem is to compare the situation with the U.S. Customs Service. The challenge for customs is to impose a minimum amount of inconvenience for people who are waiting in line for the import and export regulations, which constitutes are the majority of all the participants. Furthermore, it is quite important to deter people who commit the typical small violations, irrelevant whether it is consciously or not. Finally, it is very important to detect and handle appropriately the professional violators. In order to serve all of the stakeholders without significantly interrupting the daily processes it is important that the Customs Service is designed and organized in an optimally balanced way; not too strict but still fair. In this way severe fraud and abuse will be tackled and honest decent healthcare providers will not be bothered and accused of fraud. However this is an ideal situation and this would be extremely hard to establish in reality.

3.4 The Role of Humans

Fraud detection is not something that can be automatically pursued by simply investing more in a fraud detection system. Understanding of and familiarity with the data is an essential key to effective detection (Hand, 2010). Systems cannot understand the complexity and dynamic nature of fraud detection, therefore it is essential to keep humans involved to understand what the data means and if it should be flagged as fraudulent or not. Becker *et al* (2010) emphasizes the important role of humans and mentioned it as one of the key aspects in fraud detection; people have to verify fraud and implement corrective measures with the support of technology. Fraud detection systems do not understand the complex and changing real world or what the data means. Fraud detection software is not perfect and thus is complex and difficult to interpret; therefore it is a good thing to have humans in the loop (Becker *et al.*, 2010).

Important practical advice about the complexity of fraud is stated by Becker *et al.* (2010): “there is no sharp line between the intention not to pay and not the ability to pay”. In other words, there is a fine line between the severity of fraudulent behavior. Hand (2010) depicted this with a striking example; suppose someone is shopping for Christmas presents and he is getting over enthusiastic. Once he gets home he realizes he overspent with his credit card and now he is unable to pay for all of the purchases. He comes up with a devious plan and decides to report his credit card as stolen at the local police station. Note that the transactions were legitimate at the moment of purchasing, the transactions do not become fraudulent until after the card is reported stolen. As a result, identical data describes both legitimate and fraudulent behavior, and it is merely the action of reporting the card as stolen that determines the transaction to be fraudulent. From a data perspective this is an interesting example of the delicacy of data analysis and what defines legitimate and fraudulent behavior.

Due to the huge amount of data collected humans need fraud detection systems to support the data analysis process to scan for fraud and abuse. To create solid fraud detection systems humans and computers have to collaborate and work closely together. The next section elaborates on the second aspect of that collaboration.

3.5 The Importance of Data

Whereas the intentions of the providers cannot be observed, it is assumed that their intentions are reflected in the claim data (Hollmen, 2000). The claim data could subsequently be used in describing behavioral patterns of users. Statistics and probabilistic models (supervised data mining) could be employed in learning these usage patterns from claim data. In general these models are used either to detect abrupt changes in established usage patterns or to recognize typical usage patterns of fraud. These methods are shown to be effective in detecting fraudulent behavior by empirically testing these methods with data from real mobile communications networks (Beckers *et al.*¸ 2010).

3.5.1 Data Analysis and Timeliness

Statistics and machine learning are the foundation of most of the data mining technologies: e.g. regression analysis, clustering, and heuristics (such as applying and adjusting algorithms to improve results). Data mining and machine learning are often used interchangeably and data mining is an applied form of machine learning. *Statistics* are more theory and model based and *machine learning* makes use of heuristics with advanced statistical analysis and is focused on improving performance of learning agents (Tan *et al.* 2006). *Data mining* integrates theory and heuristics and is used to find previously hidden trends or patterns. "Data mining is the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases" (Hand *et al.*, 2001). Data mining programs are capable of detecting patterns and irregularities in the data. Strong patterns can be used to make non-trivial predictions about new data. The technical basis of data mining is algorithms that acquire structural descriptions from examples. The methods often originate from statistics, artificial intelligence and database research. Data mining is a multidisciplinary field which combines techniques and models from the other disciplines such as data mining, statistics, machine learning, visualization, information science, and database technology.

3.5.2 Statistical and Computational Challenges

The fraud detection domain has to deal with certain challenges surrounding the available data. Sudjianto *et al.* (2010) comprehend these fraud detection data challenges in a framework presented in this section.

I. Volume and Complexity of Data

Volume results often in extremely large databases and due to the complexity of the data researchers repeatedly have to dig deeper into the data in order to truly understand the meaning of a record. As result of the size, volume and complexity Sudjianto *et al.* (2010) realized a tight integration of data analysis and data management of the data sources at AT&T. Once the call database is established; it serves many needs beyond providing information for the fraud detection system. Retrospective analysis can be performed and new algorithms can be tested on the historic data. It also provides an historic view of calling behavior and individual calling patterns (Becker *et al.*, 2010).

As discussed in the papers of Beckers *et al.* (2010) and Sudjianto (2010) transaction fraud detection typically involves very large datasets. The former pointed out that the AT&T data warehouse was identified as the world's largest in 2005. Many fraud detection problems involve huge data sets that are constantly evolving. For example the Royal Bank of Scotland, which has the largest credit card merchant acquiring business in Europe, carries over a billion transactions a year while AT&T carries around 275 million calls each weekday (Cortes and Pregibon, 1998). The credit card company Barclay Card carries approximately 350 million transactions a year in the United Kingdom alone (Hand, Blunt, Kelly and

Adams, 2000). In 2005 the estimated credit card fraud rate was 0.07% (Nilson report, 2006) and although it is quite a low percentage, with 350 million transactions the losses are quite severe. To process these transactions in a search for fraudulent transactions the use of statistical models is not sufficient, data mining techniques with fast and efficient algorithms are required. The endless line of transactions never ceases; "It simply keeps coming like water from a hose" (Tasoulis *et al.,* 2008). Due to the large volumes in the credit card fraud detection often decisions need to be made in real time; it is only useful if the fraudulent transaction can be identified and stopped immediately. Algorithms aimed at real time detection will often need to make a decision solely on the data present in the current transaction. "A credit card fraud detection system with a perfectly set up classification system that classifies every fraudulent transactions would be useless if it would take a couple of months to set up" (Hand, 2010). In terms of volume, large data streams bring greatly increased complexity.

II. Class Imbalance

Financial (Bolton and Hand, 2002), telecommunications (Fawcett and Provost, 1997), intrusion and insurance (Derrig, 2002) crimes are rare events and the result is severe class-imbalance; compared to the number of legitimate transactions the criminal transactions constitute a very small percentage (Phua *et al.,* 2004; Vinciotti and Hand, 2003). Unbalanced class sizes impose fundamental limits to the classifier performance since the significant difference of legitimate and fraudulent transactions in the training group (Sudjianto *et al.,* 2010).

III. Population Drift

One of the main goals of a fraud detection system is to identify general patterns of suspicious behavior. In general people and customers will change behaviors due to changing competition or economic developments. Furthermore fraudsters try to bypass existing detection systems so that their behavior is dynamic and continuously evolving over time (Hand, 2006). Fraud and fraud detection is a game of hide-and-seek with the players are always changing their strategy to keep one step ahead of the competition. Fraud detection models need to be continuously validated to accommodate these changing patterns and distributions (Sudjianto *et al.,* 2010).

IV. Class Overlap

Criminals often try to make their transactions look as normal as possible in order to slip past detection systems. Therefore, it is hard to define which transaction is suspicious (e.g. section 2.5.1; superimposed fraud in the telecommunications industry). Two transactions with the same characteristics, one is fraudulent and one is legitimate, resulting in a substantial overlap between the fraudulent and non fraudulent case. In section 3.4 an example is given about a person who went shopping and afterward realizes that he overspent his credit card and decides to report the credit card as stolen. The fact that the credit card is reported as stolen implies that the legitimate transactions suddenly became fraudulent. From a data perspective nothing has changed so the consequence is that similar data can be classified in different ways (e.g. fraudulent or legitimate class).

V. Class Mislabeling

It is not always feasible to investigate each case that was detected as fraudulent. This poses an important need for machine learning, as detection algorithms will be trained to uncover possibly mislabeled data and motivate the need for robust methods that can handle mislabeled data (Sudjianto, 2010).

Becker *et al.* (2010) mention in their research two important aspects that support the solution for the five computational challenges (Sudjianto *et al.*2010):

- The need to join data analysis and data management
- Need for fast feedback loops

To be able to analyze if the data is labeled correctly and to apply adjustments quickly there is a justifiable need to join data management and analysis. To prevent misclassification, class overlap, and class imbalance the data analysis and management must communicate well. If the two are joined fast feedback loops communication would be even easier. To keep up with the population drift and to reduce the erroneous classifications as a result of misclassification, class overlap or class mislabeling fast feedback loops and joined data analysis as well as data management are required to achieve improved results.

3.6 Fraud Detection Techniques in Related Industries

Kou *et al.* (2004) conducted a survey of fraud detection techniques used in the credit card fraud detection, telecommunications fraud detection and computer intrusion detection industries. Recently, a taxonomy of the fraud detection techniques was presented by Laleh and Azgomi (2009) including an overview of the types of fraud in the related industries.

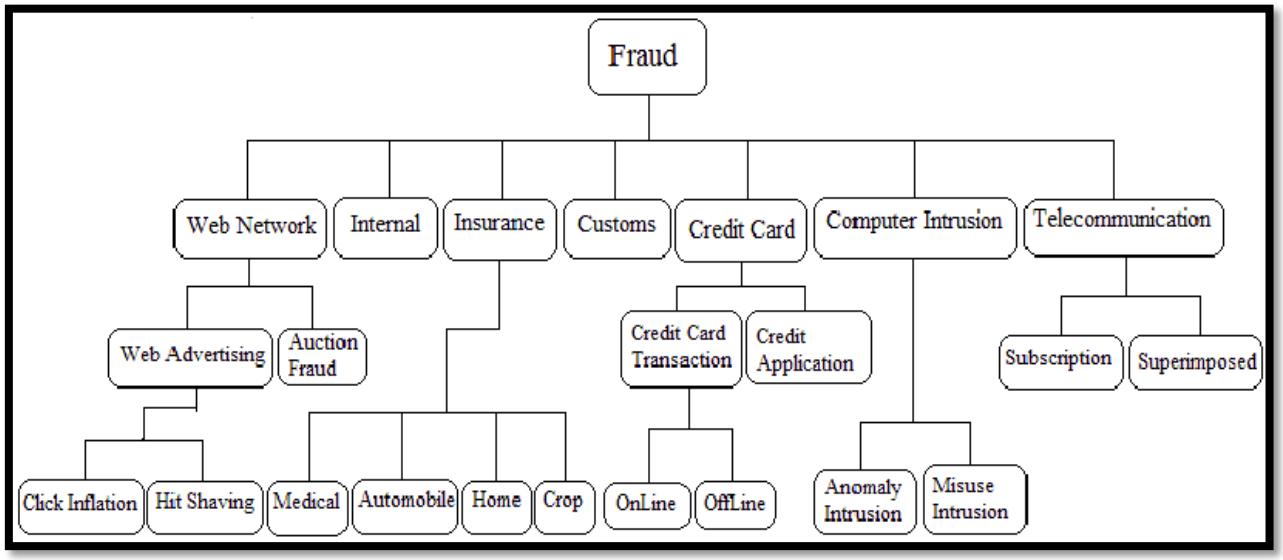


Figure 11: Taxonomy of Fraud Types (Laleh and Azgomi, 2009)

In the next section the fraud detection techniques applied in several industries are highlighted and discussed. The scientific literature review that has been conducted for this research provides an overview of the relevant papers and the applied fraud detection techniques. The next table is an overview of the fraud detection techniques and subsequently the presentation of the relevant papers.

Fraud Detection Type	Method
A	Supervised Classification Techniques
- **A1**	Linear Discrimination
- **A2**	Support Vector Machines
- **A3**	Neural Networks
- **A4**	Random Forests
B	Unsupervised Data Mining Techniques
- **B1**	Anomaly Detection
- **B2**	Cluster Analysis
- **B3**	Peer Group Analysis
C	Statistical Methods
- **C1**	Visualization
- **C2**	Profiling
- **C3**	Benford's Law
D	Rule Based
- **D1**	Online Analytical Processing
- **D2**	SQL Queries

Table 6: Overview Fraud Detection Techniques

Extensive research has been conducted in several industries about fraud detection systems and the data analysis by using data mining, statistics and rule based approaches to support the fraud detection process. An overview is provided in the following table of the conducted literature study to categorize the applied fraud detection techniques.

Title Paper	Industry	Objective paper	Type of Fraud	Technique	Results/Findings/Problems
Statistical Methods for fighting Financial Crimes (Sudjianto *et al.*, 2010)	Financial	To provide a survey of statistical techniques and data mining.	Money Laundering, Retail banking fraud	A B1 B2 C2	To provide an overview of financial fraud.
Fraud detection in the telecommunications: History and Lessons learned (Becker *et al.*, 2010)	Telecom-munications	To discuss major fraud schemes and fraud detection techniques used to address them.	Subscription and Superimposed fraud (both telecom)	C1 C2 D	Use simple understandable models, heavy use of visualization, involvement of humans.
A Comprehensive Survey of Data Mining-based Fraud Detection Research (Phua *et al.*, 2005)	General	To define existing challenges in the fraud detection domain for diff types of large data sets.	Insurance fraud, Credit card fraud, Tele-communications	A B2 B3 C2	Overview of Supervised, semi-supervised and unsupervised techniques.
Novel Techniques for Fraud Detection in mobile telecommunications Networks (Moreau *et al.*, 1997)	Telecom-munications	To explore the detection of fraudulent behavior based upon a combination of absolute and differential behavior.	Telecom fraud	A C2 D	Obtaining a significant amount of fraudulent data and labeling it as such is a significant effort and often a problem.
EFD: A Hybrid Knowledge /Statistical based System for the Detection of Fraud (Major and Riedinger, 1992)	Health care insurance	Electronic fraud detection.	Medical insurance fraud	C D	With the applied set of heuristics true positive rates are approximately 50%.
A Taxonomy of Frauds and Fraud Detection Techniques (Laleh and Azgomi, 2009)	General	Provide a taxonomy of (new) frauds and fraud detection techniques.	Internal, customs, insurance, credit card, computer, telecommunication	High level overview of A & B	The result is an overview of several types of fraud and different fraud detection techniques on a high level. High-level overview of (un)supervised and semi-supervised techniques.
Survey of Fraud Detection Techniques (Kou *et al.*, 2004)	Credit card Computer Intrusion Telecom-munications	To provide a comprehensive review of different fraud detection techniques.	Credit card Computer intrusion Tele-communications	A3 B1 C1 D	Neural network is an important tool however difficult to implement due to a lack of data. Profiling to detect fraud from call pattern is effective.

Title Paper	Industry	Objective paper	Type of Fraud	Method	Results/Findings/Problems
Adaptive Fraud detection (Fawcett and Provost, 1997)	Telecom-munications	To describe a design of user profiling methods.	Superimposed fraud (Telecom)	C2 D	Fraud detection systems must be adaptive and people must determine (trial-and-error) how to profile and which rules are effective.
Unsupervised Profiling Methods for fraud detection (Bolton and Hand, 2001)	Credit Card	To apply unsupervised techniques because labeled data is not always available.	Credit card transaction fraud	B3 C2	Both analysis and visualization have the ability to detect anomalies and detect changes in spending trends.
Statistical Fraud Detection: A Review (Bolton and Hand, 2002)	Financial Computer intrusion Telecom-munications	To describe the statistical tools available in the different areas.	Credit Card Money laundering Computer intrusion Telecom	A B1 C1 C2 D	The speed of detection is important so the time of detection should be measured. How effective a statistical tool is depends on the type of problem.
Neural Fraud Detection in Credit Card Operations (Dorronsoro *et al.*, 1997)	Credit card	To present an applied on-line fraud detection system (Minerva).	Credit card transactions fraud	A3	Positive result; it detects 40% of all the fraudulent transactions and, can be used as a basis for other models.
Establishing Fraud Detection Patterns Based on Signatures (Ferreira *et al.*, 2006)	Telecom-munications	To detect deviate behaviors within a useful time span	Superimposed Fraud (Telecom)	B1 C2	The anomaly detection with the signature as a basis supports the detection of telecom fraud.
Data mining for credit card fraud: a comparative study (Bhattacharyya *et al.*, 2010)	Credit card	To evaluation Random Forests and Support Vector Machines.	Credit card fraud	A2 A4	Random forests based methods are able to obtain good (the best of the 3) overall performances.
The application of data mining techniques in financial fraud detection (Ngai *et al.*, 2010)	Financial	To review data mining techniques in order to discover financial fraud.	Financial and insurance fraud	A B1 B2 C1	A review of 49 articles to categorize financial fraud and an overview of applicable data mining techniques.

Table 7: Overview of Fraud Detection Papers

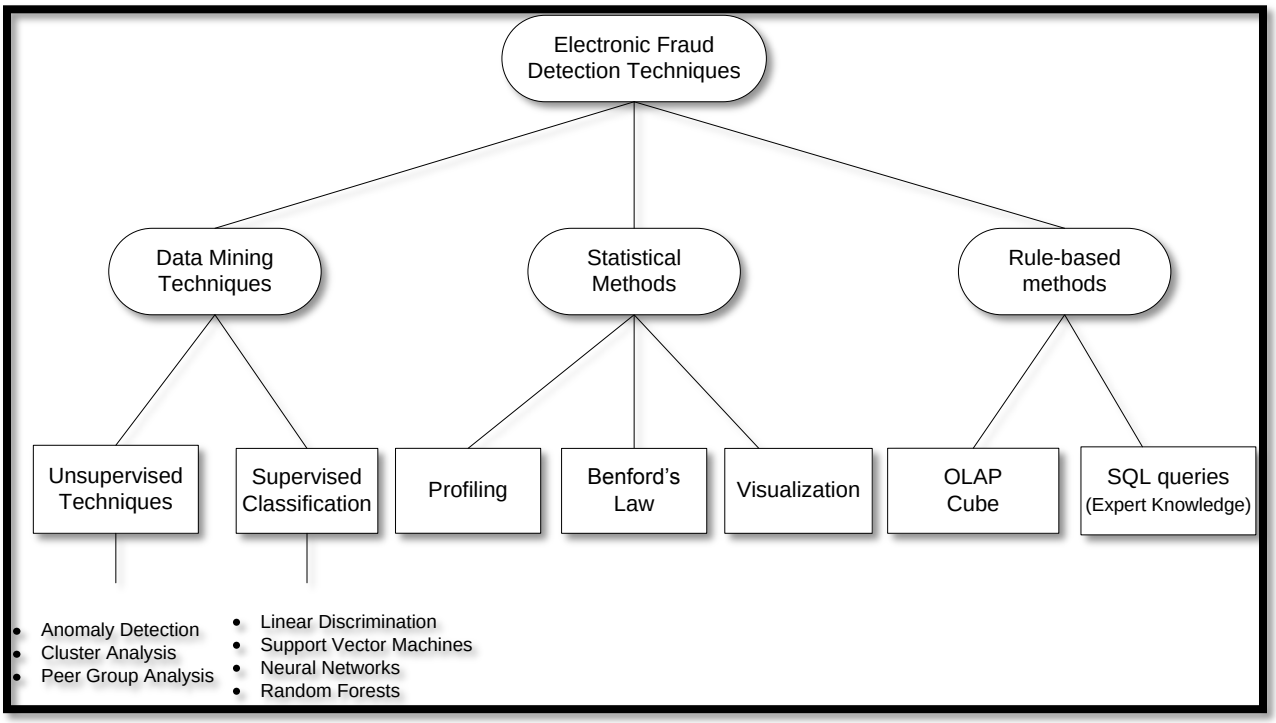


Figure 12: Framework of Electronic Fraud Detection Techniques

3.6.1 Supervised Data Mining Technique: Classification

Data mining is defined as a process of identifying interesting (possibly suspicious) patterns in databases which can be used for decision making (Ngai et al, 2010). Zhang and Zhou (2004) state that classification and prediction is the process of identifying a set of common features and models that describe and distinguish data classes or concepts. Supervised learning methods are trained to discriminate between fraudulent and legitimate behavior so that new observations can be assigned to a class as to optimize some measure of classification performance (Bolton and Hand, 2001). Supervised techniques use prior information on class membership and similar fraud instances from both classes are needed as a learning set (Vinciotti and Hand, 2003).

Classification models are supervised methods that describe and distinguish classes or concepts for future predictions. It is the discovery of a predictive learning function that classifies a data instance into one of several predefined classes. Classification can be done in several ways; naïve Bayes classifiers, neural networks, logistic regression models, decision trees and support vector machines are all alternative ways of representing conditional probability distribution (Witten and Frank, 2005). The representation power differs per technique, for example naïve Bayes and logistic regression models can only represent simple distributions whereas decision trees can represent arbitrary distributions. Fadlalla and Lin (2001) compared several statistical models, namely linear and multiple regression analysis, with neural networks in the financial industry. The result of the comparative study shows that generally neural networks outperform statistical-econometric models. Neural networks are not guaranteed to outperform under all circumstances but if there is enough data to train the network to learn and validate, neural networks become a good contending model. According to Caruana and Niculescu-Mizil (2006) achieve learning methods such as boosting, random forests, and SVMs excellent performance". Based on these analyses the conclusion can be made that on average the memory-based learning, single decision trees, logistic

regression and naïve Bayes are not competitive with the best models. The most relevant and recent techniques of classification are outlined in section 3.5 to provide an overview of the different classification techniques.

The problem with classification in fraud detection is that these methods suffer from unbalanced class sizes. Criminal or fraudulent transactions are rare and consequently legitimate transactions greatly outnumber the fraudulent transactions. As a consequence the frame of reference frame for the fraudulent class is small which results in mislabeling and misclassification. "Ultimately severe class imbalance imposes fundamental limits to the classifier performance" (Sudjianto et al., 2010). Supervised learning methods make use of training sets wherefrom the features of the different classes are learned. Training data often represents a sample that is to small to generalize from and accurately make predictions about unseen data (Vapnik, 2000). It is more useful for a classifier to learn a decision boundary that provides the largest separation between the classes. As a result of an analysis of the available and applicable techniques in the fraud detection domain an overview is provided of effective classification techniques. In the next section several learning methods are presented and described in more detail:

1) Support Vector Machines (SVM)

Support Vector Machines is a kernel method where the kernel represents the similarity between two objects and is represented in the form of a dot in a vector space. SVM selects a small number of critical boundary instances, called support vectors, from each class and uses these support vectors to build a linear discriminant function that separates the two classes as widely as possible. SVM do well in classifying non-linear separable groups; they do not require a large training datasets and training converges to a unique global solution (Sudjianto *et al.*, 2010).

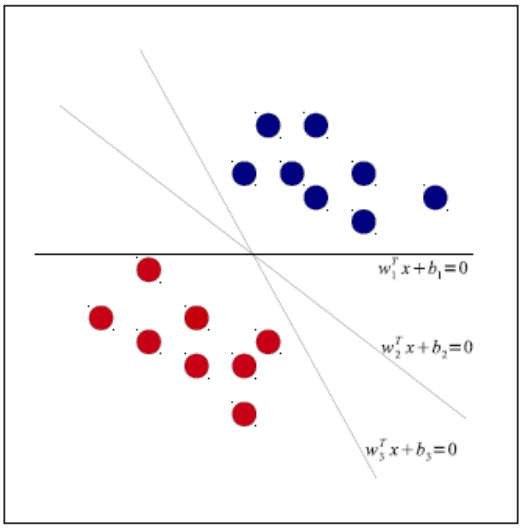


Figure 13 Seperating Hyperplanes (Hamilton, 2010)

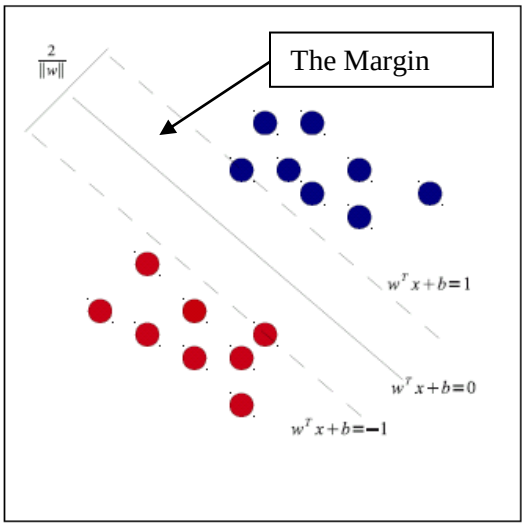


Figure 14 Maximizing margin principal illustrated (Hamilton, 2010)

Interest in neural networks appears to have declined since the arrival of support vector machines, perhaps because the latter generally requires fewer parameters to achieve the same accuracy (Witten and Frank, 2005). "As compared with techniques like neural networks which are prone to local minima, overfitting and noise, SVMs can obtain global solutions with good generalization errors" (Bhattacharyya *et al.*, 2010). Illustrated in figure 14 and 15 (Hamilton, 2010) is how SVM hyperplanes separate the linearly separable data. "SVM's work with a larger, transformed version of the feature space and find a maximum margin hyperplane that separates two classes of data" (Vapnik, 2000). Hamilton (2010) shows in figure 15 that the margin is defined as the distance between the hyperplanes.

2) Artificial Neural Networks

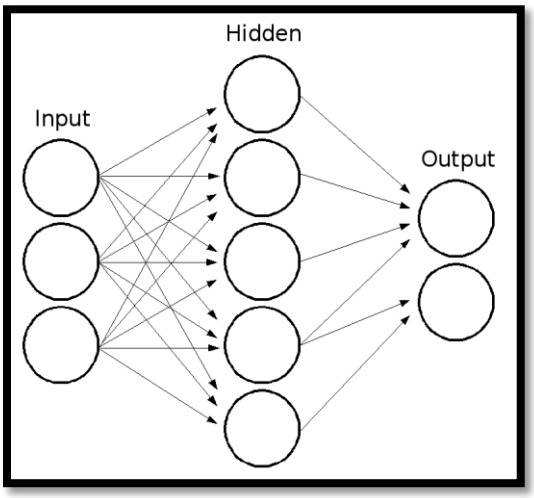


Figure 15: Artificial Neural Network (Kou *et al.*, 2004)

"The neural network is a set of interconnected nodes designed to imitate the functioning of the human brain" (Kou *et al.*, 2004). As shown in the figure artificial neural networks are made connecting artificial neurons; it can be thought of as a black box where both the input and the output variables are known. The advantages according to Ngai *et al.* (2010) are that; it is an adaptive technique, it can generate robust models, and that; the classification process can be modified if new training weights are set. "Neural networks are chiefly applied in the credit card, automobile insurance, and corporate fraud industries" (Ngai, *et al.*, 2010). Neural networks have proven to be commercially interesting due to their success in the credit card industry in the 1990s (Bhattacharyya *et al*, 2010; Sohl and Venkatachalam, 1995). In the nineties Dorronsoro *et al.* (1997) implemented a real time neural network fraud detection system called "Minerva". In this research it is shown that by applying "Minerva" neural networks are technically feasible and highly intriguing from an economical point of view. According to Maes *et al.* (2002) it is often assumed that neural networks are a fast, easy, and reliable technique capable of obtaining good results in different applications. In practice, however, it is found that applying neural networks leads to great difficulties.

Neural nets are more commonly used for predicting numeric quantities, however the techniques suffers from the disadvantage that the structures they produce are opaque and cannot be used to help us understand the nature of the solution (Witten and Frank, 2005). However the popularity in the credit card industry during the 1990s is an indication that under certain circumstances neural networks can contribute positively to detecting fraud. The circumstances in the credit card industry are 1) real time data and, 2) labeled data to determine which transactions in the past were fraudulent and to train the technique to keep up-to-date with the latest fraud schemes.

Neural Networks have the ability to deal with highly skewed data and the great accuracy makes them suitable for multi-class classification. The downside is that neural networks are like a black box and the result are hard to interpret. Furthermore the training time for a large data set is quite expensive and they are prone to over fitting. Maes, Tuyls and Vanschoenswinkel (2002) reported that the performance of the classifier is very sensitive to the vector features chosen. Removing one of the correlated features had a big impact on the outcomes of the network.

3) Decision Trees and Random Forests

Single decision tree models are quite popular in the data mining application for their simplicity and ease of use (Bhattacharyya *et al.*, 2010). To create a decision tree, an attribute needs to be placed at the root node while each branch represents a possible value. The various branches split up the example set into subsets, one for every value of the attribute. Now the process can be repeated for each branch of the tree as long as the instances actually reach the branch. Trees with tests involving more than one attribute are called multivariate decision trees like CART; classification and regression trees. They are often more accurate and smaller than univariate trees but take much longer to generate and are also more difficult to interpret.

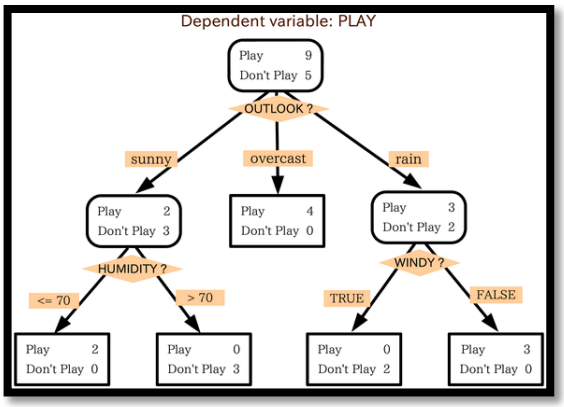


Figure 17: Example Classification Tree (Witten and Frank, 2005)

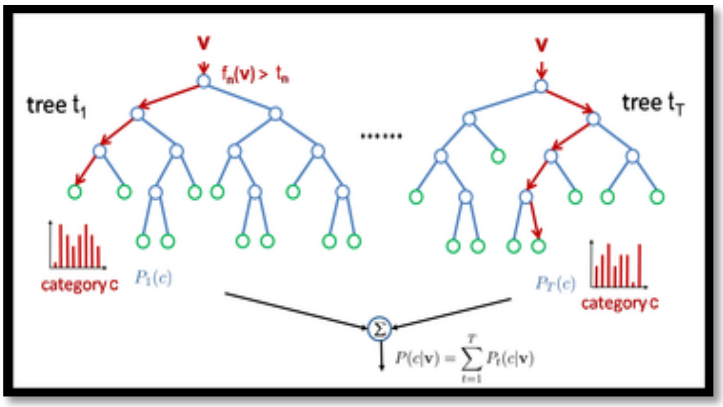


Figure 16: Example Random Forest (Whitrow et al., 2009)

A related and more sophisticated data mining technique which has been noted in recent years to show superior performance across different applications is Random Forest (Whitrow *et al.*, 2009; Bhattacharyya *et al.*, 2010). Random forests is an ensemble model consisting of multiple (sometimes up to 200) decision trees. They are computational efficient and robust to noise and it has been shown that random forests obtain good overall performance scores using a variety of performance measures (Bhattacharyya *et al.*, 2010).

4) Logistic Regression

Logistic regression is widely applied to classify the binary fraud problem. It is a relatively simple and well understood technique which performs well (Bhattacharyya *et al.*, 2010). Over the years logistic regression has been a standard technique in many real-life data mining applications (e.g. credit card fraud detection). The results are easier to understand than SVMs or Random Forests since simple techniques are easier to generalize when applied to new data sets. Accordingly, Bhattacharyya logistic regression performs consistently across different training data which indicate that the model yields similar rankings.

3.6.2 Unsupervised Data Mining Techniques

Techniques employed here are usually a combination of profiling and anomaly detection (Bolton and Hand, 2002). In unsupervised methods, there are no prior class labels of legitimate or fraudulent behavior (Laleh and Azgomi, 2009). An advantage of using unsupervised methods over supervised methods is that new fraud patterns and previously undiscovered fraud may be detected. Unsupervised models look for transactions, behavior, accounts, or customers that are behaving in an unusual or exceptional manner (Bolton and Hand, 2001). A baseline distribution of normal behavior is modeled to detect observations

that lie an abnormal distance from this norm. These observations can be examined more closely to verify the origin why it is considered as an outlier. Outlier detection is a basic form of an unsupervised method that can be used for fraud detection. A fundamental point is that by statistical analysis alone you can never be certain that an outlier or anomaly is fraudulent. The objective is to return a certain suspicion score; where a high number is more unusual or more likely to be fraudulent compared to historic observations. The idea behind a suspicion score from unsupervised methods is that unusual transactions or behavior can often be an indicator of fraud.

Clustering Analysis

Clustering is used to divide objects into conceptually meaningful groups (clusters), with the objects in a group being similar to one another but very dissimilar to the objects in other groups (Ngai et al., 2010). According to Tan *et al.* (2010) clustering is also known as data segmentation or partitioning and is regarded as a variant of unsupervised classification. In the telecommunication industry profiling is extensively applied and this is described in detail in section 3.8.

Anomaly Detection

Anomaly Detection or Outlier Detection is an outlying observation, or outlier, that appears to deviate markedly from other members of the sample in which it occurs" (Grubs, 1969). A further definition of outliers is from Barnett & Lewis (1994): An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data. Similar definitions are found for anomalies and anomaly detection. "Anomalies are patterns in data extensive use in wide variety of application such as fraud detection for credit card, that do not conform to a well defined notion of normal behavior" (Chandola *et al.*, 2009). *Anomaly Detection* is applied in the intrusion detection for cyber-security and military surveillance for enemy activity and is similar to outlier detection. According to Chandola *et al* (2009) an important aspect of anomaly detection is the nature of the anomaly, which can be classified into different categories:

- Point anomalies – if an individual data instance can be considered anomalous compared to the rest of the data, then the instance is termed as a point anomaly. For example, consider credit card detection; for simplicity the data is defined using only one feature: *amount spent.* A transaction for which the amount spent is very high compared to the normal range of expenditure of that person will be a point anomaly.
- Contextual anomalies – if a data instance is anomalous in a certain context, then it is termed as a contextual anomaly, also referred to as *conditional anomalies* (Song *et al.* 2007). A simple example of a contextual anomaly from the credit card fraud detection domain is the *time* of a certain purchase. During Christmas the credit card purchases are in general higher than in the middle of June for example. If an individual usually purchases on average $100 per week but during Christmas it will reach $1000. A new purchase of $1000 in the middle of June would be considered as a contextual anomaly because it does not conform to the expected spending behavior. However, a peak of $1000 around Christmas time will be considered normal spending behavior.
- Collective anomalies – if a collection of related data instances is anomalous with respect to the entire data set, it is termed as collective anomaly. It should be noted that the collection of events is an anomaly but the individual events are not anomalous. For example, a single transaction of

$550 is not anomalous, however 30 transactions (in a row) of exactly $550 are anomalous and likely to be fraudulent.

Anomaly detection approaches attempt to identify abnormal behavior in patterns and can make use of supervised or unsupervised methods to detect the anomalies. To gain further insights in the anomaly detection technique the computer intrusion detection systems have been analyzed. Attempts of hackers who unauthorized try to access information on computer systems cause unusual patterns which can be recognized with anomaly detection; this is highlighted in section 3.9.

Profiling

Profiling is an established method in the credit card and telecommunication fraud detection fields. User profiling is the process of modeling the characteristic aspects of the user and it often works well with low-usage subscribers because unusual behavior is very prominent (Fawcett and Provost, 1997). The digital signature is defined as a nonparametric estimate of the distribution. Profiles must be updated to reflect the dynamic patterns of criminal activity as well as changes in legitimate user behavior (Sudjianto *et al.*, 2010). Unsupervised profiling is concerned with detecting behavioral fraud through the analysis of longitudinal information. This information typically changes overtime so it is important that the profiling technique is flexible enough to capture these changes and detect changes in behavior.

Due to the dynamic nature of the fraud detection game the importance of flexibility in the detection system is highlighted in the inadequacy of large models which have a hard time adapting. There is a need to be flexible and a modular approach of several different analytical techniques at the transactional and the account/provider levels. According to Sparrow (2002) an ideal detection system should work on data at three levels:

- Transaction level –information about the transactions such as amount, date, time, and location.
- Account level – the history of the account, such as the balance.
- Customer level – credit score, stated income.

Even when real-time decision making is not required, transaction levels, account levels and customer levels are high-dimensional data which differ and change over time (Sudjianto, 2010). It is necessary to extract features that summarize a customer's transaction history, such as average transaction amounts, the average number of transaction in a certain period, the total amount of transaction etcetera. Post payment controls are important to detect certain fraud types so that in the future these schemes can be detected sooner and proper adjustments to the fraud detection system can be made.

3.6.3 Rule Based Techniques and Statistical Methods

Rule based systems have been developed based on the experience of field experts (Bolton and Hand, 2002). An example of a rule based method in the financial industry: if an account receives a large deposit (> $10,000) from country X and is subsequently followed by a withdrawal of the same amount, then this accounts will get flagged and investigated further. This type of money laundering is known by the field experts and a sequence of transactions is known to be more likely to be criminal. Another example of a rule detected by financial fraud detection experts: in the credit card industry the likelihood of fraudulent transactions is significantly higher when first a small amount is spend at a gas station, and subsequently a large amount is spend on electronic goods (or other products which are easy to sell like jewelry) in one or

more transactions. This must be reflected in an increased suspicion score and before the electronic goods are bought the credit card should be blocked to prevent financial losses.

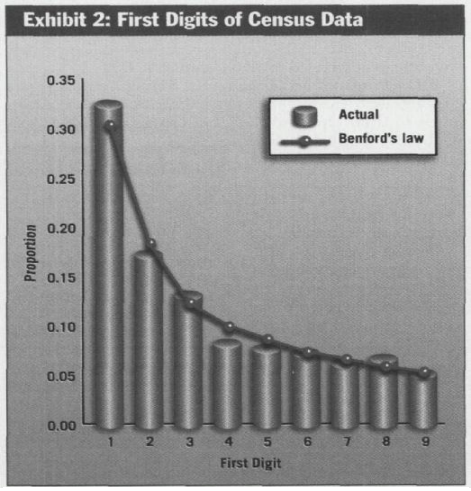


Figure 18: Benford's law Applied on the First Digits of Census Data (Hill, 1995)

As introduced in section 3.6.2, unsupervised techniques are often a combination between profiling and anomaly detection. Profiling is considered a statistical technique used to provide an overview or summary of behavior. Sudden changes in behavior can be identified when an up to date profile is available to compare the customer's unexpected behavior against their historic behavioral patterns. It is a widely applied tool in the telecommunications (Fawcett and Provost, 1999; Ferreira, 2006; and Becker *et al.*, 2010) and credit card fraud detection domains (Bolton and Hand, 2001; Sudjianto *et al.*, 2010). An interesting statistical tool originated from the accountancy field and is called "Benford's law" (Hill, 1995). It says that the distribution of the first digits of numbers of random distributions will always have a predictable form, as presented in figure 15. As shown, the digit 1 appears significantly more often than the digit 9 in the Census data of 1999. Nigrini (1999) showed that Benford's law could be applied by Certified Public Accountants as a data analysis tool to spot irregularities indicating possible error, fraud or manipulative bias. If the actual dollar amount exceeds the line that plots Benford's law the suspicion score that the numbers are made up increased. According to Nigrini (1999) data analysts require some professional judgment to identify which anomalies are worth investigating. It is an interesting tool to further investigate in the fraud detection application.

Online Analytical Processing Cubes (OLAP Cubes)

An Online Analytical Processing Cube is a data structure that allows fast analysis of data (Codd *et al.*, 1993). It makes use of a multidimensional set of data for dynamic analysis. Furthermore, it is a tool that provides visualization by generating, and analyzing reports as well as graphs to capture the behavior of providers and beneficiaries. It is considered to be a tool able to recognize patterns, increase analytic capability, and find patterns of behavior while at the same time easily presenting and summarizing data. OLAP Cubes are able to apply multiple dimension analysis, but how would analysts ideally want to apply this tool? Some examples of the opportunities to utilize OLAP Cubes in Medicaid became clear after interviewing several business intelligence experts:

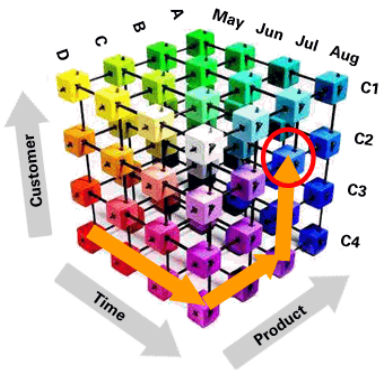


Figure 19: Example OLAP Cube (Relativity, 2010)

- The top 10 beneficiaries in a region per year
- The top 10 providers in a region per year
- $ paid per zip code per week per state
- $ paid per drug code per region/state per month
- Amount of claims per drug code per recipient

Visualization

Visualization refers to the easily understandable presentation of data and to the methodology that converts complicated data characteristics into clear patterns, allowing users to view the complex patterns or relationships uncovered in the data mining process (Turban *et al.*, 2010). Visualization is a clear presentation of complex patterns through visual characteristics such as color, position and size. According to Ngai et al (2010) visualization is best used to represent complex patterns in data.

3.7 Credit Card Fraud Detection

Credit card fraud falls broadly into two categories; behavioral fraud and application fraud (Bolton and Hand, 2001; Laleh and Azgomi, 2009). The latter type is comparable with subscription fraud found in the telecommunications industry and the "hit-and -run" strategy in the Medical insurance domain section 2.4.4. The credit card is obtained by using false identification documents and the card is used until the company blocks it. Behavioral fraud is comparable with superimposed fraud found in the telecommunications industry and the "steal a little all the time" strategy described in section 2.4.4. In this the details of legitimate credit cards have been obtained and purchases are made without the presence of the legitimate cardholder. If the cardholder does not notice the fraudulent activity on his card then the fraudulent activities can continue for a longer period without someone noticing.

Credit card fraud detection has two peculiar characteristics (Dorronsoro *et al.*, 1997). First the time span in which the system has to approve or reject a transaction is very limited. Second, there is a huge amount of credit card transactions that have to be processed at any given time. Mining such massive amounts data requires efficient techniques that are scalable (Chan and Fan, 1999). An important factor in fraud detection is the highly skewed data; there are many more legitimate transactions than fraudulent transactions in the credit card industry, and only a small number of all transactions are fraudulent. Sudjianto *et al.* (2010) conducted research to test the applicability of statistical methods in the financial industry. Especially in the credit card industry there has been extensive research done to combat fraud with a data driven approach since the quantity of data is so enormous.

3.7.1 Supervised Classification Techniques

Supervised classification has been extensively applied to credit card fraud detection (Bhattacharyya *et al.*, 2010). There are several techniques which effectively classify and detect fraudulent transactions at the moment of processing. Timing is essential in the credit card industry and therefore the classification techniques need to operate within the available real time data to determine if a transaction is fraudulent or not. In this section several classification methods applied in the credit card industry are highlighted to provide insight into the opportunities and limitations of the technique.

Support Vector Machines

As described in section 3.6.1 Support Vector Machines do well in classifying non-linear separable groups; they do not require a large training datasets and training converges to a unique global solution (Sudjianto *et al.*, 2010). These characteristics make SVM's attractive solutions to problems such as credit card application fraud, particularly because millions of transactions per day that need to be processed, of which only a limited amount are fraudulent making it a problem of finding a needle in a gigantic haystack.

Neural Networks

Neural networks are widely used for detecting fraud in the credit card industry. One of the major reasons for this is their ability to deal with the highly skewed class distributions that arise in this application (Sudjianto *et al.*, 2010). Dorronsoro *et al.* (1997) reported a positive result of the Minerva system with ratios of fraud-to-legitimate transactions of 1:150. Furthermore, once the network is trained it can analyze new data relatively quickly which is a vital aspect in the credit card fraud detection. In the credit card fraud detection Maes *et al.* (2002) conducted a comparison between neural networks and Bayesian networks. A Bayesian belief network (BBN) represents the joint probability distributions over a random set of variables; each node is a variable and each arrow represents correlation between variables. BNN's turned out to be more accurate and faster to train, however the neural networks were much faster to execute and given the importance of the small time span, the neural networks are preferred.

Regression Trees and Random Forests

Classification trees are extensively used in detecting credit card fraud; however some related techniques are more even more effective and applicable in the fraud detection. Sudjianto *et al.* (2010) provide an overview of applied classification techniques to detect financial fraud and the following figure illustrates the results of the applications.

Method	Training error	Testing error
Logistic regression	8.9%	9.9%
CART	14.3%	14.4%
C4.5	1.8 %	13.5%
Random forest	0.0%	10.8%
Boosting (AdaBoost)	2.7%	10.0%
Boosting (LogitBoost)	0.0%	9.1%

Figure 20: Applied Classification Techniques (Sudjianto *et al.*, 2010)

"Tree ensembles attempt to overcome the limitations of simple tree methods by combining the outcome of multiple models in a single classification decision" (Sudjianto *et al.*, 2010). According to Caruana and Niculescu-Mizil (2006) learning methods such as boosting, random forests, and SVMs achieve excellent performance. Bhattacharyya *et al.* (2010) support this; from the standpoint of fraud detection, random forest (Breiman, 2001) and boosting (Freund and Shapire, 1997) are the two leading learning methods most commonly used. This is supported by Sudjianto *et al.* (2010) since "Random forest" and "Boosting" (LogitBoost) have a training error of 0.0%. Boosting is a sophisticated algorithm used for assigning weights. Boosting will generate a sequence of classifiers, where each consecutive classifier in the sequence is an "expert" in classifying observations that were not well classified by those preceding it. During deployment (for prediction or classification of new cases), the predictions from the different classifiers can then be combined (e.g. some weighted voting procedure) to derive a single best prediction or classification (Statsoft, 2010).

3.7.2 Unsupervised Techniques

Bolton and Hand (2001) state; "unsupervised credit card detection has not received adequate attention in the academic literature." Credit card fraud detection heavily focuses on supervised techniques to

discriminate between legitimate and fraudulent transactions based upon previously found fraud. Outlier detection is employed to measure the distance between data objects to detect those objects that are grossly different from or inconsistent with the remaining data set (Han and Kamber, 2006).

Anomaly Detection

The unsupervised anomaly detection approach overcome the problem of misclassification by making use of data clustering algorithms, which make no assumptions about the labels or classes of the patterns. The patterns are grouped together based on a similarity measure and the anomalies, or attacks, are the patterns in the smaller clusters. Two assumptions need to be made for this to be relevant: the normal patterns or connections are much more numerous than the attacks, and that the attacks differ from the normal patterns. An advantage of both of these approaches is that they can detect new emerging threats. The drawback of data clustering for anomaly detection is a potentially high false alarm rate.

Clustering Analysis & Peer Group Analysis

Bolton and Hand (2001) propose Peer Group Analysis (PGA) as a candidate method for an unsupervised fraud detection technique; it is a technique that combines profiling and clustering analysis. Cluster analysis is a common descriptive task to identify a finite set of categories or clusters to describe the data set (Kantardzic, 2002). The basic idea is to group similar objects together to reduce the size of large data sets (Tan *et al.* 2006). To minimize the inter-cluster distances and maximize the intra-cluster distances, groups of objects with similar characteristics are combined and a group average is calculated. If the observation (the target) is dissimilar from the other observations in the cluster, it is regarded as an outlier (local outlier detection). This comparison could be with other accounts or with the same account at a previous time (Sudjianto *et al.*, 2010, Bolton and Hand, 2001, Weston *et al.*, 2008). In PGA, the behaviors are characterized as a profile and function as a basic of the observations. Then clusters of similar observations (peer groups) are identified and clustered and the mean of all the observations of the clusters form the basic. Subsequently the individual behavior (observation score) is compared to the cluster's behavior (the mean). These peer groups are monitored over time and any member with deviating behavior can be flagged for further investigation. Bolton and Hand (2001) show two examples in the following figure; the peer group members are represented in the orange graphs and the black graph is an individual from the peer group with deviating behavior when compared to the peer group. The peer group consists of 40 members.

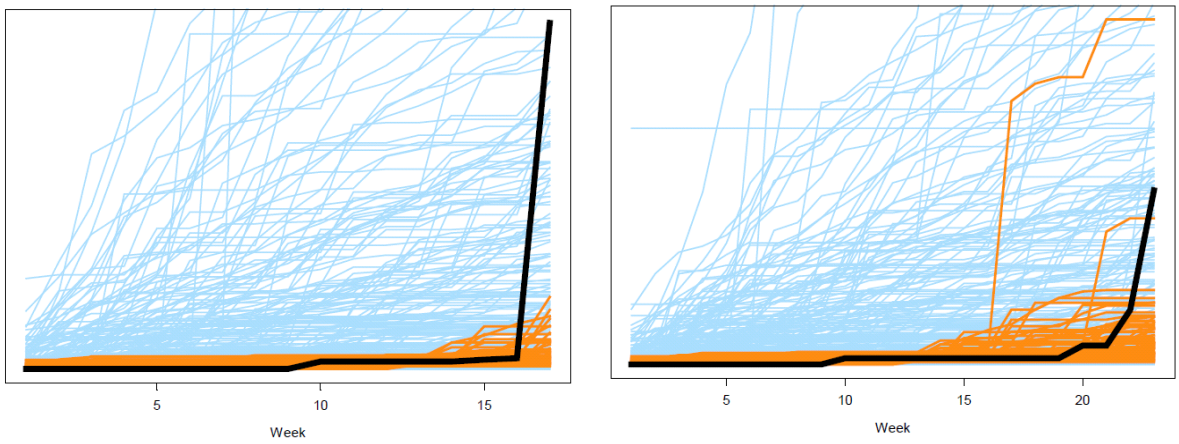


Figure 21: Peer group analysis plots (Bolton and Hand, 2001)

PGA is a clustering tool used for monitoring behavior over time in data mining situations. The nearest-neighbor method can be employed to combine transaction information from accounts that exhibit similar information. A peer group performs segmentation on known characteristics to verify and compare with an individual target. Each object is selected as a target object and compared with all of the other objects in the database on the basis of either internal or external information (information from other sources such as geographic or demographic information). Once a peer group has been defined its future behavior is used as a basis to compare future behavior of target objects. The general principle is that the cluster should be small enough to catch the local structure but not too small to make summaries (estimations) of peer group statistics unreliable. It is a compromise between how similar the members of the group are to the target and how small the group is. Once a peer group is established then the peer group statistics can be calculated (Bolton and Hand, 2001). The distinguishing feature of Peer Group Analysis is the focus on local patterns and outliers rather than on global models (Hand *et al*, 2000; Hand, Manilla, and Smyth, 2001).

3.7.3 Profiling

Spending patterns in a certain country or region are captured in the signature of the customers. In the time before the online era profiles could provide a proper indication if a certain transaction belongs to a particular customer. Currently users can pay online with their credit card and location is becoming a less important traceable indicator. However, the spending pattern of a customer can be captured when the card is used with a certain frequency. The frequency of detection must be kept under a reasonable level to limit the alarms in practice tuning thresholds regulates the rules when an alarm is issued. Fawcett and Provost (1997) proposed an adaptive user profiling method that uses user-specific rules and thresholds in detection. User profiling is attractive since it does not depend upon special hardware capability or customer input. The statistics applied in the credit card industry used to compare spending between accounts is the mean amount spent over the time window. Hand and Blunt (2000) noted that cumulative credit card spending trends over time are remarkably linear. Sudden jumps or changes in these curves merit investigation. Aggregation also has the advantage of not requiring precisely labeled data and may be more robust to the effects of population shift.

3.7.4 Money Laundering and Network Analysis

Money laundering is closely related and integrated as an extension of the fraud detection in the credit card industry. In financial fraud detection literature, link analysis is successfully applied (Zhang et al, 2003) and it is not mentioned in the fraud taxonomy presented by Laleh and Azgomi (2009). To discover money laundering it is important not to focus on individual behavior since it is often committed by a group of people working together as an organized crime organization. The individual transactions appear to look legitimate and are conducted with small monetary transactions which are not detected by the system. However when the transactions are reviewed in the context of a pattern of activity, often involving several related individuals, the criminal behavior can become more apparent (Sudjianto *et al.* 2010). Network analysis starts with a known entity of interest and finds meaningful relationships with other entities. Often the attributes used to indentify related individuals are defined by investigators based on their experience.

One option for finding these criminal groups is by using clustering algorithms to identify customers with similar behavioral patterns. However, these criminal groups are often spread out their economic activities over multiple customers, across accounts in several countries, and often over lengthy periods of time. Regular cluster analysis may be unable to detect these activities in such complex networks. Link analysis

may be able to detect these groups of people working together and these techniques are recently being applied in financial crime detection, especially for the purpose of detecting money laundering (Goldberg and Senator, 1995; Goldberg and Wong, 1998; Zhang, Salerno, and Yu, 2003). Zhang, Salerno, and Yu (2003) propose a method of using link discovery on correlation analysis, which they then applied to investigate money laundering crimes.

3.8 Telecommunications Fraud Detection

Two types of telecommunications fraud are described (Cahill *et al*, 2002; Laleh and Azgomi, 2009): superimposed fraud and subscription fraud. With superimposed fraud; the details of the account have been obtained and the phone calls are made by someone other than the plan owner. On the other hand, subscription fraud is when the account is obtained by using false identification documents and then the plan is used until the company blocks it.

3.8.1 Neural Network Classification

For fraud detection in the telephone networks supervised neural network engines are developed worldwide. Neural networks can be trained with expert knowledge and available data of fraud patterns for pattern recognition and then produce the necessary alarms. The high flexibility for pattern recognition is an appealing characteristic of neural networks and therefore is widely applied in the credit card and telecommunications industries (Dondorroso *et al*., 1997; Fadlalla and Lin, 2001; Kou *et al*., 2004). "Neural networks are systems of elementary decision units that can be adapted by training in order to recognize and classify arbitrary patterns" (Moreau *et al.*, 1997). Neural networks can calculate user profiles in an independent manner and adapt to the user behavior while reducing the operation costs substantially (Kou *et al*., 2004). However, neural networks are difficult to implement and function like a black box with an input and an output, which makes it hard to understand and interpret the results of neural networks. General disadvantages of neural network classification includes the lack of available labeled data and the fact that it is not clear how the system will handle new fraud strategies (Moreau *et al*., 1997).

3.8.2 Unsupervised Techniques: Anomaly Detection

In the telecommunications industry anomaly analysis has been applied to detect extremes and outliers and to analyze calling behavior of customers. Observations marked as outliers are not necessarily fraudulent, however, further investigation might be needed to determine this. The reason why it is an outlier explains a major part of a certain observation is fraudulent, someone with a legitimate distinctive calling pattern is not a fraudster. However, when the duration of international calls (< 2 hours) are analyzed and outliers exist it might be a wise decision to take a closer look at those outliers. International calls concern a lot of money so therefore special attention is paid to this area. Analysts specialized in international telecommunications fraud detected that certain foreign countries are particularly more used. Extensive international phone calls to that particular country should receive more attention to prevent financial losses due to fraud and abuse.

The outcome of anomaly detection is usually one of the two types: a label or a score. A *label* classifies an instance as either normal or anomalous. The *score* technique assigns a score to each instance to give an indication of the degree to which a certain instance is considered to be an anomaly. A score technique allows the output to be ranked and with a set parameter it can be hypothesized when an anomaly is fraudulent.

3.8.3 Profiling

Profiling and account signatures are techniques designed to capture the calling behavior of customers. Real time data is available and provides an overview of the behavior of the customers. How often and to which countries does a customer call to in a certain time period? Summarizing account activity is a major step in designing a fraud detection system because it is rarely practical to access all the call records of a certain account every time it is evaluated for fraud (Cahill *et al.* 2002). Ferreira *et al.* (2006) mention that fraud analysts are not interested in call details but rather in the customer behavior patterns to detect fraudulent behavior. In order to capture the behavior of a customer the characteristics of a user are captured in a signature. A signature corresponds to a set of information that captures the typical behavior of a user such as:

- Calling rate (Calls/hour)
- Most frequently called countries
- Regions of the world called
- Distributions of calls by the day of the week
- Average call duration /Longest call duration
- Most frequently called phone numbers
- Average number of calls received
- Distribution of the call duration

Additionally customer data can be added for a complete profile construction. According to Ferreira *et al.* (2006) relevant additional customer information in the telecommunication industry includes:

- Age
- Location
- Job
- Price plan

Account summaries can be compared to arbitrary thresholds (rules) for each period to verify if an account summary exceeds the threshold. A threshold exceeding the standard deviation is a simple example which warrants further inspection. Murad and Pinkas (1999) use profiling and an alert is raised if the daily profile's call duration, location, and quantity are exceeding the thresholds and the standard deviation of the profile. Fawcett and Provost (1997) propose a similar model with account specific thresholds rather than generic thresholds that apply to all accounts. The result is that the fraud detection system is more sophisticated and fewer false alarms go off. In the telecommunications industry it is rather common for a customer to have multiple phone planes. To analyze the problem at a customer level these accounts need to be aggregated or compared to look for fraudulent behavior. At a higher level, statistical summaries of call distributions (often called profiles and signatures) are compared to a certain threshold determined by experts (Fawcett and Provost, 1997).

Summarization and Break Point analysis

Activity monitoring is the task of analyzing the data streams in order to detect the occurrence of interesting behavior (Fawcett and Provost, 1999). One form of activity monitoring is "*break point analysis*" which identifies changes in behavior based on the transaction information in a single account (Bolton and Hand, 2001). In the figure the cumulative minutes spend by a certain customer is monitored

and when it passes a predefined threshold an alarm is issued. Recent transactions are compared with previous spending behavior to determine if the transaction fits the historic behavior. Detection of an increase in spending and rapid spending; features which would not necessarily be captured by outlier detection tools, are detected by activity monitoring. Cellular phone fraud detection is a typical example of activity monitoring; the task is to scan a large set of accounts, examining the calling behavior of the clients, and to issue an alarm when an account appears to be defrauded. The accounts need to be monitored continuously to identify fraud as soon as possible. When an alarm is raised a human fraud analyst is notified to examine the account carefully and to decide upon the next step. Therefore, it is important for the system not to create too many false alarms. Fawcett and Profost (1999) state that the definition of "as soon as possible" and "not too many false alarms" depends on the situation and industry and it varies with the amount of fraud, the workforce and other related factors.

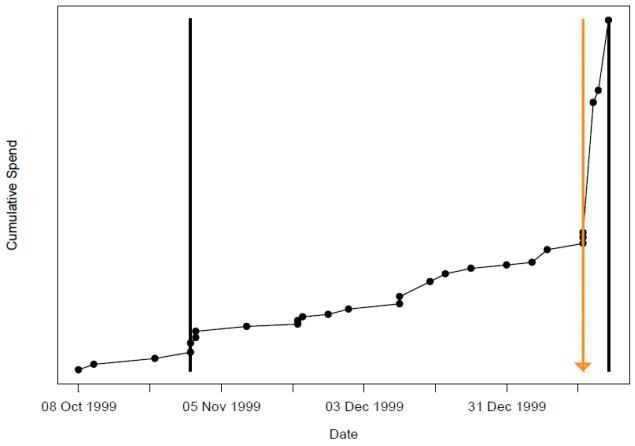


Figure 22: Break Point Analysis (Fawcett and Provost, 1999)

3.8.4 Visualization

To discover trends pattern recognition visualization is an effective tool to analyze a huge amount of data. Graphs and pictures can contain a lot of information and provide the user with useful patterns and trends. Becker *et al.* (2010) investigates telecommunications fraud and highlights the use and effectiveness of visualization and this is illustrated in the figure shown below. The light colored blocks show calls to a specific foreign country out of hours, starting at the 14th of January 2007 continuing until the 28th of January 2007. Note that the figure below represents 7,000 phone calls also that the uniform height of the calls to a certain country at 45 minutes. This is an example of how visualization can contribute to fraud detection; it can support fraud detection by detecting trends and outliers to discover irregularities in the data.

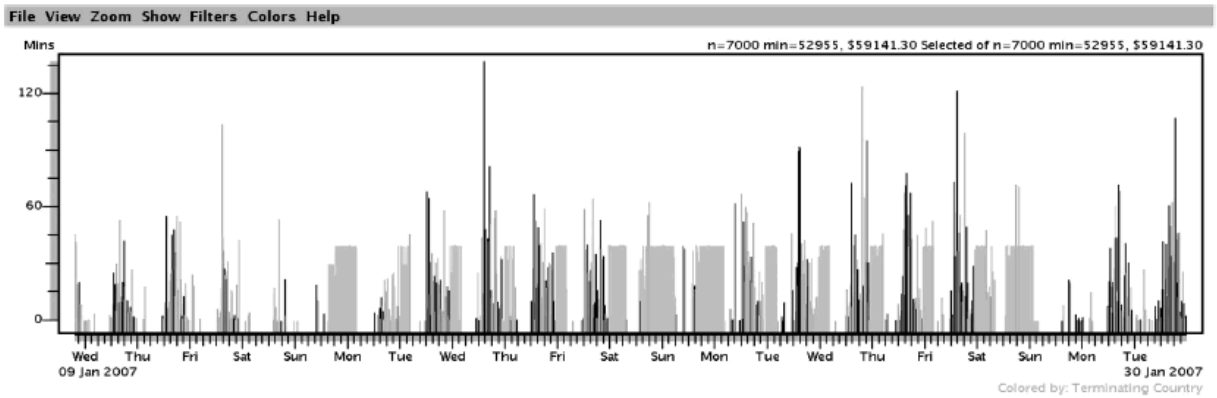


Figure 23: Visualization of 7000 phone calls (Becker *et al.*, 2010)

In low dimensional data representation outliers and unusual patterns are more subjective and may be naturally occurring. For example the heights of adult humans contains outliers, someone can be taller than 7 feet. Additionally box plots make no assumptions about the data distribution model but are reliant on humans to interpret the extreme points plotted on the box plot. As the dimensionality increases, the data point are spread out through a larger volume and become less dense; this is also known as the "curse of dimensionality". For low dimensional data representation visualization is a useful and effective method, however if the dimensionality increases the applicability declines.

3.8.5 Rule Based Methods

An effective method for detecting fraud is to check for suspicious changes in user behavior. A fraud detection system can be trained up front by using rules that raise alarms when certain (consecutive) activities occur that are highly suspicious of fraud. Examples of these rules are collision detection and velocity detection. *Collision detection* is a method which involves analyzing call data for overlapping calls. If a call is made from Chicago and at the same time the account is used to make a phone call from New York it is an indication that someone else is using the customer's account, while it is a requirement is that only one person is registered on the contract. *Velocity checking*, involves analyzing the locations and times of consecutive calls to determine if a single user could have placed them while travelling at a reasonable speed (Davis and Goyal, 1993). When a phone call is made from Los Angeles it is physically impossible to be in Miami ten minutes later making the next phone call.

Collision detection and velocity check are both considered to be accurate in the telecommunications industry (Fawcett and Provost, 1997). However, they share the disadvantage that the applicability depends on a moderate level of legitimate activity. Low usage subscribers will rarely cause collision or velocity alarms and therefore a complementary technique would be profiling. The addition of profiling often works well with low-usage subscribers because unusual behavior is very prominent. Profiling is a good complement to velocity and collision checking section because it covers aspects that the others might miss. It is necessary to discover indicators corresponding to the change of behavior that are indicative for fraud, rather than absolute indicators for fraud. It is essential to profile the behavior of individual customers in order to characterize their normal behavior. Consequently, the analysis should be regarded as alerting the fraud detection team if anomalies occur, this might be an indication for fraudulent behavior and further investigation is required. Ideally, a fraud detection system should be able to automatically learn and use rules to discover fraudulent behavior.

Fawcett and Provost (1997) applied the Detector Constructor framework (DC-1) that combines rule based indicators with the profiles of customers. Once a customer passes a certain threshold the system generates an alarm and analysts are notified automatically. For example, a "high usage detector" generates an alarm when a customer suddenly has a huge jump in their calling behavior. This type of monitoring is closely related to the break point analysis and activity monitoring. The DC-1 combines a rule-based approach with customer profiling.

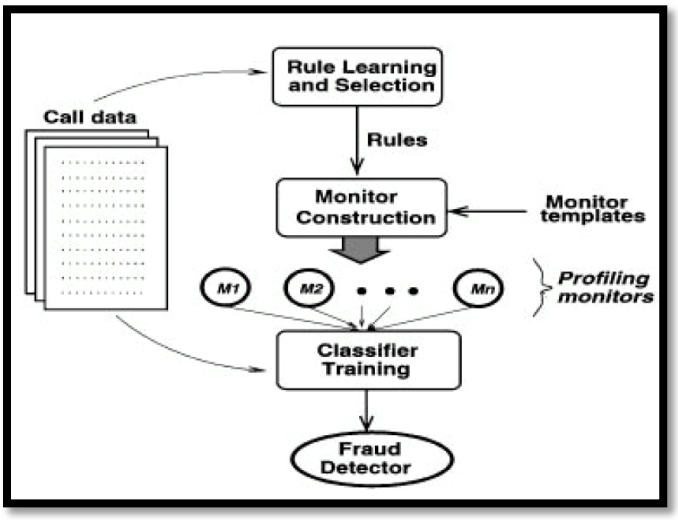


Figure 24: The DC-1 framework for fraud detection (Falcon and Provost, 1997)

In section 3.9.2 an attempt to apply this DC-1 fraud detection system to the intrusion detection domain is described with the corresponding difficulties.

Evaluating the success of the fraud detection framework and compare the results against existing expert systems is difficult. Fraud detection departments carefully protect information about how effective their system is and how much fraud they have discovered. Likewise, vendors protect the details of their detection systems since it contains trade secrets. As explained earlier, that is one of the disadvantages of fraud detection research, techniques and fraud schemes cannot be published openly or else the fraudsters would immediately change their strategy and the impact of the detection technique would be highly reduced. In reality it is impossible to know how much fraud actually exists or how greatly fraud detection strategies actually reduce it.

3.9 Computer Intrusion Detection

This section introduces hacking and computer intrusion designed to access information secured in computer networks and systems. In this discipline anomaly detection has been extensively applied since hacking often creates outliers. These include attempted break-ins, masquerade attacks, leakage, denial of service and malicious use (Kou et al, 2004).

"Intrusion is defined as the potential possibility of a deliberate unauthorized attempt to access information, manipulate information, or render a system unreliable or unusable" (Kou *et al.*, 2004). An intrusion can be defined as any set of actions that compromise the availability, confidentiality of integrity of resources such as user accounts and file systems. The models characterize the normal and legitimate behavior of these resources; the detection techniques compare the actual system activities with the established models, and intrusive and abnormal behavior is identified (Laleh, 2009). Intrusion detection fraud can be classified into two categories; misuse intrusion and anomaly detection. *Misuse detection* is an attempt to recognize the attack of previously observed intrusion in the form of a pattern or a signature (e.g. login credentials) and directly monitored for the occurrence of these patterns. *Anomaly detection* attempts to establish a historical normal profile for each user and then uses a sufficiently large deviation

from the profile to indicate possible intrusions (Lee and Stolfo, 1998). In the next two sections the applicable techniques for both categories will be discussed.

3.9.1 Supervised Classification Techniques

In the context of intrusion detection supervised methods are sometimes called misuse detection (Bolton and Hand, 2002). Supervised classification techniques as described in section 3.6.1, are applied in the credit card industry and are based on the principle of comparing new data with historic classified data to detect fraud while it is happening. Misuse detection is similar with supervised classification techniques. Mukkamala *et al.* (2002) compare Support Vector Machines with Neural Networks for intrusion detection of information systems with the goal to detect both misuse and anomalies. To learn normal behavior versus attack behavior the systems are trained with historic audit data gathered after previous attacks. "SVMs classify data by determining a set of support vectors, which are members of a set of training inputs that outlier a hyper plane in feature space" (Vladimir, 1995).

3.9.2 Unsupervised Techniques

Within intrusion detection anomaly detection systems characterize the behavior of individual users and issue alarms of intrusion when abnormalities in the behavior are detected. Unsupervised techniques are generally methods of anomaly detection, based on profiles of usage patterns for each legitimate user (Bolton and Hand, 2002). Examples of anomalies in computer intrusion are an unusually high number of network connections within an interval of time or unusually high CPU activity (Kumar and Spafford, 1994). To recognize anomalies by using profiles is an approach that is facing similar problems as the credit card industry where the amount of data is comprehensive. Recognizing and matching patterns to determine if certain behavior belongs to a specific profile is subject to several difficulties. The high amount of generated data by CPU's and the opportunity that users operate under several user identities makes recognition problems very complicated.

Supervised models are trained to detect the difference between legitimate transactions and previously known fraud. The unsupervised anomaly techniques make an implicit assumption that normal instances are far more frequent than anomalies (Chandola *et al.*, 2009). A well known problem expressed in literature is that if the assumption is not true the technique suffers from a high false alarm rate (Bolton and Hand, 2001). Fawcett and Provost (1999) report that applying the fraud detection model DC-1 in the intrusion detection domain had disappointing results. DC-1 is designed for cellular fraud detection and the work of Fawcett and Provost (1999) prove that despite some similarities the transferability of fraud detection techniques should not be taken for granted. Every industry has unique specifications and therefore fraud detection models cannot easily be transferred. However, the fundamental components of fraud detection techniques, detecting outliers and irregularities, are general aspects which can be applied in every industry.

3.10 Overview

The most important lesson the health care industry can learn from this analysis is the importance of electronic fraud detection and the potential it posesses. The foundation of the fraud detection in the other industries is the fraud detection systems that point out which transactions are suspicious and indicates that analysts should further investigate. The huge amount of data cannot be processed without the support of sophisticated fraud detection systems. Systems alone cannot understand the complexity and dynamic

nature of fraud detection, therefore it is essential to keep humans involved to interpret what the data means and to judge if a transaction should be flagged as fraudulent or not.

Lessons from fraud detection in the credit card, telecommunications, computer intrusion industries

Electronic fraud detection is even more important in the Medicaid program than most other related industries given the fact that the input of beneficiaries is reduced to a minimum. Therefore the dependence on electronic fraud detection is much greater within the Medicaid program than the credit card and telecommunications industries. In these other industries customers immediately report fraud and abuse and both industries possess real time data so labeling the data is quite easy task. The Medicaid data available is not labeled and there are no signals that this will change in the near future. Multiple stakeholders and fragmented responsibility are also hampering the process of labeling fraudulent and abusive data. Therefore, the most important data analysis opportunity in the form of supervised classification is severely restricted. Supervised classification is the most sophisticated data analysis technique because the model can be trained and adjusted and therefore best suited to detect sophisticated fraud schemes. In the credit card industry supervised classification, neural networks in the 1990s, and currently support vector machines and random forests, form the basis for sophisticated and effective fraud detection. Certainly the argument that new fraud schemes are not detectable since the training data does not contain the newest fraud schemes is legitimate. However, in the Medicaid situation the most important aspect to control is the fraud and abuse because currently it is a worrisome situation.

In the telecommunications industry unsupervised techniques such as profiling and anomaly detection are applied to further complement the supervised fraud detection. Given the current extensive and intensive use of the telephone a lot of data is available; as a consequence an accurate profile can be constructed. A downside is that unsupervised techniques suffer from a high false alarm rate; the reason is that an outlier does not imply that fraud or abuse occurred. The consequences of supervised and unsupervised techniques are confirmed in the computer intrusion industry where supervised techniques, to detect known patterns, are supported by anomaly detection to detect new intrusions.

All of these industries have an important difference with the Medicaid program; they all possess real time data to monitor the process closely. Furthermore, these industries are supported by the users who report unusual events and behavior because it affects them directly. These commercial industries and their customers do not want to lose profit and therefore they are willing to allocate the necessary resources so that appropriate measures are taken. These industries and companies realize that fraud detection is a vital aspect of doing business and that is what went wrong when the Medicaid program was established. On top of that is the number of stakeholders involved and the fragmented responsibility which complicates the situation tremendously. The U.S. government can learn from the private industry to effectively fight fraud and abuse and although the situation may be more complicated where Medicaid is concerned, a great deal of progress can be made regarding fraud detection in the Medicaid program.

4. Electronic Fraud Detection in the Medicaid Health Care Program

'The human immune system does not consist of a single, generally applicable, detection system for alien bodies, but instead builds a new detector and weapon whenever it is confronted by a new treat'

- Professor David Hand (President Royal Statistical Society)

As outlined in chapter 3, data analysis has been proven to successfully detect telecommunications fraud, credit card fraud and computer intrusion fraud. How can the lessons learned from these industries be applied to improve fraud detection in the Medicaid Program? According to the overview presented in chapter 2 (§ 2.5) the known fraud schemes the feasible fraud detection techniques will be verified.

Fraud Type	Fraud Scheme	Short Explanation	Strategy
I	**Identity Theft**	Stealing identification information from providers and beneficiaries and using that information to submit fraudulent bills to Medicaid.	Fraud
II	**Fictitious Practitioners**	Using false documents and identification information to submit fraudulent bills to Medicaid.	Fraud
III	**Phantom Billing**	Submitted claims for services not provided.	Fraud
IV	**Duplicate Billing**	Submitting similar claims more than once.	Fraud/ Abuse
V	**Bill Padding**	Providing unnecessary services and the submitting these claims to Medicaid.	Fraud/ Abuse
VI	**Upcoding**	Billing for a service with a higher reimbursement rate than the service which was actually provided.	Fraud/ Abuse
VII	**Unbundling**	Submitting several claims for various services that should only be billed as one master claim that includes ancillary services.	Fraud/ Abuse

Table 8: Overview Fraud in Medicaid

4.1 Medicaid; the Lessons Learned

The lack of labeled data in the Medicaid program results in the situation that supervised data mining techniques are not easy to implement. This is a tremendous disadvantage for the Medicaid fraud detection team fighting fraud with the available resources. Strongly recommended is a joint effort of the federal government and the states supported by the commercial industry to improve the data supply and to enable labeling of the fraudulent and abusive data. This would enable supervised data mining and can lead to major improvements of the Medicaid fraud control. In this section the importance of applying electronic fraud detection in Medicaid program is outlined and briefly discussed.

4.1.1 Incentive to Report Insurance Fraud

Compared to the credit card or telecommunications industries, the U.S. Medicaid program represents a slightly different situation. In commercial industries the companies try to avoid inconveniencing their

customers. Fraud detection, which tends to inconvenience the customer in one way or another, is part of the game and that fact is often not understood by customers. When it comes to insurance fraud, providers and beneficiaries should realize that fraud control is part of the process, and be willing to readily offer more information so that the detection teams can improve their detection techniques while limiting the inconvenience factor.

In the Medicaid program, the government directly reimburses providers so the patients do not have to pay the (often expensive) bills. The government is the "payer of last resort," meaning all other potential insurance and other potential third parties that could be liable for the medical services would pay prior to Medicaid paying the balance due. The fact that Medicaid will pay the balance leaves little incentive for consumers of healthcare to worry about fraudulent billing. If customers notice a mistake on their bill they are inclined to think, "the government is paying it anyway so why would you worry" (Sparrow, 2002). This is a striking difference compared to the telecommunications, credit card, and insurance industries in which customers would immediately complain and report any instance if fraud because it is their own money that is being stolen. Even though it is mandatory for state Medicaid agencies to send an Explanation of Benefits (EOB), people rarely read and verify that Medicaid was billed correctly (Sparrow, 2002).

4.1.2 High Dependency on Electronic Fraud Detection

Given the fact that the U.S. government is the payer of last resort with little feedback from the consumer of services, the dependency on electronic fraud detection is significantly greater than the related industries. As learned from the case studies of the credit card, telecommunications and computer intrusion, fraud detection using supervised classification is the most effective technology. However the requirements for this approach are currently not available in the Medicaid program, and the successes from the other industries are mainly because of the possession of real time data and the opportunity to label data. While the implementation of classification techniques is complex and expensive and the results hard to interpret, it is a sophisticated technique capable of detecting intelligent fraud schemes. Benefits must be weighed against the costs to determine whether these fraud detection techniques are profitable and could represent a significant improvement to the fraud detection system. In chapter three, the analysis showed that supervised techniques are necessary for an effective fraud detection system. And, the extensive application of classification techniques in various domains proves the effectiveness and utility to contribute to fraud detection. However, no one technique is applicable to discover the extended fraud strategies and schemes. Therefore, a fraud detection system consisting of multiple techniques to maximize the changes to detect the fraud and abuse is needed – a flexible modular approach capable of adapting to the continuous changes in the fraud detection field.

4.1.3 Multiple Techniques Approach

Provost (2002) makes an important point, noting that fraud detection problems can be approached in various ways. These approaches include two-class supervised classification with training samples of fraudulent behavior to construct an assignment rule, unsupervised outlier detection in which one looks for patterns in data that do not conform to the expected behavior of that account, peer group analysis which compares the current behavior with previous similar accounts, and non-statistical rule-based techniques which match transactions to configurations known to be of high risks. All of these methods can be run in parallel or can be combined to form a basic input for a fraud detection system. Any one technique is not entirely comprehensive, and therefore no single technique is suitable to find all kinds of fraudulent

behaviors. It is important that a viable fraud detection system consists of several different techniques to cover all possible aspects. Sparrow (2002) emphasizes the use of several detection techniques at different levels to create a solid detection framework, as exemplified by the credit card industry. Becker *et al.* (2010) outline the importance of simple understandable models so that it is easy to adapt to the flexible fraud detection environment. Designing a mix of detection techniques, empirically testing these techniques, and comparing the results will achieve better results. A current example is random forests, a mix of decision and classification trees, which achieves better classification results compared to the established classification techniques (Sudjianto *et al.*, 2010).

4.1.4 Knowledge Discovery

Knowledge discovery can lead to new insights concerning fraud detection. According to Fayyet *et al.* (1996), knowledge discovery is the non-trivial process of identifying valid, novel, potentially useful and understandable patterns in the data**.** Three criteria to successfully apply data mining techniques include:

1. The data is available.
2. The past is a good predictor of the future.
3. The data contains what we want to predict.

The assumption in this research is that the Medicaid case meets these three criteria. However, a severe limitation of the Medicaid fraud detection system is the number of stakeholders involved and the lack of incentive for beneficiaries and practitioners to report fraud. In addition, fraud techniques evolve over time, sometimes very quickly, meaning the past is not always a good predictor of the future. The goal is to reduce the time to discovery of new fraud schemes and ensure they can be identified and prosecuted quickly upon discovery. In the following sections a proposal per fraud scheme is given how certain fraud types can be combated based on the conducted systematic literature review.

4.2 Fraud Type I: Identity Theft

Submitting claims with the identity information of others obtained in an illegal way constitutes identity theft. In order to submit a claim to Medicaid, the patient's Medicaid ID number is required. In order to deter identity theft type fraud, Medicaid providers and beneficiaries must be encouraged to verify their Medicaid EOBs and report irregularities. Without the support of the individuals involved in the treatment, it is a challenging task to verify if a claim is submitted fraudulently. Offering incentives to both providers and beneficiaries to report fraud or abuse committed with their identity would be a straightforward method to detect identity theft fraud, though a large, dedicated support infrastructure must be in place to both interact with the reporting party and follow up with their report of fraud. Given this high, expensive barrier to implementation, the focus of this research has been on electronic fraud detection techniques.

4.2.1 Profiling for Detecting Identity Theft

Creating a profile of a person involves capturing the characteristics of their medical claims over a certain period. As stated in section 3.7.3; summarizing account activity is a major step in designing a fraud detection system because it is rarely practical to access all of the records of a certain account every time it is evaluated for fraud (Cahill *et al.* 2002). A single transaction might not attract attention at a first glance, however, when put in perspective of the claiming behavior of a provider or recipient it might appear to be aberrant and possibly fraudulent.

The profiling technique is designed to capture the behavior of individuals and to discover abnormalities. This profiling is the claiming signature corresponding to the actions of the individual. Identity theft implies that a different individual or organization is using the identification; it is considerable that the subsequently submitted claims are aberrant compared to further submissions. Profiling is designed to detect these abnormalities and therefore profiling seems appropriate to detect identity theft.

Profiling can take place in several manners, and no profile is perfect, it depends on the situation. Several characteristics can be included in the profile, and possible new features should constantly be analyzed to verify if the profile can be improved to more accurately characterize the claiming activities to better detect fraud and abuse. A potential profile for the Medicaid Program recipients and providers is suggested in the following summarization:

- How many first time recipients/providers have been reported per month (large amounts of first time recipients/providers may indicate, a stealing 'a-little-all-the-time' strategy).
- Amount of submitted claims per month (A spike in claims may indicate a 'hit-and-run' strategy).
- Distribution of claims per month (Can be analyzed to see if a certain claim code is being exploited)
- Average number of claims by provider type (Possible indication for fictitious claims and/or unbundling schemes).
- Most frequent billed recipient or provider (Identity theft scheme and/or fictitious practitioners billing obliviously on a certain recipient).
- Average claiming amount per month (Possible indication for fictitious practitioners).
- Highest 20 claims per month (Possible indication for fictitious claims scheme, identity theft or upcoding scheme).
- Distribution of how many claims/providers or claims/recipient are reported (exploiting recipients/provider personal information).

Ferreira *et al.* (2006) suggest using additionally relevant customer data for a complete profile construction. Applying this in the Medicaid Program the following criteria might be useful to include in the characteristics of the profile:

- Age
- Gender
- Location (e.g. zip code)
- Profession
- Specialty (e.g. a back pain specialist will probably submit numerous back related claims)

The knowledge of medical subject matter experts is required to improve and optimize the profiles to characterize the claiming behavior as accurately as possible from a data analysis perspective. For example traveling for more than 100 miles for a specialist in cancer treatment might not be unusual, however to travel that far to visit a general practitioner might be a suspicious act. The expert knowledge of medical claiming experts is needed to define the fine line between normal and suspicious behavior. Furthermore is it important to encounter the timeliness of the profile, things change overtime and the construction of the

profile should be flexible enough to change accordingly. The time frame should move along as time passes by, e.g. a profile that is based upon the data of the last two years.

Provider Profiling for Detecting Identity Theft

Provider profiling would be a clear opportunity to identify identity theft fraud. Since the profile is a signature of a certain provider; the claiming behavior of the past is captured in their profile. When their Medicaid identification is stolen and used to defraud the program the submitted claims may well be distinguishable when compared to the established profile. The difference in claiming activity should be flagged and a data analyst should examine the sudden change in behavior. Are there more divergent aspects in the claim data that require further investigation to verify if the provider identification is not stolen and used for fraudulent purposes?

Beneficiary Profiling for Detecting Identity Theft

A profile for beneficiaries serves the goal of fraud detection, as well. However, the amount of claims submitted per beneficiary is significantly smaller than that of a provider. Most providers submit several claims per week, and, consequently, the available data to base the profile on is more extensive. Therefore trends and patterns might be harder to recognize in the profile of a beneficiary due to the smaller subset of available data. If a beneficiary is the victim of identification theft then the sudden change in the profile could stand out as being significant, or it could simply be an example of a new acute medical condition or the onset of a new chronic condition. Profiling groups of beneficiaries with similar demographics and environmental conditions could lead to better profiles for predicting anomalies. Beneficiary profiling could be especially effective when the strategy of the fraudster is a 'hit-and-run', where as much claims as possible are submitted by the provider in a short time, they receive payment, then disappear.

Limitations Profiling for Detecting Identity Theft

Sparrow (2002) is rather skeptical in his analysis about the profiling technique since profiling cannot cover the whole spectrum of fraudulent schemes found in the medical programs Medicaid and Medicare. Only fraud that produces anomalous billing patterns is discovered with this technique because only those outliers will be detected. The majority of Sparrow's (2002) critique is that profiling can only take place as a post payment review and that it will take at least a number of months before the data is analyzed and profile updated.

Activity Monitoring for Detecting Identity Theft

Activity monitoring is a form of profiling and can raise the suspicion score or flag possible fraudulent transactions if the claim is divergent compared to the established profile. When fraudsters use the 'steal a little all the time' scheme profiling and activity monitoring might not be an appropriate tool. The fraud might be included in the corresponding profile with a result that the fraudulent transactions do not attract the attention of fraud detection systems. Since the difference in behavior might be too subtle for profiling and activity monitoring to detect, they are often not an appropriate technique to discover identity fraud.

Analyzing and using longitudinal information to compare changes in behavior is a powerful application of activity monitoring. If the claiming behavior of a provider during a certain month is compared with their behavior throughout previous months, then rapid increases can easily be detected. For example, a rapid increase in the number of submitted claims or, a sudden increase in the total amount of submitted claims per month constitute suspicious phenomena. Activity monitoring includes scanning a large set account to detect fraudulent and abusive signs. Brake point analysis is a tool that combines profiling and activity

analysis with rule based techniques. The fraud detection system is designed to issue an alarm if certain thresholds are exceeded and a human should further investigate the case. In this way the system supports the humans in the fraud detection process and enable effective monitoring so that the people can invest their time in the cases flagged as possible fraudulent or abusive.

The effectiveness of the monitoring process is dependent on the rules of the fraud detection system. Rules determine when an alarm is issued and the objective is to keep the amount of false alarms to a minimum. As mentioned previously, the effectiveness of the rules is dependent of the field in which it is applied and the support of field experts is required. The generation and selection of rules is the key aspect of a rule based approach. It is obvious that men do not need pregnancy treatments and experts are needed to acknowledge this fact. However many situations, such as finding the fine line of when it is useful to provide psychiatric treatment to children, are complex cases. It might be extremely unusual that children under the age of 16 receive psychiatric help; and so a health expert can provide his experience and opinion to set a reasonable and plausible age to set a guideline. Nevertheless, analysts should keep in mind that outliers exist and that odd treatments can happen, therefore it is prudent for analysts to further investigate the flagged cases. Subsequently a list of possible suggestions of fraud detection rules is presented:

- Old enough for psychiatric treatment
- Demographic data (such as, visiting a provider which is more than 100 miles from home)
- Drugs that can be prescribed 3 times per year (subscriptions > 3)
- Pregnancy treatments for men
- Anti depression medication prescribed for children

The provided list is a suggestion for a rule based approach in the Medicaid Program. The construction and selection of rules can be designed in various ways. Fawcett and Provost (1997) suggest personal thresholds for a more sophisticated approach to detect fraud in order to reduce false alarms issued by the fraud detection system. This is considered the next step in the field of profiling in Medicaid and should be further investigated once the general profiles with thresholds are implemented. The results can be compared with the beneficiary or provider profiling to note if it is an improvement.

4.2.2 Anomaly Detection and Cluster Analysis for Detecting Identity Theft

Anomaly Detection is an outlying observation, or outlier, that appears to deviate markedly from other members of the sample in which it occurs (Grubs, 1969). Analyzing the submitted claims outliers of beneficiaries or providers with excessive claiming behavior stand out. If individuals with stolen Medicaid identification get greedy and bill excessively, it is not their account and the hit and run strategy (section 2.5.1) would lead to outliers as a consequence of their claiming behavior. An example of excessive claiming behavior is provided in an article of the Washington Post where outliers were detected using outlier detection (see figure 26). The next step is to further investigate the actual outliers to find out what the reason or explanation is for their uniqueness.

An analysis of the data with the anomaly detection technique can result in discovering beneficiaries or providers with unusual behavioral patterns. For example, if the amount of submitted claims is investigated as shown in the following figure where an outlier was detected using outlier detection.

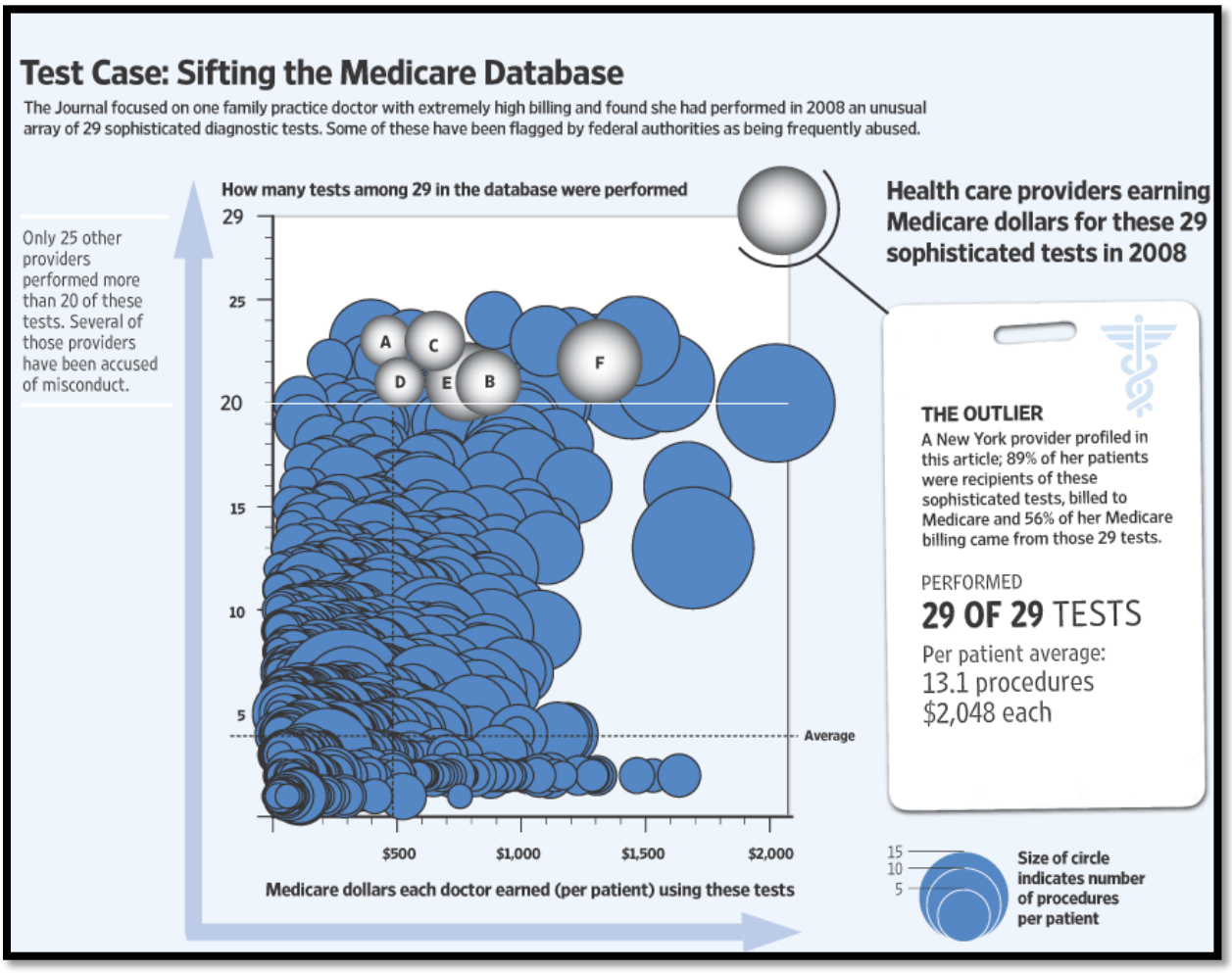


Figure 25: Anomaly detection in Medicare (Washington Post, 2010)

Cluster analysis is a localized form of outlier detection in which, the behavior of a Medicaid participant is grouped together with participants exhibiting similar behavior. If changes occur compared with the cluster they belong to, further research is needed to show the reason for the deviation. A (local) outlier is not a guarantee that fraud is committed, especially in the health care industry where various specialist outliers will occur. A certain backache expert might provide an exceptional amount of a certain back treatment and compared to other specialist in a certain region he is considered to be an outlier. However that does not imply the backache expert is a fraudster with criminal intentions. After applying outlier detection or cluster analysis a next step is always needed to verify the reason why the outlier occurs. Both techniques can be used to identify and flag unusual behavior as a first step in the fraud detection process. The next step is to determine if the outliers correspond to fraudulent or abusive behavior.

4.3 Fraud Type II: Fictitious Practitioners

As stated in chapter 2, fictitious health care providers are a real threat to the program. Phantom corporations that solely exist on paper attempt to drain Medicaid using a fictitious provider name to submit fake claims without the intention to deliver any service. The fraudsters are criminals who deliberately choose to defraud Medicaid and acquire their fictitious identity with falsified documentation. The criminal intention to defraud the program with his type of fraud has a lot in common with identity

theft; to defraud the system without any intention to deliver any kind of health care service. When an identity is stolen a sudden change in behavior could be expected and a profile would be an appropriate tool to detect that change. In this case, the fictitious identity is used from the beginning to defraud the program. The fraudster is captured in the profile which might look just like a thousand other providers.

It is important to prevent providers with fictitious identities from entering the Medicaid program in the first place. Eligibility criteria need to be met in order to enter the program, and using falsified documentation is a violation of the law. However, fraud detection teams need to realize and keep in mind that fraudsters with false documentation are able to enroll themselves into the program. Fraud detection techniques can play an important role to fight this fraud type of fictitious identity.

4.3.1 Rule-Based Methods for Detecting Fictitious Practitioners

Fake treatments and services are submitted and therefore it is likely that providers use the information of beneficiaries only once to avoid being detected by the beneficiaries. The chances of being discovered are smaller if only the beneficiary's information is used inappropriately once. As a result these providers have a high number of patients who receive treatment for 'the first time'. A couple of interesting queries could be useful by detecting fictitious fraud are presented:

- How many patients are seen for the first time in a certain period and how does it compare to other similar providers in the same region/period/specialism etc.
- The amount of patients a provider sees per month.
- The distribution of services provided.
- The distribution of submitted bill amounts.

These examples of applicable queries are a suggestion to begin thinking along this direction and it does not mean to imply that this is an exhaustive list to fight fictitious fraud. These queries need to be tested in several states with corresponding Medicaid data to verify their applicability and to see how results can be improved by applying heuristics. For example, changing the variables might lead to a better result, however, in this research no fundamental conclusion can be given.

4.3.2 Statistical Methods for Detecting Fictitious Practitioners

Since the provider needs to fabricate all of the medical services and beneficiaries who received these treatments; the statistical Benford's law might be an interesting and applicable technique to utilize in this case. Providers need to make up the costs for the fake treatments and since made up numbers are slightly different, and therefore detectable by Benford's law, it should be an option to consider. Sparrow (2002) mentions the existence of lists with information about Medicaid beneficiaries which are available on the black market that are used bill the Medicaid program. If one of these list is discovered then a query could be developed to verify which providers provided several services to these patients known to be on a black market list. In the telecommunications industry there are currently fraud detection lists of numbers that have been used for fraudulent activities.

4.3.3 Supervised Data Mining Techniques for Detecting Fictitious Practitioners

Classification techniques which are trained to detect the difference between fraudulent and legitimate claims present an opportunity for the use of electronic fraud detection in Medicaid. "Classification has been the most popular and the only way used so far to identify fraudulent financial statements" (Dianmin *et al.*, 2007). In section 3.7 and 3.8 the use of supervised techniques in the credit card and

telecommunications fraud detection industries is outlined, and the results are positive. Focus on certain medical segments such as hospice care where it is harder to verify if the presented procedure was necessary and actually took place. However further research is required to judge about the applicability of supervised data mining techniques.

Identity fraud is a form of fraud which is hard to detect since no particularly suspicious changes have to occur. The fraudsters submit fake claims which are similar to legitimate claims and a verification method is required which verifies if the treatment actually took place. This fraud scheme is quite sophisticated and when no extreme billing behavior is demonstrated techniques like outlier detection and profiling fail to detect. Unsupervised data mining techniques such as outlier detection focus on outliers and abnormalities which are not caused if "normal" billing takes place. Next profiling and activity monitoring have a strong focus on detecting changes in behavior and in this scenario that is not the case, therefore profiling does not seem an effective tool to combat this type of fraud.

4.4 Fraud Type III: Phantom Billing

Billing for services not provided is quite similar to fraud type I: identity theft. Providers committing phantom billing steal and abuse beneficiaries' identification information in order to bill Medicaid for services which were never provided. Phantom billing concerns the illegal use of beneficiary information (with or without the knowledge of the involved beneficiaries) to bill Medicaid for procedures never executed.

4.4.1 Peer Group Analysis for Detecting Phantom Billing

Phantom billing is billing for services never provided. This is rather hard to discover when only the claiming behavior of the particular provider is analyzed. An interesting question would be if other experts in the same medical field are showing approximately similar treatments. Therefore to find out if a provider is exhibiting a suspicious claiming behavior the claiming behavior of the provider should be compared to other specialist in the same field of expertise. Peer Group Analysis (PGA) is an interesting tool used to find out how the behavior of a certain provider relates to the behavior of other providers within a certain group. An important aspect of this analysis is how the providers are grouped together and why they are grouped together. Especially in the medical industry extremely specialized specialties exist and it not an easy task to group some of them together approximately.

Therefore the use of medical experts is needed to form the groups to subsequently apply Peer Group Analysis. The assumption is made that grouping providers is possible and then the comparison of the behavior can start. As an example 20 dentists in a certain zip code are grouped together and the distribution of treatments are vetted; the result is a group average of the amount of times certain treatments are provided per group of patients (e.g. 1000):

Treatment	Average per month	Reimbursement Rate
Cavity treatment	150	$ 50
Pulling teeth	15	$ 300
Root canal	3	$ 1200

Table 9: Group average example for PGA (These numbers are not a reflection of the reality)

The summarization of the group averages are shown in the table and this is an example of how PGA could be applied. Now the profiles of a particular dentist can be compared to the group averages. If a

certain dentist performed an unusually high number of root canals during a particular period it does not mean that fraud has been committed; however it might be an indication that this dentist needs to be investigated. Since a root canal is considered to be an oral surgery (Medicaid Florida, 2010) the reimbursement rate is significantly higher than the other treatments in the example and therefore it is likely that dentists with devious intentions use this treatment to bill Medicaid.

4.4.2 Anomaly Detection supported by Visualization for Detecting Phantom Billing

The detection of anomalies and outliers is heavily used in fraud detection and an example of visualization of an anomaly analysis is provided by the San Diego Supercomputer Center; the national Medicaid database. Outliers and trends are easier recognized when presented in graphs; it is easier for the human mind to grasp details when it is visually represented. Since fraud detection is mainly about outlier detection, visualization is an appropriate tool to create insight about the data distribution.

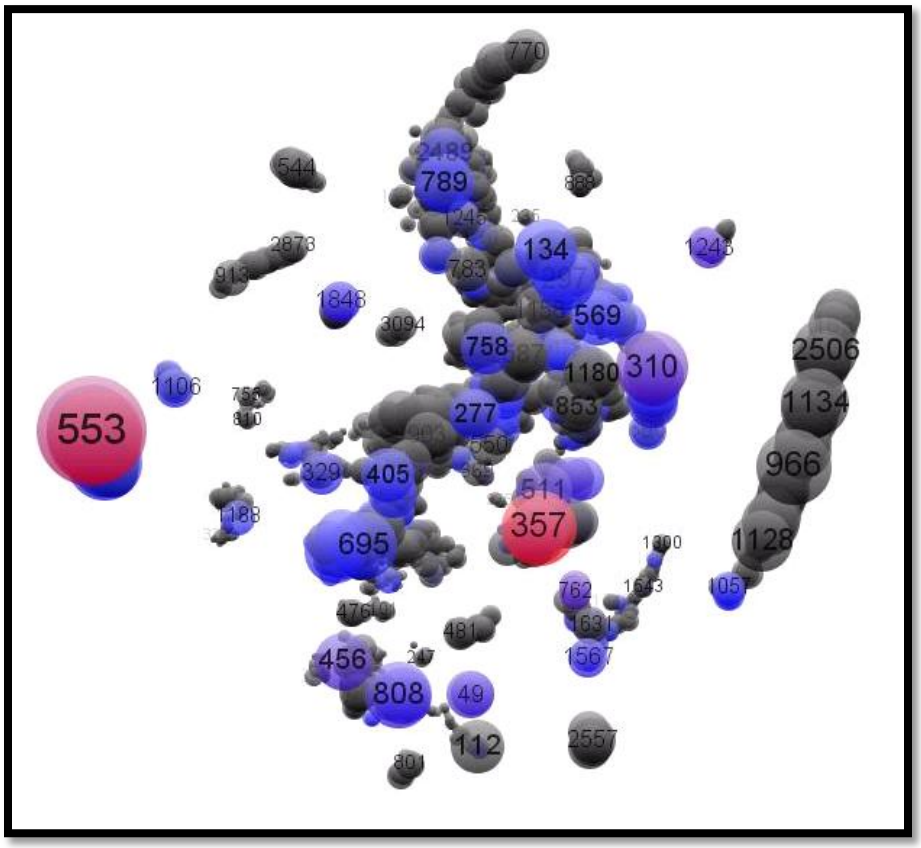


Figure 26: Example of Anomaly Detection supported by Visualization

In this example, each circle represents a provider and the size of the circle represents the activity of this provider in dollars. The red and blue colors are related to the amount of billing errors; red representing the highest number of errors while blue represents the lowest number of errors. The numbers in the circles are arbitrary labels assigned as provider identification numbers (these numbers differ from the real provider IDs). It is possible to quickly glance across the graph and to explore the clusters of providers.

4.4.3 OLAP Cubes for Detecting Phantom Billing

Online Analytical Processing (OLAP) could be applied to analyze the Medicaid Data. If the claims are organized by zip code and provider it is possible to easily identify outliers. For example, if 1,000 claims are submitted in a certain zip code during a particular month and 800 come from one single provider, this means that one provider is responsible for 80% of all the submitted claims. It might be that this provider is really popular, however this might be a case of phantom billing and further investigation is required. A second example is if a certain drug should generally only be prescribed once every three months and an instance arises when the submitted bills indicate, that a single recipient received more than 4 prescriptions per year. Certainly it is possible that in a specific case five, six or seven prescriptions are provided, however 20 times per year is highly suspicious. In an OLAP Cube outliers are easily detectable and an overview of the distribution of this drug can be generated in the form of an OLAP Cube.

In general, Online Analytical Processing is a tool to organize data by state/time/provider/recipients/ type of service/claim/zip codes etc. There is no optimal OLAP cube and the field of application determines which cubes are successful and which are not. This tool can be used to provide analysts with the trends and outliers that exist in the data. The result of the cubes is highly dependent on which variables and defined rules are utilized. Therefore, a process of trial and errors is needed to investigate which variables provide the best result. Applying OLAP cubes will provide advancing insights which should lead to improving fraud detection through the use of trends and outliers.

4.5 Fraud Type IV - VII: Billing Errors / Creative Billing

Creative billing is a miscellaneous section which encompasses the several types of Medical billing fraud schemes. These fraud types are grouped together, since the difference between fraud and abuse is a fine line. In general these are fraud schemes committed by medical providers who actually provide health care services as their main activities. Given the current set up of the Medicaid system tempting opportunities exist for providers to illegally increase their requested billing amount by using a form of creative billing. Instead of billing for one single treatment they find it pays more if they bill Medicaid for separate services. Instead of billing for the actual simple procedure a claim for a more complicated treatment is submitted because the more complex procedure corresponds with a higher reimbursement rate. All of the five fraud schemes are ways for practicing providers to illegally upgrade their submitted bills. Therefore this section of fraud schemes is classified as white collar fraud. If the intention of the provider is not fraudulent their actions are a result of a mistake it is considered to be abuse. The intention is barely measurable and that is embodied in the difficulty of determining the severity of the crime. If a bill was submitted twice by mistake it is considered as abuse, and not as fraud. From the perspective of electronic fraud detection both of the acts are considered to be equally illegal and equally important to be discover.

4.5.1 Fraud Type IV: Duplicate Billing

Billing Medicaid several times for the same service and receiving several reimbursements is considered to be an illegal act. An opportunity to prevent the program from reimbursing a provider several times for the same service is a preprocessing edit. Before the claim is processed the claim history is consulted to see if a similar prior claim exists. According to Sparrow (2002) these so called edits are already in place, however in the national Medicaid database in San Diego similar claims do exist in the data of 2006 in certain states.

This edit may not be applied in all states and therefore there is still room for improvement in order to present a more united front to combat this type of fraud. Similar claims are quite simple to detect in a post payment analysis since the query simply needs to focus on the content of specific fields such as:

- Provider ID
- Beneficiary ID
- Service code
- Date

Duplicate billing should be detected with the prepayment edit and if that edit does not function correctly then the post payment analysis in the form of a query is an adequate solution to detect the fraudulent behavior. Currently the national Medicaid database enables data analysts to use SQL querying for detecting duplicate claims.

4.5.2 Fraud Type V & VI: Bill Padding and Upcoding

Bill padding and upcoding are quite similar fraud schemes. Bill padding is abusing patients in person by giving them treatments that they do not need. Upcoding is (only) misusing the patients on paper by claiming to have provided certain treatments which were not administered at all. In both cases the provider illegally uses the patient's information to bill Medicaid. In the case of bill padding it is quite a challenge for patients (no matter their level of education) to understand what exactly their doctor is doing. There are occasions when patients are not capable of understanding what medical treatments they received since they are unconscious during the procedure. Upcoding is closely related to bill padding from a data perspective since the corresponding of the claims is similar; however it is without the actual execution of the procedure.

Profiling and Peer Group Analysis for Detecting Bill Padding and Upcoding

To provide an overview of the claiming behavior of a provider profiles as presented in section 4.2.1. Discovery of these fraud schemes is a challenge when the patients barely understand what is happening to them even though they are directly involved. The origin of both fraud schemes is providers attempting to upgrade the procedure to receive a higher reimbursement. So if the treatment distribution of providers with similar medical expertise is compared then extensive use of submitting the "expensive" version of the claim will become apparent if the rest of the provider's peers are performing a different (and probably cheaper) treatment. If the procedure is relatively equally provided over the amount of patients compared to the other specialist in the group then apparently this is how often a certain procedure in this specialty and region is provided. In section 4.3.2 an in-depth description of how Peer Group Analysis could be effectively applied in the Medicaid program.

Rule Based Methods for Detecting Bill Padding and Upcoding

If an analyst knows what he is looking for then obtaining the desired results is simply a matter of querying a database. An example of fraud often detected in Medicaid is to bill for services outside of the hospital (outpatient) while the patient was in fact in the hospital (inpatient). This form of upcoding is occurring because the reimbursement rate for outpatient services is higher than for inpatient services. If the patient was in the hospital at the same time that the claim states he was receiving outpatient care then it is a signal that fraud was committed by the provider. When a patient is in the hospital he cannot be charged for services which were said to have taken place outside of the hospital. This is comparable to the

velocity check (outlined in section 3.8.5) which is often applied in the credit card and telecommunications fraud detection.

4.5.3 Fraud Type VII: Unbundling

Instead of billing for one procedure fraudulent providers submit several claims to increase the overall reimbursement amount that they receive. Since the procedure would be spread out over several visits the total amount of visits should be analyzed to verify if it would be possible and or feasible. Sparrow (2002) mentions the example of a particular provider who submitted so many claims to Medicaid that after his behavior was analyzed the conclusion was that this provider had billed for more than 24 hours a day. Furthermore, summarizing statistics captured in a profile could show how often a certain sub procedure occurs. These profiles can be compared to a peer group to show how the distribution is related to the other providers in the peer group.

4.6 The Opportunities and Limitations of Fraud Detection Techniques in the Medicaid Program

After analyzing industries that must combat fraud on a day to day basis, it is apparent that the opportunities afforded by electronic fraud detection techniques are numerous. Electronic fraud detection techniques support the fraud detection process and the dependency of various industries on these systems is significant. There is no universal solution capable of combating all of the existing fraud schemes, however each of the fraud detection techniques have unique features to contribute in the battle against fraud. The potential application of the several electronic fraud detection techniques in Medicaid is discussed in the following three sections.

4.6.1 Classification Techniques

Classification techniques as artificial neural networks, support vector machines and classification trees are widely applied throughout several industries with satisfying results. Especially in the credit card industry supervised classification is one of the main drivers of electronic fraud detection. Due to the accuracy of the trained models supervised approaches are preferred if labeled data is available. The credit card company can label certain data as fraudulent and the availability of real time data makes it possible to immediately update the data as well as the fraud detection system. Due to the popularity of classification techniques in the 1990s a lot of research has been done about neural networks (Bhattacharyya, 2010). The successful application of neural networks in the credit card fraud detection industry is an indication of the potential of applying classification to detect fraud and abuse in the Medicaid system.

Opportunities Classification Techniques

"Support Vector Machines and random forests are sophisticated data mining techniques which have been noted in recent years to show superior performance across different applications" (Bhattacharyya *et al.,* 2010). The advantages of SVMs include their speed and applicability; they do not require a large training data set and are applicable to large data sets. The scalability of SVMs; the classification complexity does not depend on the dimensionality of the feature space (Mukkamala *et al*., 2002). This is supported by Joachims (1998) that SVMs do not suffer from the "curse of dimensionality" so they can learn a potentially larger set of patterns and be able to scale better than neural networks. Downsides include the fact that SVMs are computationally intensive, however compared to neural networks the training and running time is considerably shorter for SVMs (Mukkamala *et al.*, 2002). Furthermore, the results are not easy to interpret, which is comparable to the results of neural networks.

Bhattacharyya *et al.* (2010) use random forests, support vector machines, and logistical regression based data mining methods to distinguish fraudulent from non-fraudulent credit card transactions by using real time data. They show that random forests based methods are able to obtain good overall results by using a variety of performance criteria. Random forests and support vector machines are two complex classification techniques applied in several fraud detection industries with the best results. Currently supervised data mining is currently not applied to detect fraud in Medicaid; therefore a start to develop a more simple technique such as regression analysis is recommended. Fraud detection is a complex process since fraud schemes are sophisticated and hard to recognize. Supervised classification in the form of random forests and support vector machines are able to deal (partly) with the fraud and therefore these techniques are suitable to be applied. Starting with a more simple technique is recommended to gain experience with supervised classification.

Limitations and Requirements of Supervised Classification

For supervised classification, labeled data is required to train the system to recognize the difference between fraudulent and non-fraudulent behavior. Analyzing data with supervised and unsupervised techniques may yield results that were previously unknown. For example, unexpected associations can reveal new relationships and patterns that were previously unknown. It is a challenge to set up a properly trained data set with sufficient data for non-fraudulent and fraudulent transactions. Additionally, fraudulent transactions are scarce since the non-fraudulent transactions generally far outnumber the fraudulent ones (Bolton and Hand, 2002). With such a skewed data set it is a challenge for the classification technique to achieve a high accuracy rating. Neural networks, support vector machines and random forests have been shown to be successful in their early stages (Zhou and Kapoor, 2010; and Bhattacharyya *et al.* 2010). An important limitation of finding fraud with supervised classification techniques is that the techniques are only able to find fraud that has been discovered before. The training set contains the known fraud schemes and unknown forms of fraud cannot be incorporated in the training set. In contrast, unsupervised methods simply seek out those customers who behave in an 'unusual' manner (Bolton and Hand, 2001). Anomaly detection is looking for outliers in the data and therefore new types of fraud can be discovered if it produces an anomaly. There is no perfect technique that covers the whole fraud spectrum and therefore several techniques need to be applied in unison and updated since fraud is continuously changing. The need to be flexible and adaptive is a key aspect in fraud detection and a modular approach is presented by Becket *et al.* (2010) to realize a flexible fraud detection system. The fraud detection techniques are modular components which can be individually updated to enable the system to adapt quickly to the newest fraud trends.

A main problem with classification is that these methods regarding fraud detection suffer from unbalanced class sizes. Criminal or fraudulent transactions are rare and legitimate transactions usually outnumber the fraudulent ones. As a consequence, the reference frame of the fraudulent class is small which often results in mislabeling and misclassification. "Ultimately severe class imbalance imposes fundamental limits to the classifier performance" (Sudjianto et al., 2010). However, the most important constraint of applying supervised classification in the Medicaid program is the absence of labeled data and the lack of opportunity to label the data. It is unknown if a certain claim is fraudulent or not and that is an essential criteria for supervised classification.

4.6.2 Unsupervised Data Mining Techniques

Supervised classification models are trained to recognize the difference between fraudulent and legitimate transactions based upon the past and therefore cannot cope with new fraud schemes.

Opportunities

Looking for patterns and outliers that are unknown for the data analysts is an aspect where data mining can contribute to the field. Unsupervised data mining would be an effective addition to contribute to the fraud detection process. When dealing with unsupervised techniques the analysts basically look for outliers that are significantly different from the majority. Given the tremendous amount of outliers which exist for various reasons the number of false alarms is quite high. Not every outlier is an indication of fraud or abuse and therefore clustering analysis and peer group analysis are useful techniques that focus on smaller clusters to put outliers in perspective. Although establishing clusters may lead to fewer false alarms establishing peer groups is an intensive task.

Limitations

The high amount of false alarms is quite a limitation of anomaly detection; not every outlier or cluster is an indication for fraud. The development of peer group analysis and cluster analysis is a development to reduce the false alarms by putting them in perspective. If a certain outlier is cause by a provider submitting 100 claims on a day an alarm might go off, however if all the providers in the peer group submitted around the same amount of claims it is probably a "normal" procedure in that specialism. Outliers occur and numerous false alarms costs money and resources so false alarms are a severe limitation of outlier detection.

Recommendation to compose Peer Groups

In peer group analysis each cluster is composed out of practitioners with similar submitting patterns. A datacenter is not capable of dividing practitioners into relevant and useful groups that contribute to revealing unusual behavior. A special task force should be established made up of members who are capable of grouping practitioners together in a meaningful way. Furthermore, the task force needs to have the responsibility to form those groups and perform data analysis on the Medicaid data to verify the effectiveness. Concerning the analysis of Sparrow (2002) that analysis not only should take place on a provider level but on higher levels as well (e.g. practitioners working for the same health care center, nursing home, or hospital) to look for patterns on organizational levels. The task force should look at this information as well regarding the establishments of clusters and peer groups.

4.6.3 Statistical Methods

A statistical method such as profiling is a potentially valuable technique to apply. Investigating the cumulative claiming amounts per practitioner in a certain time frame can offer valuable insights. Furthermore, summarizing at different levels is a challenging opportunity; vetting the claiming behavior of practitioners working in the same health care facility might lead to interesting patterns and insights. However every transaction that differs from the established profile does not guarantee fraud occurred. In the telecommunications industry it occurs that an exceptional phone call is made to a foreign country, this is applicable for the health care as well. From time to time an exceptional claim might be submitted and similar to unsupervised data mining techniques, exceptions (outliers) exists and false alarms are a limitation of these techniques. Profiling has been successfully applied in credit card and

telecommunication fraud detection and it can be a useful addition to the current rule-based fraud detection system.

4.6.4 Rule-based Techniques

Rule-based approach is useful technique to verify if easily detectable fraud and abuse, however it is crucial to be aware that rule-based techniques are rather limited in their detection capability. Intelligent fraud schemes are not detectable with a rule-based approach since these schemes mimic legitimate behavior and are hard to discover. Nevertheless, rule-based approaches have proven to be useful, especially in combination with other techniques (e.g. profiling), and are an important component of the Medicaid fraud detection system.

4.7 Proposal to Verify the Effectiveness of Fraud Detection Techniques

Concerning this research, the applicability and effectiveness of the fraud detection techniques are discussed in theory. The main reason for this is that there is not an opportunity or plausible way to actually apply the techniques given current constraints. The following table is a possible matching scheme to verify the applicability and effectiveness of various detection techniques for the several Medicaid fraud schemes.

	Identity Theft	Fictitious Practitioners	Duplicate billing	Phantom Billing	Bill padding	Up-coding	Un-bundling
Supervised Classification							
Anomaly Detection							
Profiling							
Clustering Analysis							
Peer Group Analysis							
OLAP Cube							
Visualization							
Benford's law							
Rule Based Querying							

Table 10: Feasibility of fraud detection techniques for Medicaid fraud

Legend:	
+	Effective
+/-	Neutral
-	Not Effective

Based upon the systematic literature research and the presented Medicaid fraud schemes the applicability of the various fraud detection techniques is discussed in the sections 4.2 – 4.5. However, further research is required to empirically test the techniques on the Medicaid data to obtain quantitative proof about the

effectiveness of the various techniques. This matrix is a proposal for a matching scheme to provide an overview of the current fraud schemes in Medicaid and the available fraud detection techniques in related industries. The fraud detection techniques are not tested on real Medicaid data to verify the effectiveness of the techniques. With the use of a sample data set including several types of fraud the fraud detection techniques can be empirically tested. The effectiveness can be determined by measuring the percentage of correctly detected fraudulent cases and how many transactions have been flagged incorrectly. Furthermore, the robustness, the data accuracy, speed, and the interpretability of the various detection methods are important components to take into account.

If the detection techniques are tested and proven to be applicable and effective, then small pilots need to be set up for further testing. For example, a certain state could function as a test case for the rest of the U.S. and successes would ideally then be shared and implemented in other states. The unique legislations per state need to be taken into account as well as the differences in data sets (fields etc). This is a comprehensive project where many stakeholders need to collaborate. However the potential of electronic fraud detection techniques and the financial estimations of the losses due to fraud and abuse are both too evident to ignore.

5. Case study: Medicaid Integrity Program

'Food gained by fraud tastes sweet to a man, but he ends up with a mouth full of gravel.'

- Proverbs 20:17 (New International Version)

Given the problematic situation in the health care industry in the United States (see Chapter 2) the federal government started up the Medicaid Integrity Program (MIP), a project initiated in 2005 to improve fraud detection in the U.S health care program Medicaid. The Centers for Medicare and Medicaid (CMS) are responsible for the program which is based on post-payment analysis of the Medicaid data. The fraud detection process is a review of the data to verify that all of the submitted Medicaid claims are legitimate and billed correctly. It is a reactive process which takes place after the actual reimbursement of the providers. This post-payment analysis strives to discover irregularities and mistakes in the processing and payment process and provide a framework for audits and collections of these overpayments. To perform this master's thesis, a case study at the San Diego Supercomputer Center was conducted. The findings of this case study are presented in this chapter, and some recommendations are given according to the literature study described previously.

5.1 Medicaid Statistical Information System Data

CMS stores the Medicaid Statistical Information System (MSIS data) in the national Medicaid database to support CMS fraud, waste, and abuse analysis efforts and to assist the data analysts to perform post payment provider analysis and audits. The nationwide Medicaid database is located at the University of California San Diego's San Diego Supercomputer Center (SDSC). SDSC is responsible for making the data available for data analysts of CMS as well as for the related contractors. The CMS requires that all the states in the U.S. send their data sets in a standardized format known as MSIS. To improve the timeliness of the MSIS files' release, CMS encourages the states to submit their files electronically instead of via magnetic tapes and magnetic cartridges, as was done in the past. As of July 2009, 34 States were sending their MSIS files electronically (OIG report, 2009), while 16 states were still delivering their data on magnetic tapes. The CMS is collecting the MSIS data from States to establish an accurate and comprehensive database containing standardized enrollment, eligibility, and paid claims statistics to be used for administration at the federal level.

Alongside hosting the Medicaid data, several information technology applications are developed at the SDSC to improve data analysis and to support the process of fraud detection. The opportunities and benefits of IT applications, such as business intelligence and workflow management systems, are important in the battle against fraud. The business intelligence tool IBM Cognos is now in place to address the CMS needs to "produce statistical reports, support Medicaid-related research, and assist in the detection of fraud and abuse in the Medicare and Medicaid programs" (OIG report, 2009). The workflow management system, based in Oracle's Siebel product, is currently implemented to support the process of fraud waste and abuse detection and the recovery of money which has been lost to fraud and abuse.

Applied SQL Queries in the Hospice Care at SDSC

By using Structured Query Language (SQL) information can be obtained from the database if you know what you are looking for. Basically, querying is making use of expert knowledge; the analysts are field experts and know what they are looking for. If the experts find something by querying they can be rather sure that it is actual fraud because they are looking for patterns which violate with the restrictions. Rule based SQL querying involves developing an algorithm that looks for fraudulent and abusive activity. If a provider is detected with the algorithm then the analysts know it is behavior which is not allowed. Whether the flagged action was caused by accident or was intentionally instigated is not always verifiable and therefore further investigated is needed.

In a SDSC industry study to find fraud and abuse in the U.S. Medicaid hospice care several queries are developed and presented. For the period between January 2004 and December 2008, an initial analysis was performed to determine potential suspicious hospice care providers filing claims which were:

1. Top 20 hospice providers where at least one quarterly difference in total hospice billing is above $50K, or at least one quarterly difference is 50% greater than the previous quarter by total amount paid by Medicaid.
2. Top 20 hospice providers with beneficiaries having lengths of stay (LOS) greater than six months.
3. Top 20 diagnoses codes by total amount paid by Medicaid; includes daily average based on service dates and daily average based on quantity of service billed.
4. Top 20 providers who did not have any inpatient claims and who also did not have any outpatient claims totaling over $10K within sixty days of the beneficiary being admitted for hospice care.
5. Top 20 hospice providers whose beneficiaries had suspicious Local Coverage Determinations (LCD) diagnostic codes.
6. Top 20 hospice providers with the lowest mortality rates within six months of hospice care.
7. Top 20 hospice providers by number of prescription and Durable Medical Equipment (DME) claims in prescription claims.
8. Top 20 hospice providers who billed a single beneficiary on the same day.
9. Top 5 providers by total amount paid by Medicaid; total Medicaid amount paid by Medicaid per provider and per provider's top 5 diagnostic codes detailed.

The second query is based on the rule that a patient should not be enrolled in a hospice care center for longer than six months. Hospice care is designed for terminal patients and if a patient lives longer than six months then they should not be considered terminal. There are known cases of fraudulent schemes in the hospice patients were billed for more than a year. This example illustrates the fact that background knowledge is required to understand why specific algorithms are developed.

Consequences of the Rule-Based Approach

Using SQL querying based on expert knowledge is one way to detect health care fraud and abuse, however, what about fraud and abuse the analysts do not have any knowledge about? With querying obvious fraud schemes as unbundling and duplicate billing are detectable and it is not a complex procedure. Given the obvious feature of these schemes the results of the queries can be labeled as overpayment with a significant certainty. The data analysts know what they are looking for and if an overpayment is discovered the fraudulent providers will be questioned for verification. So the advantage

is that is a rather simple and effective technique to detect fraud and abuse. The disadvantage is that the analysts only look for obvious fraud schemes they know about and that that fraud is detectable by using the current MSIS data.

Supervised and unsupervised data mining would offer a great deal of new insights; data mining is the automatic extraction of new high-level information (knowledge) from low-level data (Fayyad *et al.*, 1996). Data mining offers new results, relations, and knowledge that might lead to new insights. However, the downsides of supervised data mining include; the complexity of the techniques, the requirement of labeled data, and the interpretation of the results. In the Medicaid Integrity Program the requirement for supervised data mining, along with the related challenges and consequences, is outlined in the next section.

5.2 Labeled Data: Closing the Loop

The most important restriction that limits the use of supervised methods in the Medicaid Integrity Program is the lack of labeled data available. To overcome this problem feedback information is required. The MIP will generate some of this labeling itself through the audits performed on suspect providers and claims. This will, for the first time, allow for labeled data. Labeling the data would improve the accuracy of the data since more information would then be known about the transactions. Even if the data is partially labeled (e.g. 'under investigation'), a start could be made to develop and apply supervised learning techniques.

Conflict of Interest

Another point of interest noticed during this masters research is the inherent conflict of interests between the state and the federal government. As outlined in section 1.2, the federal government partially funds the Medicaid program while the states are responsible for detecting fraudulent and abusive transactions. However if a state labels or marks a certain transaction as fraudulent, the federal government demands its part of the recovered fraud within 90 days of detection, because Medicaid is partially funded by federal dollars. Therefore, states are reluctant to label a transaction as fraud since they must then bear the burden of recovering the actual money and repaying the federal government in a short timeframe. There is a big difference between recognizing fraud and actually being reimbursed. According to Sparrow (2002) often a fraudulent provider disappears, the money is gone, and the federal government will demand its share back from the state. Therefore, states have the incentive to label all the fraudulent and abusive transactions as "overpayments"; which, under current legislation, provides them a greater time window to seek out repayment from providers before having to repay the federal government its share. "Overpayment" is a term used to define a mistake made by the provider or the transaction system, or, in other words, a billing error.

5.3 Siebel Workflow Management Implementation

Evaluation feedback is a complex process due to the multiple stakeholders in the Medicaid health care program. The federal government oversees the states while every state regulates Medicaid independently and employs different Medicaid contractors who review and analyze the data for fraudulent and abusive behavior. Every state submits summary MSIS data sets quarterly to the Center of Medicaid and Medicare (CMS), and the data is then hosted at SDSC. With the implementation of a workflow management system, the MIP processes will be automated and collaboration between stakeholders (CMS, states, contractors, and providers) will be supported and simplified. The process and data is standardized and

centrally stored so that all of the stakeholders have appropriate access, explained in detail in the next section. When dealing with health care data, privacy is a critical aspect and all sensitive data must be well secured from outsiders. The workflow management system operates as a secure central information source which can be well controlled with concerns to privacy and visibility.

Transferring the historic claim data into useful information by analyzing the data is an important step in the process. However, the information must be available to the right and responsible people at the right time. Information technology is an enabler to support the business process and improve the information for the involved stakeholders (see Appendix A). The goal of the implementation is to improve the fraud detection process in order to reduce the time it takes from responding to actually recovering the funds.

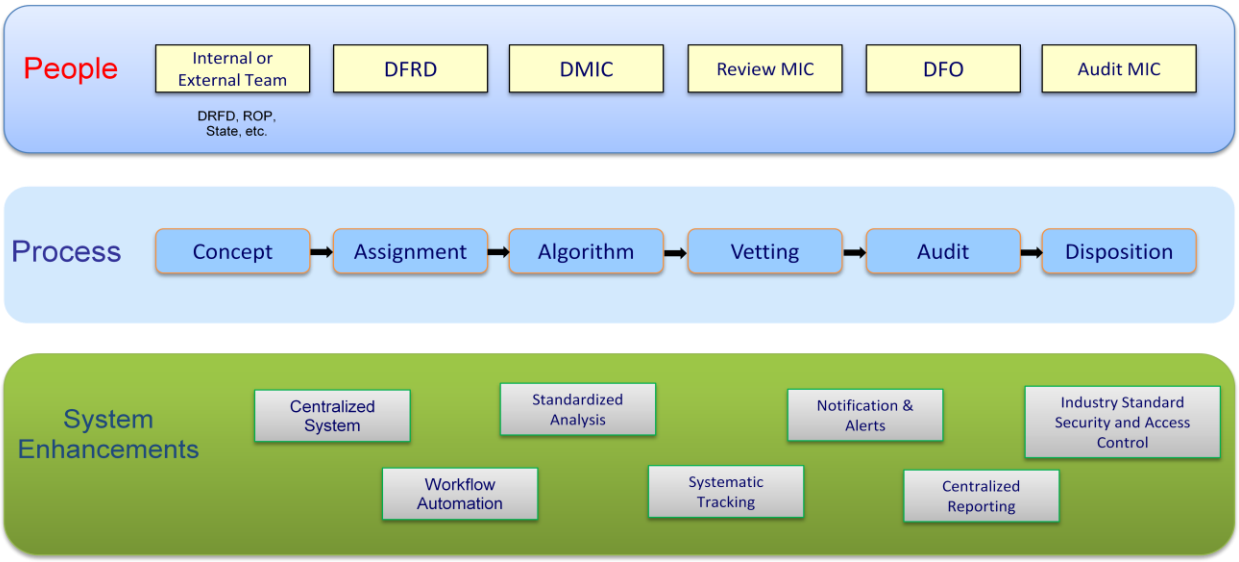


Figure 27: Fraud Detection Process Medicaid including stakeholders and system functionality (SDSC, 2010)

The Siebel workflow management is a standardized and centralized system that enables the stakeholders to systematically track the progress of the investigations of providers. The investigation process which goes after potential fraudsters is an important step in a fraud detection system. Once the fraud is detected the perpetrators need to be brought to justice and the funds must be recovered (see Appendix B). The workflow management system supports the process and keeps track of all the changes and results. Once the result of an investigation is known then the system stores the information and the SDSC can make use of this information later. Labeling fraudulent transaction yields a knowledge database with enormous potential to contribute to further fraud detection techniques. Supervised classification techniques can be developed and applied to the dataset to verify if the new labels lead to improved fraud detection. A simple classification technique, such as linear regression, can be a first step in the direction of supervised techniques.

6. Discussion

'It is not the strongest of the species that survives, nor the most intelligent that survives. It is the one that is the most adaptable to change.'

- Charles Robert Darwin (English Naturalist, 1809 – 1882)

This chapter formulates the conclusion (§ 6.1) and the limitations of the fraud detection analysis (§ 6.2). Based on the performed literature study conclusions are drawn and the implications of the findings are presented. Subsequently the identified limitations are discussed and this research is finalized by giving suggestions and recommendations for further research (§ 6.3).

6.1 Conclusion

A description of the current situation surrounding the U.S. health care program Medicaid coupled with its corresponding problems is provided below. Furthermore, an analysis of the fraud detection domains in various industries and the characteristics of fraud and the detection of fraud are presented. By analyzing diverse industries that successfully utilize electronic fraud detection techniques, a framework of electronic fraud detection techniques emerged. An additional case study at the national Medicaid database in San Diego is included to complement the literature study. The main research question was:

> **How do electronic fraud detection systems facilitate security in other industries and how can the fraud detection of the U.S. health care program Medicaid be improved?**

The conclusion of this thesis answers the sub questions as formulated in section 1.4; subsequently the implications of the findings are discussed. Every section concerns a sub questions and all sections contribute to answer the main research question.

6.1.1 Critical Aspects of the Medicaid Health Care Program

The U.S. health care program Medicaid is outlined in chapter 2 and the current situation is described with its corresponding problems. Due to the magnitude and the current setup of Medicaid several fraud types are detected and described in detail. The sub question regarding the Medicaid program is formulated as following:

> *What are critical aspects of the U.S. health care program Medicaid and what is the current situation regarding fraud and fraud detection?*

In the Medicaid system beneficiaries do not automatically report fraud and abuse committed on their Medicaid accounts, unlike the situation in the credit card and telecommunications where customers report misuse almost instantly. Critical aspects of the Medicaid program include the involvement and incentives of beneficiaries to contribute to detect fraud in the program. Because the government is the party paying for the treatments, the beneficiaries do not pay enough attention to notice if the EOBs match the treatments they received. Verifying the EOB is also impeded due to the computer generated codes and the names of services which often include complicated medical terms unintelligible to most healthcare consumers. As Sparrow (2002) mentioned in his analysis, even when beneficiaries reported forms of fraud the government was not able to go after the fraudster. Due to the magnitude of the program (49

million people were enrolled with a yearly budget of $321 billion dollar in 2008) the amount of fraudulent reports might easily become overwhelming if there are not enough resources devoted investigating the suspicious reports.

Automating the claim process has been a critical decision in the development of fraud in the Medicaid program. Since there are not any people involved in the processing and verifying of the truthfulness of the claims it has been quite easy for criminals to drain the system (Sparrow, 2002). The program is governed by the state governments resulting in numerous different processing systems and unique legislation in every state. This increases the complexity of the problem since there cannot be one generally applicable solution. Due to the magnitude of the program and the various legislation protocols many stakeholders (with divergent interests) are involved and have to work together efficiently to effectively fight fraud and abuse. Given the fraud schemes that have been discovered in the literature study, the fraud detection systems currently in place are suboptimal. Seven identified fraud and abuse types are shown in the following table:

Fraud Type	Fraud Scheme	Short Explanation	Strategy
I	**Identity Theft**	Stealing identification information from providers and beneficiaries and using that information to submit fraudulent bills to Medicaid.	Fraud
II	**Fictitious Practitioners**	Using false documents and identification information to submit fraudulent bills to Medicaid.	Fraud
III	**Phantom Billing**	Submitted claims for services not provided.	Fraud
IV	**Duplicate Billing**	Submitting similar claims more than once.	Fraud/ Abuse
V	**Bill Padding**	Providing unnecessary services and the submitting these claims to Medicaid.	Fraud/ Abuse
VI	**Upcoding**	Billing for a service with a higher reimbursement rate than the service which was actually provided.	Fraud/ Abuse
VII	**Unbundling**	Submitting several claims for various services that should only be billed as one master claim that includes ancillary services	Fraud/ Abuse

Table 11: Medical Fraud Schemes

The current situation is worrisome, and the estimation of Kelly (2009) is that $700 billion dollars is lost due to fraud waste and abuse in the whole U.S. health care industry. Medicaid is responsible for a significant amount of this estimation (as well as Medicare); clearly an effective Medicaid fraud detection system needs to be developed.

6.1.2 Electronic Fraud Detection Systems

A systematic literature study has been conducted to identify the applied electronic fraud detection techniques for the purpose of fraud detection. The credit card and telecommunications fraud detection systems have been examined even as the computer intrusion detection systems. The sub questions have been defined as following:

What are the important aspects and requirements for an effective electronic fraud detection system?

Fraudster will always try to find new ways to avoid the fraud detection systems so that they can continue to commit fraud and stay undetected. Therefore, several fraud schemes exist as presented in table 11, however there are certainly more fraud schemes which remain undetected and attack the U.S. health care industry in the hopes of draining the system. Based on the conducted literature study a fraud detection system should consist of:

- Fast response time
- Routine measurements
- Flexible modular approach
- Random audits

Fraud detection is a dynamic process and the fraud detection team must constantly react to the moves made by the fraudsters. It is essential that the reaction is as fast as possible in order to limit the losses. If the reaction time proves to be too slow then retrieving the funds is often impossible since the fraudster had enough time to disappear with the money. Routine measurement is needed to monitor the situation whether the fraud detection process is successful or not. The possibility that an increase in the total amount of committed fraud took place should be taken into account while measuring. A decrease in the fraudulent cases detected does not imply an actual decrease in the amount of fraud committed since fraud is constantly changing. Therefore, a flexible approach is required to keep the fraud detection system up to date for the latest fraud schemes. A modular fraud detection system is capable of updating parts of the system so that the whole general system is not outdated and unable to detect new fraud schemes. Random audits are required to create the impression that every submitted claim risks a thorough inspection to verify if the service was actually provided as presented. Sparrow (2002) describes that the Medicaid processing system is developed for honest providers who are notified and informed when something went wrong with the submitting process. Processing claims should be a suspicious matter since sometimes human beings display opportunistic behavior (Williamsons, 1985). The Medicaid processing systems need to be adjusted so that all of the states are able to increase security without causing significantly more trouble for honest practitioners. This is a problem comparable with the customs line at airports, not every traveler need to open their suitcases to prove their innocence; however the people with criminal intentions need to be detected without causing trouble for other passengers.

The dependence of Medicaid on electronic fraud detection is compared with the related industries quite high and the opportunities available are not currently fully utilized. The lack of labeled real time data is a crucial aspect that determines the situation in the Medicaid program and therefore makes the Medicaid fraud detection a reactive process. Nevertheless, there is still a lot of unused potential regarding electronic fraud detection; statistical methods and data mining techniques are especially potential areas in which to improve the Medicaid fraud detection system. When credit card customers become a victim of fraud they will immediately report that fraud to the credit card company because it is their own money that is stolen. Therefore electronic fraud detection is an important component for the Medicaid fraud detection and the lack of expertise and resources to develop effective data mining tools to support the electronic fraud detection process. Data mining has been proven to be effective in fraud detection settings, and the credit card industry should be an inspiring example of the implicit potential. However, given the magnitude of the program (50 states) the multiple involved stakeholders, fragmented responsibility, lack of labeled data, and the lack of real time data available to accurately monitor the problem and respond are all severe limitations. As stated in section 2.7 monitoring and controlling and measuring the fraud and abuse in the

program is essential to control the situation and make informed decisions. Electronic fraud detection techniques have been proven to be effective in other fraud detection industries and can positively contribute to safeguarding the Medicaid program.

6.1.3 Electronic Fraud Detection Techniques

The fraud detection techniques which have been applied in other fraud detection industries are also applicable in the Medicaid setting. A result of this research is the following framework which has been constructed after evaluating the findings from the various fraud detection domains. The framework contains the answer regarding the sub question about fraud detection techniques:

What important fraud detection techniques from other industries can be applied in the fraud detection of the Medicaid program?

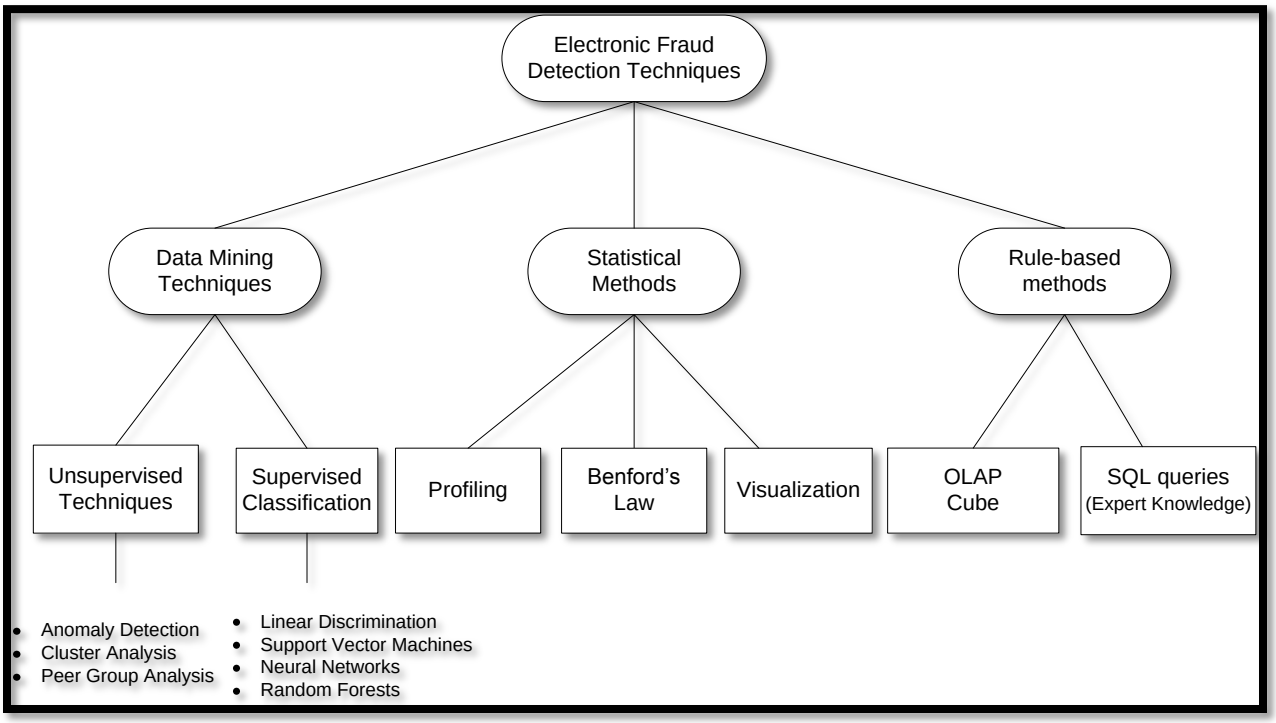


Figure 28: Framework of Electronic Fraud Detection Techniques

Data Mining Techniques

Data mining techniques are clearly separated into supervised and unsupervised techniques; where supervised models are trained with sample data to recognize the difference between fraudulent and non-fraudulent transactions, while unsupervised techniques detect outliers and anomalies. As outlined in section 3.6 and discussed in section 4.6, random forests and support vector machines are two complex classification techniques applied in several fraud detection industries with the best results. Fraud detection is a complex process since fraud schemes are sophisticated and hard to recognize. Supervised classification in the form of random forests and support vector machines are able to deal (partly) with the fraud problem and these two techniques are considered in the academic literature to be suitable and effective. Although starting with a simpler technique (e.g. logistic regression) is recommended to gain experience with supervised classification. Supervised classification models are trained to recognize the

difference between fraudulent and legitimate transactions based upon the past and therefore cannot cope with new fraud schemes. Unsupervised data mining would be an effective addition to contribute to the fraud detection process to detect new fraud schemes in the form of anomaly detection. Special forms of anomaly detection are clustering analysis and peer-group analysis, both of which put outliers into perspective which reduces false alarms and allows the entire process to operate much more smoothly.

Statistical Methods and Rule-Based Methods

Profiling is a widely applied tool in the relevant fraud detection industries used to generate a signature of the behavior of customers. If transactions are detected which significantly deviate from the profile, alarms are raised in order to enable a fraud analyst to further investigate the situation. Investigating the cumulative claiming amounts per person in a certain time frame can offer valuable insights. While summarizing and enumerating at different levels is a challenging opportunity, vetting the claiming behavior of practitioners working in the same health care facility may lead to interesting patterns and insights. Although deviations of the profiles do not automatically imply that fraud occurred and computing profiles for all providers and beneficiaries is a computational intensive project. If these downsides limit the project need to be further researched.

Applying rule-based methods, such as querying a database, is a useful technique in order to discover fraudulent transactions. With the support of Online Analytical Processing, a tool used to organize the claim data by state/time/provider/recipients/type of service/type of claim/zip codes etc., analysts are able to verify outliers or abnormal behavior. The detection capability of rule based approaches and profiling is limited to the understanding of the human mind. Data mining might offer new insights by presenting associations and relations never previously analyzed, however it is important to understand that no technique is capable of detecting all the fraud schemes. As a consequence, a flexible modular fraud detection system consisting of several techniques which all contribute to protect the Medicaid program is required.

6.2 Limitations

A major limitation of this research is its theoretical approach; a systematic literature study has been conducted to provide an overview of the current worrisome situation in Medicaid and what electronic fraud detection techniques exists in related fraud detection domains. The published fraud frameworks are limited since fraudsters would benefit from being able to easily access the information and would undoubtedly attempt to use that sensitive information to defraud the system. For that reason a literature study is not the ideal way to investigate fraud detection systems; however it does provide a proper first impression and overview of the existing fraud schemes and detection techniques currently applied. The high number of stakeholders, 50 states with unique legislation and eligibility rules, and the magnitude of the program's budget complicate Medicaid fraud detection. As Sparrow (2002) explicitly explains, fraud should be properly measured to create a realistic impression is created of the current situation and an estimation of the amount of fraud and abuse in Medicaid (and Medicare). The estimation of $700 billion lost to fraud and abuse annually in the U.S. health care industry (not only Medicaid) is the most reliable estimation found in the literature today; however politicians would never agree with this extremely high amount. To make a realistic estimation of the fraud and abuse present a proper measurement method should be applied which both the industry as well as the government can agree upon. Once that is in place the amount of fraud can be reduced to a level that is generally accepted. Although fraud can never be

completely eradicated, it can be better managed, however many systematic improvements need to be realized.

Concerning supervised data mining such as classification, labeled data is required to train the system to recognize the difference between fraudulent and non-fraudulent behavior. This training is especially important due to the fact that fraudulent transactions are scarce and that the non-fraudulent transactions generally far outnumber the fraudulent ones (Bolton and Hand, 2002). With such a skewed data set it is a challenge for the classification technique to realize a high level of accuracy. As a consequence the reference frame of the fraudulent class is small and consequently mislabeling and misclassification often occurs. However, the most important constraint of applying supervised classification in the Medicaid program is the absence of labeled data and the opportunity to label the data. Currently, it is unknown if a certain claim is fraudulent or not and that is an essential criteria for supervised classification. Given the fact that in the related fraud detection industries the supervised classification was the most accurate technique to point out fraud, is the lack of labeled data a major limitation.

6.3 Recommendations for Further Research

An important next step is empirical testing of the proposed fraud detection techniques on (Medicaid) datasets in order to quantify the results. Assuming all the stakeholders in the Medicaid case are collaborating, small pilots need to be established to empirically test the applicability and the effectiveness of the techniques. Subsequently some mix need to be designed to verify is a mix is an improvement compared to a single technique. According to this systematic literature research, no technique exists that detects all types of fraud. Thus a modular mix of techniques needs to be developed to achieve better results. However, to verify if the results are actually better, empirical testing must be performed. Evaluation criteria should include the detection rate, effort, interpretability, and the costs versus the benefits. The corresponding costs of developing a fraud detection system will have to be compensated by the resulting benefits of the fraud detection system. What are the strengths and weaknesses of the particular system and how can the overall performance be improved? Applying fraud detection techniques in practice (e.g. a pilot with one of the states) could help determine the effectiveness of these approaches.

Further research need to be conducted in Medicaid fraud detection, including case studies and interviews to provide a deeper understanding of the current Medicaid fraud situation. For example, conducting in-depth case studies at Medicaid agencies in several states will provide a better understanding of the actual effectiveness of the operations and pre-payment tools utilized. Furthermore interviews with fraud analysts in the credit card or telecommunications industries could generate more detailed and descriptive information about the actual effective implementation of fraud detection systems. While systematic literature review is not suitable to share sensitive information, cooperation projects between the government and private companies can lead to significant improvements in the fraud detection systems. The U.S. government should benefit from the experience and expertise of the credit card and telecommunications industries to combat fraud and abuse. Publishing these results is not recommended, however the information should be shared between states to improve fraud detection capabilities across the Medicaid program.

Incentive alignment

Incentive analysis and incentive alignment (Ba et al., 2001) is an interesting domain to further investigate. How are the many government agencies working together to fight fraud and what are the threats and

opportunities for improvement? What are the incentives of the stakeholders and are these incentives aligned or do conflicts of interest exist and how can they be solved? The goal should be a proactive system that prevents fraudsters committing fraud in the Medicaid program. Currently the fraud detection systems are reactive; data analysis of historic claim data analyzes claiming behavior and provides insight about the current practice. Information about fraudulent behavior should be applied to prevent fraud and abuse from occurring in the first place.

Social networks and Customer Tools

Other ideas to further investigate include the use of social media networks such as Facebook, LinkedIn, or Twitter as an input for provider and beneficiary profiling and link analysis. The Medicaid program is victim of organized crime and this may be a way to detect interconnections between individuals which form a professional network. Separately, the government could collect feedback from beneficiaries on whether services were properly billed via a web application where beneficiaries could verify their Explanation of Benefits (EOBs), including consumer-readable descriptions of services billed. As previously stated, a support infrastructure would need to be implemented to follow-up with reports of problems by beneficiaries, but this would be a tremendous source of leads to follow.

Opportunities

As described above there are numerous opportunities to improve the Medicaid fraud detection system and I would like to conclude with some positive notes. The opportunities of fraud detection are endless and the Medicaid stakeholders need to work together to enable electronic fraud detection. It is needed to establish data sets and pilots in order to test the applicability and effectiveness of the fraud detection techniques and successes need to be shared. Information technology has plenty of opportunities to improve and support the fraud detection such as data mining, statistical methods and network analysis by using social networks. When the stakeholders work together (with aligned incentives), I am absolutely positive the war against fraud can be won and I hope this research will contribute in that battle.

Reference List

Akerlof , G.A. (1970), "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics*, Vol. 84 (3) pp. 488-500

Anderson, R., Bond, M. and Murdoch, S.J. (2006), "Chip and Spin," *Computer Security Journal, Vol.* 22(6), pp. 1 - 6

Ba, S., Stalleart, J., and Whinston, A.B. (2001), "Research Commentary: Introducing a Third Dimension in Information Systems Design – the case for incentive alignment," *Information System Research 12(3),* pp. 225 - 239

Becker, G.S. (1968), "Crime and Punishment: an Economic Analysis," *76 J. Politics and Economics 169*

Becker, D., Kessler, D., and McClellan, M. (2005), "Detecting Medicare abuse," *Journal of Health Economics24 (2005)* pp. 189 - 210

Becker, R. A., Volinsky, C., and Wilks, A. R. (2010), "Fraud Detection in Telecommunications: A Historical Perspective and Lessons Learned," *Technometrics 52,* pp. 20–33

Benbassat, I., Dexter A. S., and Todd. P. (1986), "An experimental program investigation color-enhanced and graphical information presentation; an integration of the findings," *Communication of the ACM 29(11),* pp. 1094 – 1105

Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J.C. (2010), Data mining for credit card fraud: A comparative study", *Decision Support Systems 2010,* pp. 1 - 12

Bolton, R.J. and Hand, D.J. (2001), "Unsupervised profiling Methods for Fraud Detection," *Technical Report, department of Mathematics, Imperial College, London*

Bolton, R.J. and Hand, D.J. (2001), "Peer Group Analysis – Local Anomaly Detection in Longitudinal Data," *Technical Report, department of Mathematics, Imperial College, London*

Bolton, R.J. and Hand, D.J. (2002), "Statistical Fraud Detection: A Review," *Statistical Science 17 (3),* pp. 235-255

Boyle, F., Aytug, H., Koehler, G.J. (2007), "Systems for strategic learning," article is part of the "Handbook on Decision Support Systems, *Springer,* pp. 206 – 220

Breiman, L. (2001), "Random forest," *Machine learning (45),* pp. 1 – 32

Cahill, M.H., Lambert, D., Pinheiro, J.C., and Sun, D.X. (2002), "Detecting Fraud in the Real World," *Handbook of massive data sets book contents*, pp. 911 - 929

Caruana, R. and Niculescu-Mizil, A., (2006), "An Empirical Comparison of Supervised Learning Algorithms," in *Proceedings of the 23rd International conference on Machine Learning Pittsburg, PA,* pp. 161 – 168

Chan, P. K. and Fan, W., (1999), "Distributed Data Mining in Credit Card Fraud Detection," *IEEE November/December 1999,* pp. 67-74

Chandola, V., Banerjee, A., Kumar, V., (2009), "Anomaly Detection: A Survey," *ACM Computing Surveys,* pp. 1 – 72

Chapterhouse, LLC (2008), Medicaid Overview – Size and Segments; www.chapterhouse.com/Healthcare_Innovation_Submenus/PDF%20Files/Medicaid%20Overview-Size%20and%20Segments.pdf

Centers for Medicaid and Medicare (2010), Retrieved 9.6.2010, from "www.cms.gov,"

Codd, E.F., Codd, S.B., and Salley, C.T. (1993), "Providing OLAP (Online Analytical Processing) to User-Analysts: An IT Mandate", *Codd and Date, Inc.* retrieved at 23 November 2010.

Cortes, C. and Pregibon, D. (1998). Giga-mining," *In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining,* pp. 174-178.

Center for Regulatory Effectiveness (2010); http://www.thecre.com/fedlaw/legal17/medicaid.htm

Davis, F. D., R.P. Bagozzi, Wrshaw P. R. (1989), "User Acceptance of computer technology: A comparison of two theoretical models," *Management Science 35(8),* pp. 982 – 1003

Davis, A.B., and Goyal, S.K. (1993), "Knowledge-Based management of cellular clone fraud," *Proceedings PIMRC '92*, pp. 230 - 234

Davis, K., Schoen, C., and Stremikis, K. (2010), "Mirror, Mirror on the Wall – How the performance of the U.S. Health Care System Compares Internationally 2010 Update, the Commonwealth Fund

Derrig, R. A. (2002), "Insurance Fraud," *The journal of Risk and Insurance 69(3),* pp. 271-287

Dianmin Y., Xiaodan, W., Yunfeng, W., Yue, L., and Chao-Hsien, C. (2007), "A review of data mining-based financial fraud detection research," *International Conference on Wireless Communication, Networking and Mobile Computing 2007,* pp. 5519 – 5522.

Dorronsoro, J.R., Ginel, F., Sanches, C., and Cruz, C.S., (1997), "Neural Fraud Detection in Credit Card Operations," *IEEE Transaction on Neural Networks, Vol 8 (4),* pp. 827 - 834

Douma, S. and Schreuder, H. (2002), Book: Economic Approaches to Organizations, Chapter 4, 7 & 8

Eisenhardt, K. M. (1989), "Agency Theory: An Assessment and Review," *Academy of Management Review 14(1),* pp. 57-74

Fadlalla, A. and Lin, C-H., (2001), "An Analysis of the Applications of Neural Networks in Finance," *Interfaces, Vol 31 (4)* pp. 112-122

Fawcett, T. and Provost, F. (1997), "Adaptive Fraud Detection," *Data Mining and Knowledge Discovery 1,* pp. 291-316

Fawcett, T. and Provost, F. (1999), "Activity monitoring: Noticing interesting changes in behavior," *Association for Computing Machinery*

Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy (1996), "Advances in Knowledge Discovery and Data Mining," *AAAI/MIT Press 1996 Chapter 1*

Federal Bureau of Investigations (2010) - FBI.gov: http://dallas.fbi.gov/dojpressrel/pressrel09/dl021209.htm, retrieved at 28th September 2010

Ferreira, P., Alves, R., Belo, O., and Cortesao, L. (2006), "Establishing Fraud Detection Patterns Based on Signatures," *New trends in Artificial Intelligence, 13th Portuguese Conference on Artificial Intelligence EPIA 2007,* pp. 428 - 440

Frieden, J. (1991), "Fraud squads target suspect claims," *Business & Health 9(4),* pp. 21-33

Freund, Y., and Schapire, R. (1997), "A Decision-Theoretic Generalization of Online Learning and Application to Boosting," *Journal of Computer and System Science 55*, pp. 119 - 139

Furlan, S, and Bajec, M. (2008), "Holistic Approach to Fraud Management in Health Insurance," *Journal of Information and Organizational Sciencesvol. 32 (2),* pp. 99- 114

GAO, United States General Accounting Office (2000), "Health Care Fraud: Schemes to defraud Medicare, Medicaid, and private health care insurers," June 25 2000

Goldberg, H. G., and Senator, T. E. (1995), "Restructuring Databases for Knowledge Discovery by Consolidation and Link Analysis," in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95),* Menlo Park, CA: AAAI Press, pp. 136-141

Goldberg, H. G., and Wong, R. W. H. (1998), "Restructuring Transactional Data for Link Analysis in the FinCEN AI System," in *Papers from the 1998 Fail Symposium on Artificial Intelligence and Link Analysis, October 23-25 Orlando, FL,* Technical Report WS-98-0 Menlo Park, CA: AAAI Press, pp. 38-46

Hamilton, M. (2010), "Thesis: Large Margin Kernel Methods for Calmodulin Binding Prediction," Department of Computer Science Colorado State University, Spring 2010.

Han, J., Kamber, M. (2006), "Data mining: concepts and techniques," *Second edition Morgan Kaufman publishers 2006,* pp. 285 - 464

Hand, D.J. (2010), "Fraud Detection in Telecommunications and Banking: Discussion of Beacker, Volinsky, and Wilks (2010) and Sudjianto *et al.* (2010)" – *Technometrics, February 2010, vol. 52 (1),* pp. 34 – 38

Hand, D.J. (2006), "Classifier Technology and the illusion of progress," *Statistical Science Vol. 21(1),* pp. 1 - 14

Hand, D., Mannila, H., and Smyth P. (2001), "Principles of Data Mining," *MIT Press, Cambridge, MA.*

Hand, D. J., and Blunt, G. (2000), "Prospecting for gems in Credit Card Data," *Proceedings if the Workshop on Statistical Modeling for Data Mining,* University of Pavia

Hand, D. J., Blunt, G., Kelly, M. G., and Adams, N. M. (2000). "Data mining for fun and profit," *Statistical Science 15,* pp. 111-131.

HHS OIG, Department of Health and Human Services, Office of Inspector General, 2002. "Improper Fiscal Year 2001 Medicare Fee-for-Service Payments," February 21, 2002

Hill, T.P. (1995), ´A Statistical derivation of the significant-digit law," *Statistical Science 10,* pp. 354 - 363

Hyman, D. A. (2001), "Health care fraud and abuse: Market Change, Social Norms, and the Trust "reposed in the workmen"," *Journal of Legal Studies,* pp. 531 – 567

Hyman, D. A. (2002), "HIPAA and Health Care Fraud: An Empirical Perspective," *Cato Journal Vol. 22 (1),* pp. 151 - 178

Jensen, M. (1983), "Organization theory and methodology," *Accounting Review 56,* pp. 319 – 338

Joachims, T., (1998), "Making Large-Scale SVM Learning Practical," *LS8-Report, University of Dortmund*

Kantardzic, M. (2002), "Data mining: concepts, models, methods, and algorithms," Wiley IEEE Press.

Kelly, R. (2009), "Where can $700 billion in waste be cut annually from the U.S. Healthcare system?," *White paper Thomson Reuters,* pp. 1 - 30

Kou, Y., Lu, C-T., Sirwongwattana, S., and Huang, Y-P., (2004), "Survey of Fraud Detection Techniques," *Proceedings of the 2004 IEEE International Conference on Networking, Sensing & Control, Taiwan 21-23 March 2004,* pp. 749 – 754

Kumar, S. and Spafford, E.H. (1994), "A pattern matching model for misuse intrusion detection," *Inforsec Computer Science*

Laleh, N., and Azomi, M.A., (2009), "A Taxonomy of Frauds and Fraud Detection Techniques," *ICISTM 2009, CCIS 31,* pp. 256-267

Lee, W. and Stolfo, S. (1998), "Data mining approaches for intrusion detection,' *Proceeding of the 7th USENIX Security Symposium, San Antonio, Texas* pp. 749 – 754

Leonard, K. J. (1993), "Detecting credit card fraud using expert systems". *Computers and Industrial Engineering 25,* pp. 103-106.

Macaulay, S. (1963), "Non-contractual relations in a business; a preliminary study," *American Sociology Review 28 (1),* pp. 55-67

Maes, S., Tuyls, K., and Vanschoenwinkel, B. (2002), "Credit Card Fraud Detection Using Bayesian and Neural Networks," *in Proceedings of the 1st International NAISO Congress on Neuro Fuzzy Technologies*, Havana, Cuba

Major, J.A., and Riedinger, D.R. (1992), "EFD: A Hybrid Knowledge/Statistical-Based System for the Detection of Fraud," *International Journal of Intelligent System, vol. 7,* pp. 687 – 703

Medicaid Florida (2010) – information retrieved on 24th of November http://www.hrsa.gov/reimbursement/states/Florida-Medicaid-Covered-Services.htm

Medi-Cal (2010), Retrieved 12.10.2010, from " www.medi-cal.ca.gov"

Moreau, Y., Preneel, B., Burge P., Shawe-Taylor, J., Stoermann, C., and Cooke, C. (1997), " Novel Techniques for fraud detection in mobilte telecommunications networks," *ACTS Mobile Summit*, Grenada Spain

Morris, M. (1999), "New Allegations Raised in Dentist's Fraud Case," *Kansas City Star,* June 12 1999

Mukkamala, S., Janoski, G., and Sung, A. (2002), "Intrusion Detection Using Neural Networks and Support Vector Machines," *IEEE 2002*

Mullins, I. M. et al (2005), "Data Mining and Clinical Data Repositories: Insights from a 667,000 patient data set," *Computers in Biology and Medicine 36,* pp. 1351 – 1377

Murad, U., and Pinkas, G. (1999), "Unsupervised Profiling for Idnetifying Superimposed Fraud," *Proc. Of PKDD99*

Ngai, E.W.T., Hu, Y., Wong, Y.H., Chen, Y., Sun, X. (2010), "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature", *Decision Support Systems 2010,* pp. 1 - 11

Nigrini, M.J. (1999), "I've Got Your Number," *Journal of Accountancy,* pp. 79 - 83

Nilson Report (2006), HSN Consultants, Nilson report issue #858, June 2006

Office of Inspector General, Department of Health & Human Services (August 26, 2009) "MSIS Data usefulness for detecting fraud, waste and abuse"

Phua, C., Lee, V., Smith K., and Gayler R. (2005), "A comprehensive Survey of Data Mining-based Fraud Detection Research," *School of Business Systems,* final version 2: 9/02/2005 pp. 1-10

Pressman, R.S. (2000), "Software Engineering: A Practitioner's Approach," *McGraw-Hill College Division 806,* Boston, MA.

Psaty, B.M., Boineau, R., Kuller, L.H., and Luepker, R.V. (1999), "The potential costs of upcoding for heart failure in the United States," *The American Journal of Cardiology 84,* pp. 108-109

Relativity Corporation (2010), - http://www.relativitycorp.com/data/article1.html, retrieved at 11th October 2010

Sappington, D.E.M. (1991), "Incentives in Pricipal-Agent Relationships," *Journal of Economic Perspectives 5(2),* pp. 45 – 66

Silverman, E., and Skinner, J. (2001), "Are for-profit hospitals really different? Medicare upcoding and market structure, *NBER working paper 8133 (February)*

Sohl, J.E. & Venkatachalam, A.R. (1995), "A Neural network approach to forecasting model selection," *Information & Management, 29 (6)* pp. 297 - 303

Song, X., Wu, M., Jermaine, C., and Ranka, S. (2007), "Conditional Anomaly Detection," *IEEE Transactions on Knowledge and Data Engineering 19 (5),* pp. 631 – 645

Statsoft – Software for Statistics, Data Mining, Predictive Analytics and Credit Scoring (2010), http://www.statsoft.com/textbook/data-mining-techniques/, information retrieved on 23rd of December 2010

Stonebraker, M., Cetintemel, U., and Zdonik, S. (2005), "The Requirements of Real-Time Stream Processing" – *SIGMOD Record, Vol. 34(4)* – pp. 42 – 47

Sudjianto, A., Nair, S., Yuan, M., Zhang, A., Kern, D., and Cela-Díaz, F. (2010), "Statistical Methods for Fighting Financial Crimes," *Technometrics 52,* pp. 5–19

Schwartz, R.B. and Russo, M.C. (2004), "How to quickly find articles in the TOP IS Journals," *Communication of the ACM vol. 47(2)*

Tan, P., Steinbach, M., Kumar, V. (2006), "Introduction to Data Mining," *Pearson International Edition,* Chapter 1 - 4, 6, 8 & 10

Tasoulis, D., Adams, N.M., Weston, D.J., and Hand, D.J. (2008), "Mining Information from Plastic Card Transaction Streams," *Proceedings in Computational Statistics: 18th Symposium Vol. 2,* pp. 315-322

Todd, P., Benbasat I. (1999), "Evaluating the Impact of DSS, Cognitive Effort, and Incentives on Strategy Selection," *Information Systems Research,* Vol.10, No.4, pp. 356-374

Turban, E., Aronson, J.E., Liang, T.P., Sharda, T. (2007), "Decision Support and Business Intelligence Systems," *9th Edition Pearson Education, 2010*

Vapnik, N.V. (2000), "The Nature of Statistical Learning Theory,," Second edition

Vessey, I., Galletta, D. F. (1991), "Cognitive fit: An empirical study of information acquisition," *Information System Research 2(1),* pp. 63-84

Vladimir, V.N., (1995), "The Nature of Statistical Learning Theory," *Springer, Berlin Heidelberg New York*

Vinciotti, V., and Hand, D.J. (2003), "Scorecard construction with unbalanced class sizes," *J Iran Statis Soc 2,* pp 189 - 205

Weston, D.J., Hand, D.J., Adams, N.M., Whitrow, C., and Juszczak, P. (2008), "Plastic card fraud detection using peer group analysis," A*dvances in Data Analysis and Classification Vol. 2(1),* pp 45 - 62

Whitrow, C., Hand, D.J., Juszczak, P., Weston, D., Adams, N.M. (2009), "Transaction aggregation as a strategy for credit card fraud detection", *Data Mining and Knowledge Discovery 18 (1),* pp. 30 - 55

Witten, I. H. and Frank, E. (2005), "Data Mining: Practical machine learning Tools and Techniques," 2nd edition, Elsevier

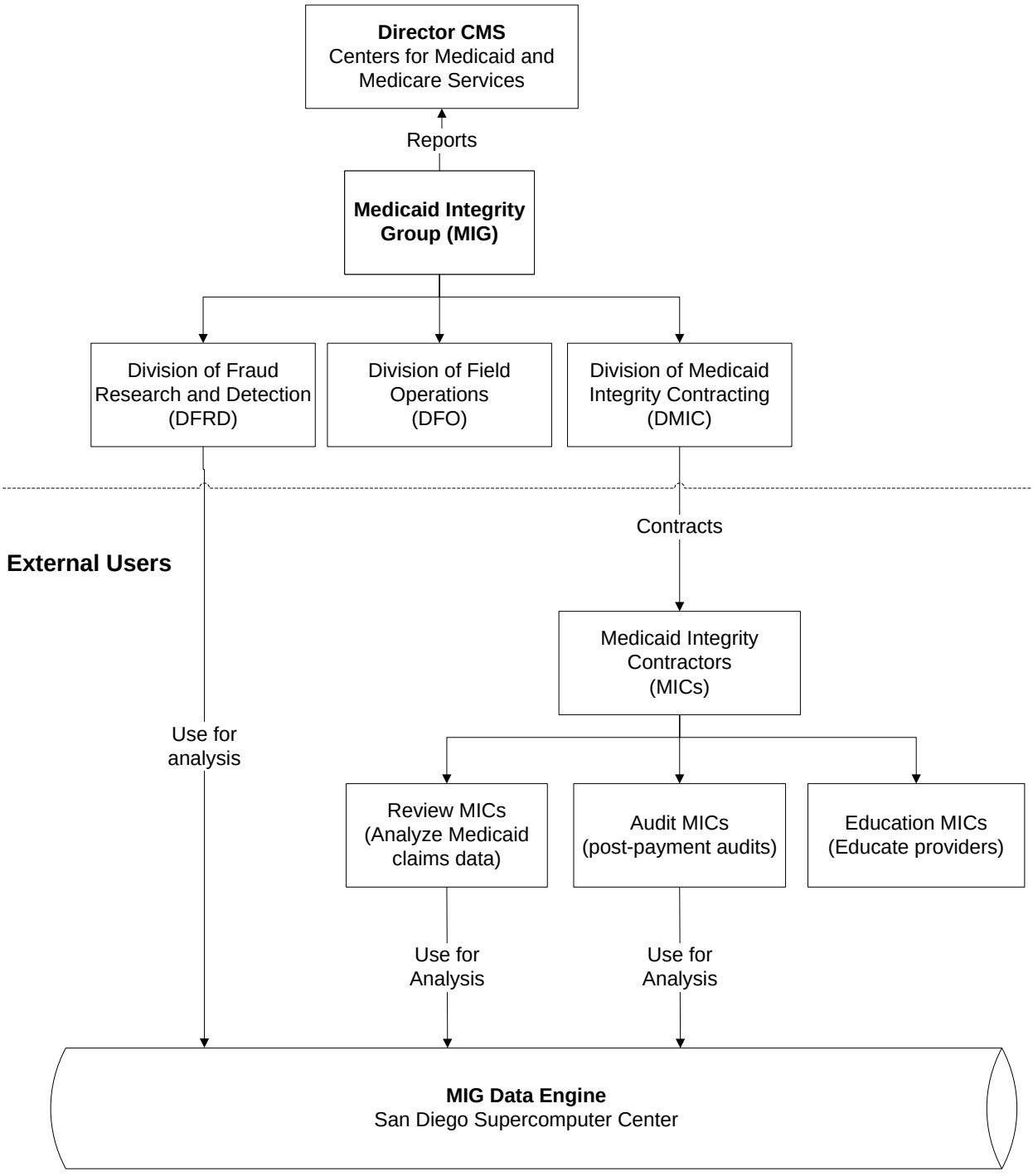
Yang, W., Hwang, S. (2006), "A process-mining framework for the detection of healthcare fraud and abuse," *Expert systems with Applications,* 31, pp. 56-68

Zhang, Z., Salerno, J. J., and Yu, P.S. (2003), "Applying Data Mining in Investigating Money Laundering Crimes," in *Proceedings of the Ninth ACM SIGKDD,* New York: ACM, pp. 747 – 752

Zhang, D., and Zhou, L. (2004), "Discovering Golden Nuggets: data mining in financial application," *IEE Transaction on Systems, Man and Cybernetics 34(4)*

Zhou, W., and Kapoor, G. (2010), "Detecting evolutionary financial statement fraud," *Decision Support Systems 2010*

Appendix A: Stakeholders overview of the Medicaid Integrity Group



Stakeholder overview MIG Data Engine

Appendix B: Fraud detection process with corresponding stakeholders

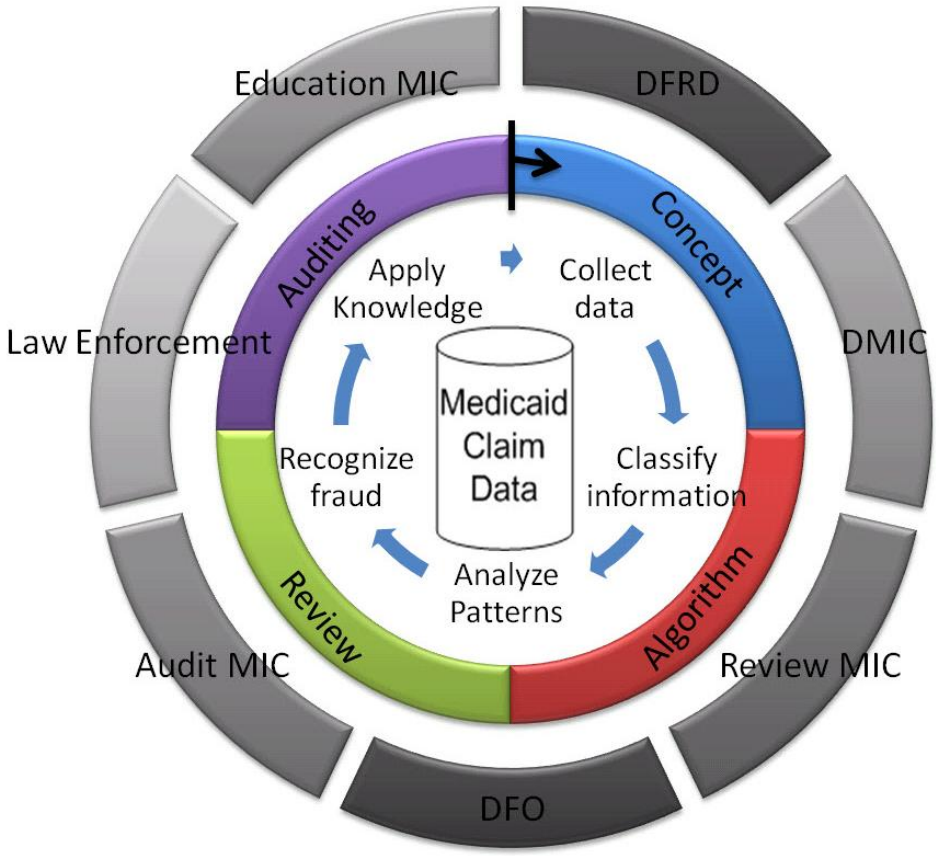


Figure 29: Fraud detection process and stakeholders involved