Fraud detection within Medicaid

Msc Thesis

Paulus Schoutsen

UNIVERSITY OF TWENTE.

UNIVERSITY OF CALIFORNIA · SAN DIEGO · LET THERE BE LIGHT

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Name	Paulus August Maris Schoutsen, Bsc
E-mail	p.a.m.schoutsen@student.utwente.nl
Student ID	s0088838
University	University of Twente
Faculty	School of Management and Governance
Study	Business Information Technology
Examination Committee	Prof. Dr. Roland Müller
	Berlin School of Economics and Law
	Prof. Dr. Jos van Hillegersberg
	University of Twente, Netherlands
Company Supervisor	Dallas Thornton, MBA
	University of California, San Diego

Preface

You are not reading just an ordinary thesis. No – this is my thesis. And it is *awesome*. And to make it even better, after eleven months of hard work it is finally done. I wouldn't have been able to do any of this if it wasn't for the support of a lot of people and the perfect location.

First of all I would like to thank my supervisors Roland, Jos and Dallas for the countless Skype sessions in which they managed to give shape and purpose to my thesis. I would like to thank especially Roland for his punctuality and feedback, Jos for his patience and Dallas for the opportunity to be in San Diego.

Off all the places that one can write a thesis I think that there is not a better place than San Diego. The weather that is always sunny, relaxed atmosphere and the ocean make sure there is always a positive vibe. And if all the American way of living got too much I could always lose my steam next doors in Tijuana, Mexico.

Next to that I would like to thank my girlfriend Anne Therese and the many friends of the "Awesome San Diego" group who did an awesome job in keeping me occupied outside of my office hours and reminding me from time to time that I came to San Diego to write a thesis.

This list of thanking people wouldn't be complete without thanking my family who has always encouraged me to enjoy life and do the things that I like.

And finally I would like to thank my colleagues at the San Diego Supercomputer Center that made it every day fun to come to work: Peter, Carlo, David and Lanre. Also a big thanks to all my other colleagues that have been able to enjoy my music in the office for the last eleven months: *"If you think that my music is too loud – you're too old".*

Enjoy,

Paulus

Paulus Schoutsen, Bsc

1 Introduction

In 2007 a total amount of $2.26 trillion was spent on health care within the United States of America (National Health Care Anti-Fraud Association, year unknown).

An exact measure of fraud is unavailable. This is due to the fact that most fraud goes undetected and if it is detected it is stopped. This means there is only undetected fraud of which the size cannot be known. The National Health Care Anti-Fraud Association (NHCAA) estimates that the losses due to health care fraud are in the tens of billions of dollars each year. The Federal Bureau of Investigation (2009) supports this claim by stating that three to ten percent of total billings within public and private health care are fraudulent.

Besides the staggering costs of health care fraud also imposes a risk for the health of patients. One of the most significant trends observed by the Federal Bureau of Investigation (2009) is the willingness of medical professionals to risk patient harm in their fraud schemes.

One of the problems that there is still so much fraud within the system is the lack of sophisticated fraud control systems. The systems are static, lack real time fraud detection and focus on fraud detection in the transaction without looking for patterns of suspicious behavior within the interactions between patients and health care providers. In these patterns might be hidden, larger, more complex and more sophisticated fraud schemes.

This thesis will focus on the health insurance that the government of the United States of America provide to people with low incomes: Medicaid.

1.1 Research focus

This research will focus on developing a method that enables quick development and deployment of effective fraud detection algorithms within Medicaid. This will consist of a method to structure the data and a method to use the structured data to find fraud.

1.2 Design science methodology

To bring structure into the development approach of the fraud detection system the Information System Research Framework by Hevner, et al. (2004) as shown in Figure 1 will be used.

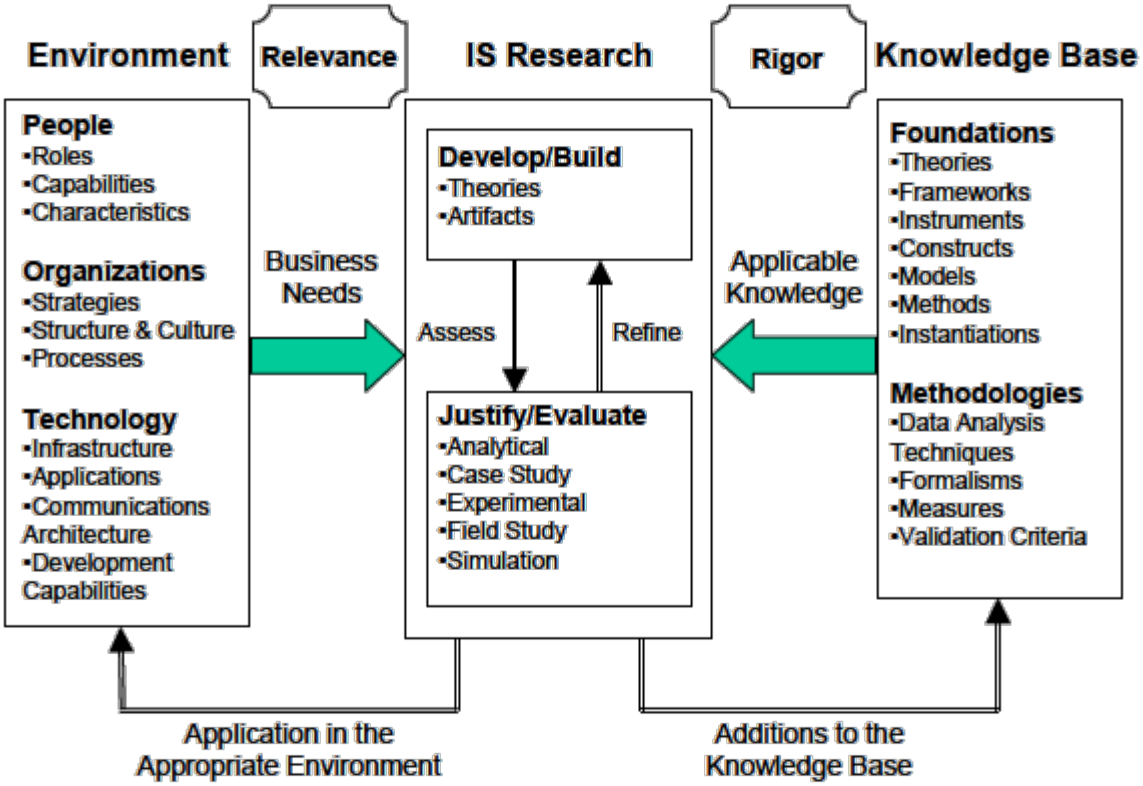


Figure 1 Information System Research Framework (Hevner, Ram, March, & Park, 2004)

The environment is represented by the fraud detection industry. The industry has big databases of health insurance claims and related information and is always in need of more advanced and flexible fraud detection algorithms and systems. The environment and its lack of advanced detection and algorithms will be further explained in chapter 2.

The knowledge base is represented by the scientific literature on fraud detection, data mining and the health care industry. This will be touched upon in chapter 2.

The artifact that is to be created for this IS research is a data warehouse optimized for fraud detection and a workflow how it can be used. This artifact and the steps taken to create it will be further discussed in chapter 3 and chapter 4.

The artifact will be evaluated by creating a prototype and do black box functional testing to detect its usefulness, failures and identify defects in chapter 5. In chapter 6 the platform will be evaluated to see if it can be used to detect the most prevalent fraud cases. Interviews will be conducted with an expert on health care fraud detection to review the work.

These methods have been chosen as they most applicable. Alternatives for example would be doing a case study. This is not a good option as it will require an actual implementation to be able to validate it which is very difficult to achieve. Also a simulation is not an option as it will let the system work with fake data which will not represent the fraud cases that are out there. Let alone is it possible to correctly simulate and test it.

If the evaluation of the developed fraud detection system is successful in finding suspicious behavior it can be used as a basis for a new, or improvements to a current, fraud detection system on the business side. It can be used as either a prepayment or a postpayment solution. Also the result of this thesis will add to the knowledge about health care fraud detection methods.

1.3 Research questions

To structure the research the goal translated into a research question as follows:

What is an efficient way of finding fraud within Medicaid?

This question will be answered throughout this thesis by answering the following sub questions:

What is fraud and what kind of pattern detection approaches are available?

Chapter 2 will contain a literature study about fraud within health care that will show how fraud remains a big challenge to solve. It will look into current attempts that are made to battle fraud and into the different approaches to prevent and find medical health care fraud.

What should be the layout of a data warehouse to support fraud-finding algorithms that cover most fraud?

Based on the literature study and gathered domain knowledge a data warehouse will be developed in chapter 3 to support the development and deployment of new algorithms to detect fraud within the health care industry. The will be on a system that will be flexible enough to be able to be used to find all possible ways of fraudulent behavior.

How can the data warehouse be used in an effective way?

Chapter 4 will detail a workflow to use the data warehouse to find fraud. The focus will be on how the usage of the system can help find fraud and refine and improve the data warehouse while doing so.

Chapter 5 will provide the results on a prototype of the data warehouse. Further evaluation of the platform will be done in chapter 6 where the platform will be tested to see if it can find the most prevalent fraud schemes according to the Federal Bureau of Investigation and let the data warehouse be analyzed by an industry-expert.

2 Fraud within the Health Care Industry

To talk about fraud and abuse the following definition of fraud and abuse in health care by Kelley (2009) will be used in this paper:

> *Fraud and abuse: A situation in which healthcare is paid for, but not provided, or a situation in which reimbursement claims are made to third party insurance companies or federal programs such as Medicare or Medicaid, and no such services were rendered. Fraud and abuse are also defined as healthcare providers receiving kickbacks, patients seeking treatments that are potentially harmful to them (such as seeking drugs to satisfy addictions), and the prescription of services known to be unnecessary.*

Within the health care system there are three parties that can commit fraud according to Li, et al. (2008): service providers, insurance subscribers and insurance carriers. Service providers are accounted for the greatest proportion of the total health care fraud and abuse and also the highest risk: health problems may be caused by patients receiving unnecessary treatments. As service providers account for the highest proportion of fraud and can do actual damage the rest of this thesis will focus on detecting fraud by this group.

The world of fraud and fraud detection is dynamic: when one method of fraudulent behavior has been found a detection system will be developed. The fraudulent health providers will move on to find the next weakness in the system to exploit. Fighting fraud is a never ending battle which requires a robust system on which new fraud detection patterns can quickly be deployed.

To make the battle on fraud more difficult one should note that some fraudulent claims are the result of mistakes made during entering the claim into the system by honest providers. For example, an honest provider could accidently claim provided services on the wrong patient or he could make a typing error when entering the service-code.

Fraud and fraud detection not only causes a lot of costs for the health insurance. It also results in high costs for the health providers who need to have a detailed administration of all services provided to show the health insurance that they are acting honestly. According to a survey performed in California by Kahn, et al. (2005) physician offices spend 27 percent of their revenue on administration and 14 percent on billing and insurance-related (BIR) functions. Hospitals spend 21 percent and 7-11 percent respectively. All these costs increase the cost of health care for patients.

2.1 Types of fraud

According to Sparrow (2000, p. p205) there are 2 different types of fraud: “hit-and-run” and “steal a little, all the time”. With a hit-and-run the fraud perpetrator doesn’t care so much about stealth. It will submit a lot of fraudulent claims, acquire large amounts of money and disappear off the grid before anyone can catch him or the acquired money. With “steal a little, all the time” the fraud perpetrator makes sure that the fraudulent claims will go unnoticed so that he can submit false claims over a long period of time. For example he will hide his false claims within large batches of valid claims and when caught will say it was an error, repay the money and continue his behavior.

Major and Riedinger (2002) have identified five categories in which health care fraud can take place:

Financial	The flow of dollars
Medical Logic	Whether a medical situation would normally happen.
Abuse	The frequency of treatments.
Logistics	The place, time and sequence of activities.
Identification	How providers present themselves to the insurer.

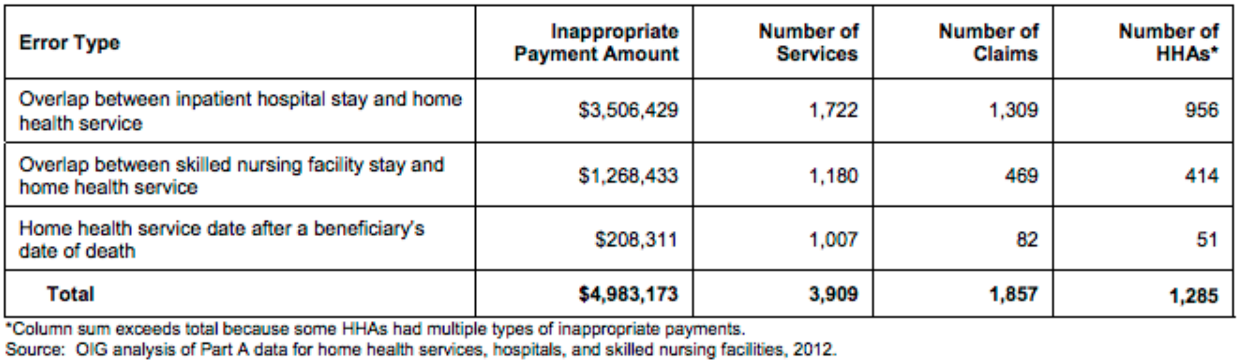
These categories are taken into account for the development of the data warehouse in the next chapter.

The Federal Bureau of Investigation is actively involved with the public and private health care industry in finding fraudulent health care providers. With assistance of the NHCAA and CMS the FBI has identified Durable Medical Equipment (DME) providers as one of the top two provider types identified within all case referrals, preliminary investigations, and suspensions (Federal Bureau of Investigation, 2009). These providers and other providers are using a great variety of fraud schemes to steal money from health insurance. The FBI has created a list of the most prevalent fraud schemes within the health care industry (Federal Bureau of Investigation, 2009) that is discussed in the following sections.

2.1.1 Billing for services not rendered (Logistics)

The provider will add extra services to a bill or even bill patients that it has never seen. This fraud is also called phantom billing. For a health provider most of this type of fraud is very difficult to track down without interference from the patient. Exceptions are for when the service provided overlap with other services or the patient died.

The United States Department of Health and Human Services published a report in august 2012 (Office of Inspector General, 2012) on analyzes performed on $19,5 billion worth of Medicare payments to home health agencies (HHA) in 2010. They found that Medicare inappropriately paid $5 million for home health claims that overlapped with claims for stays in inpatient hospitals or skilled nursing facilities.



Error Type	Inappropriate Payment Amount	Number of Services	Number of Claims	Number of HHAs*
Overlap between inpatient hospital stay and home health service	$3,506,429	1,722	1,309	956
Overlap between skilled nursing facility stay and home health service	$1,268,433	1,180	469	414
Home health service date after a beneficiary's date of death	$208,311	1,007	82	51
Total	**$4,983,173**	**3,909**	**1,857**	**1,285**

*Column sum exceeds total because some HHAs had multiple types of inappropriate payments.
Source: OIG analysis of Part A data for home health services, hospitals, and skilled nursing facilities, 2012.

Table 1 Inappropriate Medicare Payments for Home Health Services, 2010 (Office of Inspector General, 2012)

This fraud type is also exploited a lot by criminals. They use hit-and-run approaches like the Drop Box Scheme shown in Figure 2.

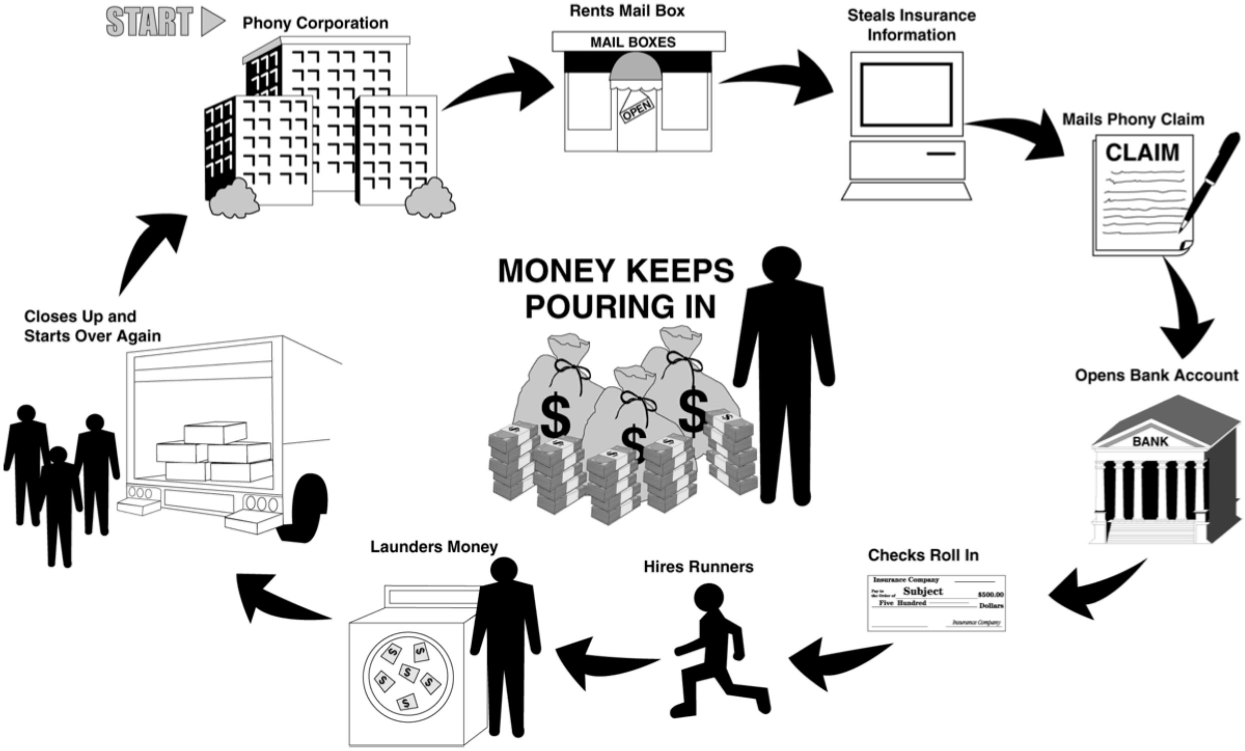


Figure 2 Drop Box Scheme (United States General Accounting Office, 2000)

This scheme exploits the fact that the current state of fraud detection is not adequate enough to catch abuse as it happens. By quickly billing a lot of patients, cashing the money and disappearing the fraud perpetrators manage to steal a lot of money.

2.1.2 Upcoding of services and items (Medical Logic)

When upcoding happens the patients are billed for an item or a service that is similar or covers the service or item delivered to the patient but is more expensive.

Psaty et al. (1999) did research to the upcoding of treatments related to heart failures. They collected a range of information on Medicare-paid hearth failure events in 1993 including hospital records, interviews patients, physicians, witnesses and death certificates. They conclude that 37,5% of hospitalizations for hearth failure may reflect incorrect diagnoses resulting in US hospitals receiving excess reimbursements from Medicare of as much as $933 million a year.

2.1.3 Duplicate claims (Logistics)

This happens when a provider bills a service or item twice on the same patient. When a provider is caught executing this fraud it is very easy for him to state that the fraud was a result of an administrative error and continue business as usual.

2.1.4 Unbundling (Logistics)

The practice of submitting bills in a fragmented fashion in order to maximize the reimbursement for various tests or procedures that are required to be billed together at a reduced cost.

An example would be blood tests. A physician orders a blood test to be done to test for X, Y and Z. This can be billed together for $60 as a package blood test or $25 per separate test.

2.1.5 Excessive services (Abuse)

This occurs when a service provider provides unnecessary services to a patient.

Copeland et al. (2011) did research on 321 DME suppliers in Nevada providing incontinence briefs, diapers, or pads. They found that suppliers whose average claim was over 272.5 diapers per patient per month fell into the 95 percentile. 270 diapers per patient per month might be requested in the beginning to build a stock. But getting this as an average would mean the patients use an average of 9 diapers per day per month!

2.1.6 Medically unnecessary services (Abuse)

This fraud happens when a specialist provides services to the patient that does not need the treatment. When a patient is exposed to medically unnecessary services there is a chance that the health of the patient will actually decrease. An example by Sparrow (2000, p. 5): In September 1998, Thomas Anderson, a dentist, pleaded guilty in Michigan 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.

In this kind of fraud it might be that the patient is involved in the fraud scheme. Cases are known that recipients receive kickback for receiving bills for services never rendered. An example of this is the Rent-a-Patient scheme. This scheme, shown in Figure 3, consists of a middleman that recruits willing patients to be examined by clinics. The clinic will perform unneeded tests and services and bill the insurance company.

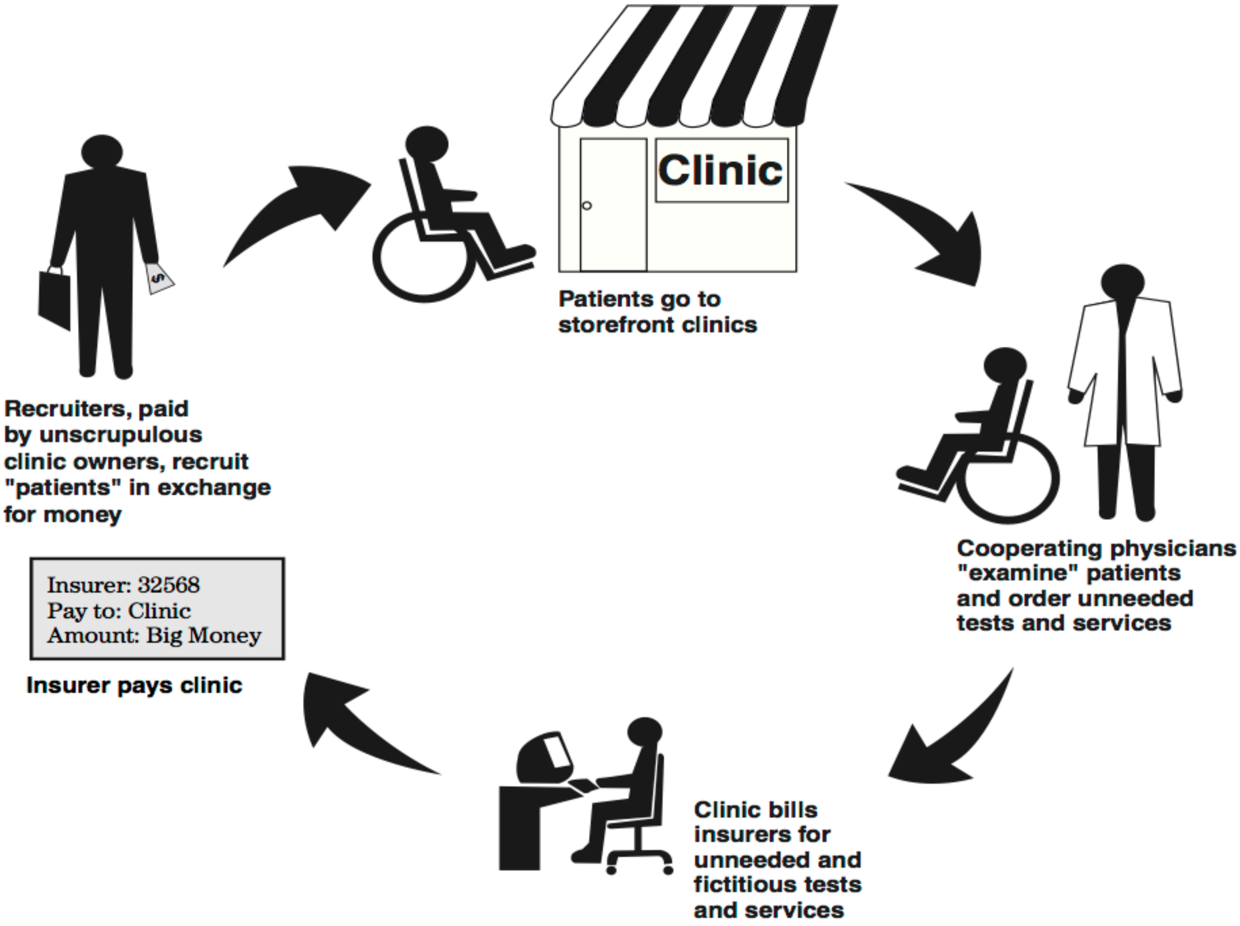

Clinic
Patients go to storefront clinics
Cooperating physicians "examine" patients and order unneeded tests and services
Clinic bills insurers for unneeded and fictitious tests and services
Insurer: 32568
Pay to: Clinic
Amount: Big Money
Insurer pays clinic
Recruiters, paid by unscrupulous clinic owners, recruit "patients" in exchange for money
$

Figure 3 Rent-a-Patient scheme (United States General Accounting Office, 2000)

2.1.7 Kickbacks (Financial)

When a health care provider accepts or offers money for referral of a patient for health care services that may be paid for by health care insurances it is called a kickback. Kickback schemes might involve false diagnoses so the patient can be treated and billed upon by other health care providers.

An example of such a fraud scheme involving kickbacks and false diagnoses is the Pill Mill Scheme as shown in Figure 4.

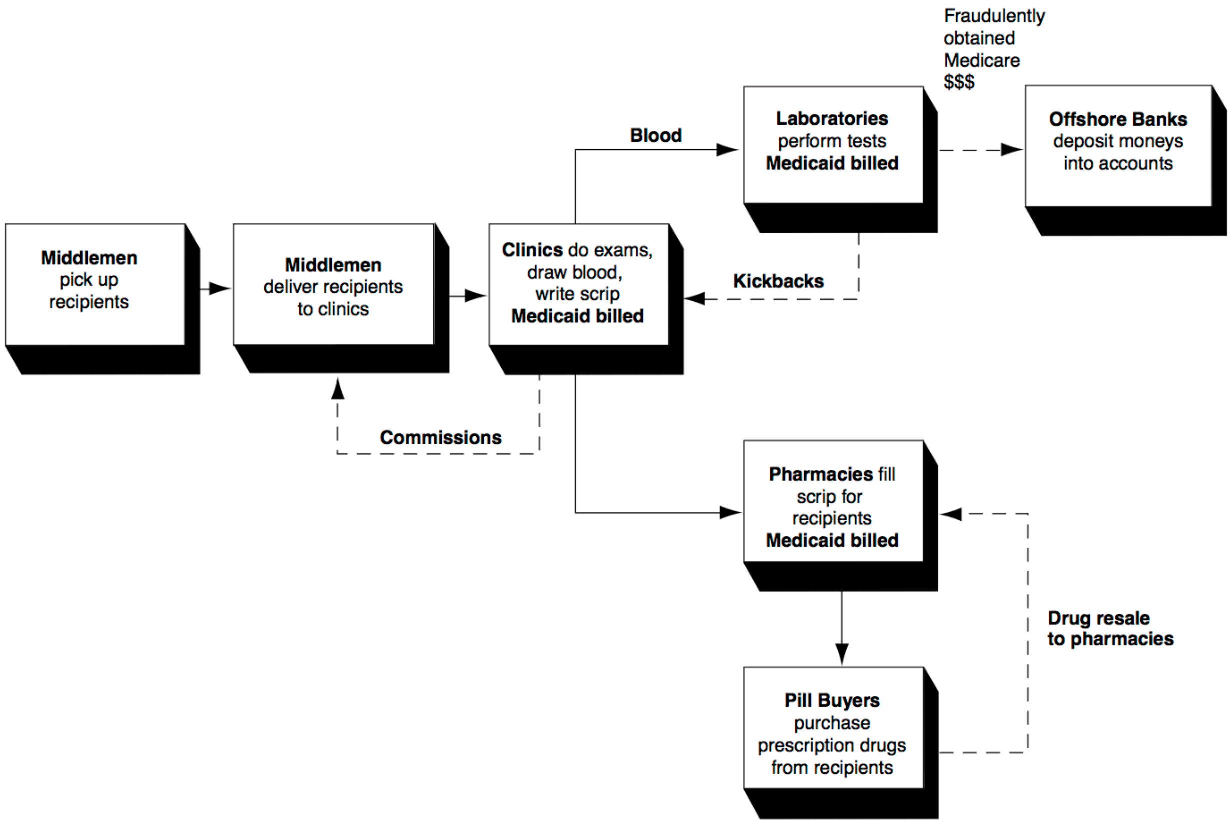

Fraudulently obtained Medicare $$$
Blood
Laboratories perform tests Medicaid billed
Offshore Banks deposit moneys into accounts
Middlemen pick up recipients
Middlemen deliver recipients to clinics
Clinics do exams, draw blood, write scrip Medicaid billed
Kickbacks
Commissions
Pharmacies fill scrip for recipients Medicaid billed
Drug resale to pharmacies
Pill Buyers purchase prescription drugs from recipients

Figure 4 Pill Mill Scheme (United States General Accounting Office, 2000)

The Pill Mill Scheme is a variation on the rent-a-patient scheme. Multiple actors are involved in guiding a patient through a set of fake diagnoses and unnecessary tests and sales. The actors are usually related through a joint or related ownership. By involving the patient in the scheme there is no one that will complain about the fraud.

2.2 Finding fraud

Looking for fraud is based heavily on data-analytics. Is it possible to find suspicious patterns within the available data? Are there claims that are just impossible to have happened? Did the provider bill on a patient that died two years ago? Did a bill come in two hundred miles away from where the patient lives while there was a provider claiming on the patient the same day but close by the patients' house? In the latter case it is unable to determine with a hundred percent certainty that it is fraud but it is suspicious and further research should be done.

The fraud detection can take place at two stages in the processing of claims: prepayment and postpayment.

Prepayment is the moment when a claim gets in at the insurance company and it is not yet paid. Finding fraud at this moment is preferred as you will not have to recover the money; it is not paid yet. It does give extra challenges because less data is known to do fraud detection on and flagging false positives will cause grief by honest providers as they will have to provide proof that they are doing honest business and have to wait for their money in the process.

Postpayment is where most fraud detection is done. It is the analyses of the data after the claim has been paid. This can happen one month after the claim has been paid but also 5 years later. The later a claim gets flagged as potential fraud the more issues might arise that will stop the recovery of the money. The provider might have vanished or no proof can be collected anymore that a patient received a treatment yes or no. Postpayment analyses might result in indicators that can be used in prepayment to declare a claim suspicious more easily based on suspicious provider behavior in the past.

Sparrow (2000, pp. 237-238) writes that the process of generating indicators postpayment to use during prepayment monitoring should not be confused with prepayment monitoring as it will not provide protection against rapid bust-out schemes because the indicators will only be created after postpayment analyses has been done – which can take months.

Categorizing the most prevalent fraud schemes shows that the categories Logistics, Medical Logic and Abuse are having the most visible fraud problems within the health care industry. One should note when reading this list that this list shows which fraud schemes are being detected the most and therefore are visible. This does not mean that they occur the most. Due to the nature of fraud trying to be undetectable one can never be certain which fraud scheme occurs most.

To detect each scheme, whether executed in a “hit-and-run”- or a “steal a little, all the time”-fashion, the payment integrity programs need to look beyond the transaction-level controls that are embedded in the claims-processing sequence as they accomplish very little for fraud control (Sparrow, 2000, p. p131). One has to look at multiple events related to patients or health care providers to be able to classify the behavior and see if something suspicious is going on.

There are two ways to find fraud in large data sets of claims through data-mining: supervised and unsupervised methods. Unsupervised statistical fraud detection methods are statistical methods that look for anomalies in data sets without being setup to look at specific attributes. It does so by determining outliers in a data set. Supervised methods look for specific, pre-defined, patterns or values in the data.

The weakness of supervised methods is that it requires prior knowledge on fraudulent claims and indicators to be able to find similar claims. According to Travaille, et al. (2011) supervised classification models are particularly appropriate as they can be trained and adjusted to detect sophisticated and evolving fraud schemes. Still, unsupervised methods are required to detect new, unknown, types of fraud before it has been discovered and a detection scheme has been implemented as a supervised method.

The reason why payment integrity programs accomplish very little on fraud detection is partly due to the way the current claim systems work. They are setup with honest users in mind providing feedback in cases of things going wrong. According to Sparrow (2000, pp. p163-165) when the system detects an error in a submitted claim it will either try to correct the claim if able to or it will reject the claim and tell the user what caused the error. Thus telling potential fraud perpetrators how to adjust the claim so they can bypass some of the transaction level controls and get it not rejected.

Sparrow (2000, p. p233) proposes that for effective fraud detection one has to look at the data beyond the transaction-level. To do this he has setup seven levels of health care fraud control that cover all fraud out there:

Level 1		Claim, or Transaction level
Level 2		Patient/Provider relationship
Level 3	a	Patient level
	b	Provider level
Level 4	a	Patient group/Provider
	b	Patient/Practice (clinic)
Level 5		Policy/Practice relationship
Level 6	a	Defined group of patients (e.g., families or residents of one nursing home.)
	b	Practice (or Clinic)
Level 7		Multiparty, criminal conspiracies

Each higher level involves larger fraud schemes with more people and money involved and an increased difficulty of being detected.

According to Sparrow (License to Steal: How Fraud bleeds america's health care system, 2000, p. 239) the bulk of the industry's detection toolkit is focused at level 1 and 3. During prepayment the transaction (level 1) and patient level (level 3a) are evaluated to see if suspicious behavior is going on. For example a second childbirth within 9 months or a duplicate claim being billed. During post payment the focus is on the provider level (level 3b). For example if there was an unusual increase in the amount of claims billed.

2.3 Current fraud finding efforts in the United States

The United States is putting a lot of effort into the finding of fraud. The reason for this is because they are paying for two health insurances: Medicaid and Medicare. First one is a free health care insurance for people with low income or resources. Medicare is a health insurance to guarantee access to health care for the elderly.

The cost of these insurances has been growing steadily over the last decades on a higher rate than the gross domestic product (GDP) as can be seen in Figure 5. In 2009, health care costs reached $2.5 trillion – nearly 17 percent of the GDP. Yet despite this spending, health outcomes in the U.S. are considerably below those in other countries (Institute of Medicine, 2010).

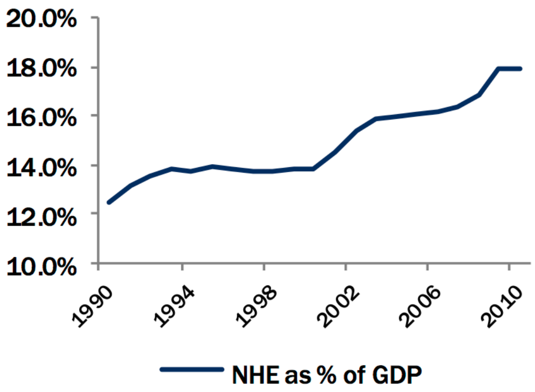


Figure 5 U.S. Health Expenditures as a Percentage of GDP (Health Care Cost Institute, 2012)

Besides an increasing impact of GDP America also suffers from very high health care costs compared to other countries as can be seen in Figure 6. It is therefore in great interest of the American government to reduce these costs.

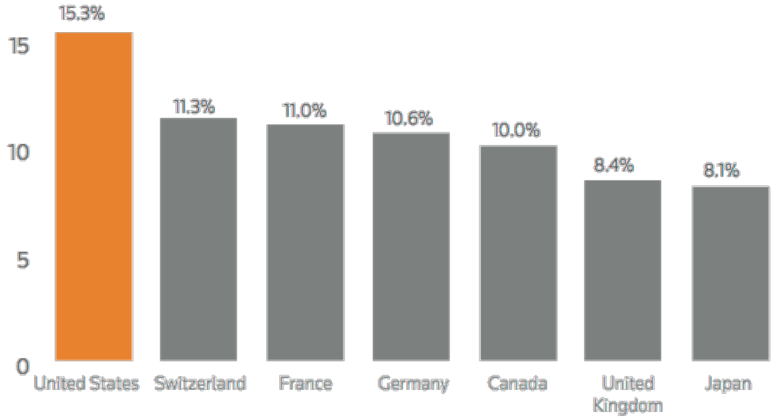


Figure 6 Health Care Spending Share of GDP in 2006 (Kelley, 2009)

The bill for the insurance is split between the federal government and the state governments through the Federal Medical Assistance Percentages (FMAP). The percentage of the bill that is paid by the federal government varies between 50 percent (i.e. California) and 74 percent (i.e. Mississippi). This might give state governments less incentive to look for fraud, as they will only receive between 26 and 50 percent of the money that they recover. Another reason that the states will try to not look for fraud is that the program managers will look bad if money is recovered from fraud perpetrators: they were tricked by the fraud and paid the claim in the first place.

One of the ongoing efforts to reduce health care costs is the battle on fraud. The government has introduced numerous legislations including increasing the number of audits done on payments and the introduction of a *qui tam* provision through the False Claim Act (31 USC § 3729–3733) in 1986. This latter one allows individuals that start a case against a fraud perpetrator to receive a part of the reclaimed money. This will increase the willingness of individuals to bring fraud cases to light.

The False Claim Act has proven to be very effective. It has led to judgments and settlements against fraud perpetrators to a total over $25 billion (Taxpayers Against Fraud Education Fund):

In 2009 Pfizer agreed to pay $2.3 billion for fraudulent marketing: $1.3 billion criminal fine, $1 billion under the False Claim Act. (United States Department of Health & Human Services, 2009)

In 2006 Tenet Healthcare Corporation, operator of the nation's second largest hospital chain, has agreed to pay the United States more than $900 million for alleged unlawful billing practices. (Department of Justice, 2006)

One of the biggest challenges in the fraud hunt is not the finding of the fraud itself but the process of proving fraud is happening and recovering the money. After a claim has been flagged as suspicious humans will have to look at the data, judge if it is worth prosecuting, request extra information from the providers and start a law suit to get the money back. This can take a long time and can result in a lot of cases piling up. The more cases that pile up might take the focus off the recent cases and take effort away from finding new ways to find fraud.

> *"The last thing in the world I need right now is to detect more fraud."*
> – Senior Health Care Financing Administration (HCFA) official. (Sparrow, 2000, p. 228)

2.3.1 Organizations and companies specialized in fraud

The government does not fight the battle for fraud alone. Private health insurers also suffer significant from fraud and this resulted in numerous organizations and companies.

One of the most recent initiatives from the industry is the Health Care Cost Center institute. This is an institute that has access to roughly three billion health insurance claims for more than 33 million individuals. This data was contributed by a set of large health insurers who collectively represent almost 40 percent of the US private health insurance market. The goal of this institute is to provide access to an unprecedented amount of health care cost and utilization data to researchers and policymakers trying to understand the factors influencing health care costs.

On the profit side only a few companies focus specifically on fraud detection within the health care. They offer services like prepayment fraud detection, outlier detection, social network analysis, advanced text analytics and data mining.

2.4 Literature about fraud-finding in the health care

The literature about fraud within the health care can be divided into three categories.

2.4.1 Overview of the field

The first category provides an overview of the field. It focuses on what kind of statistical methods can be used. For example Travaille et al. (2011) have created an overview of statistical methods used by fraud detection within other industries and how they can be applied within the health care industry (see

Figure 7). Li et al. (2008) have surveyed used methods within the health care industry. Li writes that combinations of unsupervised and supervised methods are used in combination with profiling. This category helps in getting an overview of the different methods that are possible for fraud detection.

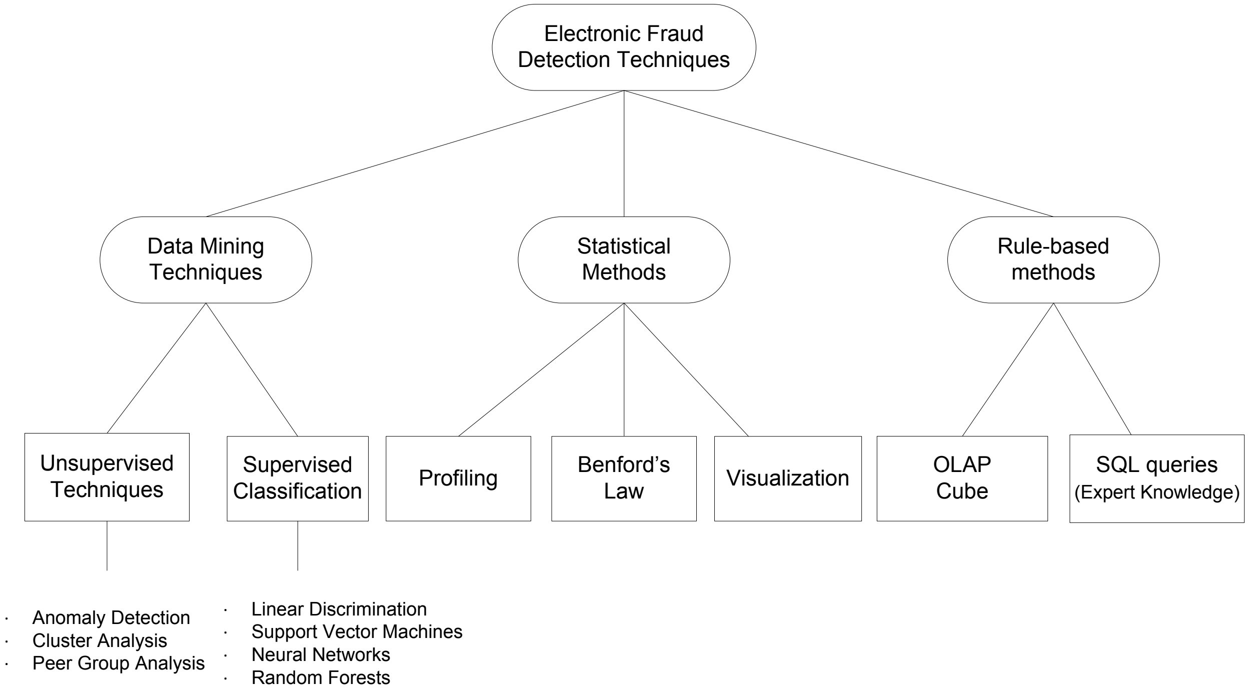


Figure 7 Framework of Electronic Fraud Detection Techniques (Travaille, Müller, Thornton, & Hillegersberg, 2011)

2.4.2 Case studies

The second category is providing results on actual applications of the methods to find its usefulness in detecting fraud. For example Copeland et al. (2011) wrote a paper about the use of unsupervised methods to find Medicaid fraud within Nevada. Yang, et al. (2006) has done research to do fraud detection by looking at the order of which services are performed. This category helps in choosing a method for fraud detection by being able to compare the results of individual methods.

2.4.3 How to do fraud detection

The third category, including this thesis, is focused on how to improve the current health care systems to prevent fraud and improve fraud detection. For example Morris (2009) describes five key components how the current health system has to change to better battle fraud. Major and Riedinger (2002) describe a workflow and system to setup fraud detection departments with results of its use in the real world. Similar work has been done by Ortega et al. (2006) who introduced a data-mine based system that decreased the time it takes to detect fraud by 76% from an average of 8,6 months to two months.

Because Major and Riedinger (2002) and Ortega et al. (2006) describe real systems that are used to find fraud they cannot go into details of the exact working of the systems. Doing this would give fraud perpetrators an advantage on penetrating the fraud defense.

3 Data warehouse to find Health Care Fraud

The system will have to support the fraud detection tools. These tools can differ between basic query tools to analytical applications like SAS. To support the detection of fraud the system will be a dimensional data warehouse. Using a data warehouse as defined by Kimball & Ross (2008) prescribe certain elements to be available as can be seen in Figure 8.

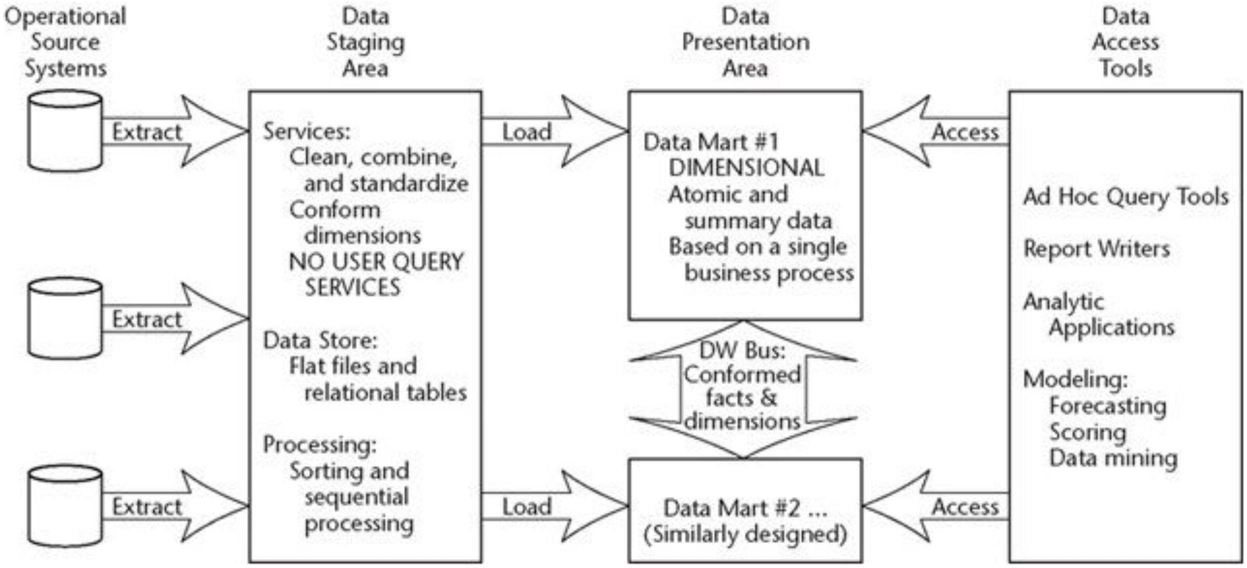


Figure 8 Basic elements of the data warehouse (Kimball & Ross, 2008)

The first step is to gather all the data from different sources in the staging area. Besides the use of the claim data from the source systems also extra information on for example providers, drugs and services will be gathered. Other activities within the staging phase are sorting of the data and pre-calculating extra attributes that can be useful for fraud detection. An example is the calculation of the distance between the provider and the patient.

To develop a data warehouse the four-step dimensional design process as defined by Kimball & Ross (2008) and shown in Figure 9 will be used.

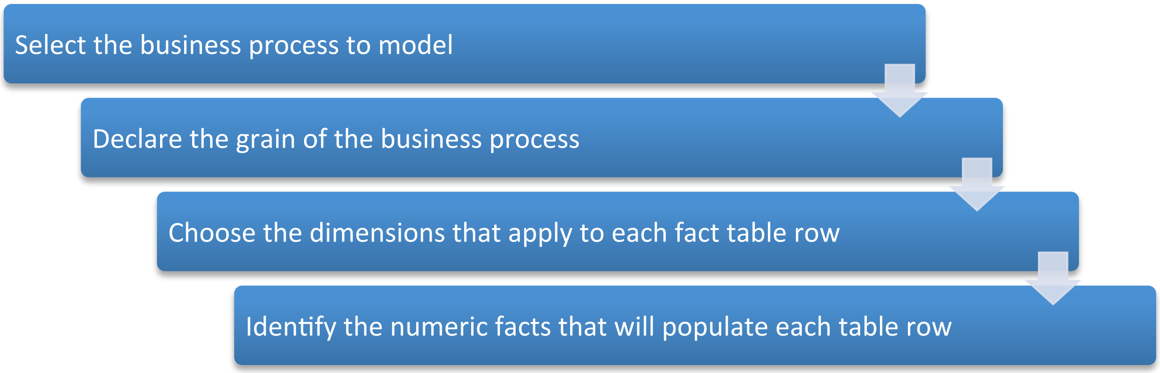

Select the business process to model
Declare the grain of the business process
Choose the dimensions that apply to each fact table row
Identify the numeric facts that will populate each table row

Figure 9 Four-Step Dimensional Design Process (Kimball & Ross, 2008)

3.1 Defining the requirements for the data warehouse

The first step is to select the business process to model. That is easy, that will be fraud detection. Within fraud detection there are analysts working with the data looking for fraudulent providers. Fraud is being found by asking questions that can be answered by analyzing the data.

3.1.1 Declaring the grain of the business process

Declaring the grain means defining the finest detail that should be able in the data. Kimball & Ross (2008) state that it is preferred to develop the models for the most atomic information captured by a business process. This means that the data cannot be subdivided any further.

A user using a data warehouse will usually start with generating an overview of the data but when it sees anomalies it will want to be able to drill down in the data and find the specific details what is causing the anomaly. This means that the finest detail has to be very specific.

3.1.1.1 Source information

The data that is available in the source systems first has to be analyzed to be able to determine the level of grain that is available.

The source data is used instead of defining a new set of data to be used for fraud detection. The reason for this is that it is nearly impossible for organizations to change the way their source systems are organized and executed. Let alone that there will be incentive to change it to optimize fraud detection. So the data model for fraud detection is limited to the information that is already collected plus external reference data.

The focus of the data warehouse is to support fraud detection within Medicaid. This is a government paid health insurance and has a lot of public data available regarding the processing of the claims.

Medicaid uses four different claim forms for providers to submit claims to the source system. These will be discussed in the following sections. See appendix A-D for paper versions of these claim forms.

HCFA-1500 (CMS1500)
This form is used by professionals (for example physicians, chiropractors) to bill for services rendered.

Information available on each claim:

- Patient information (name, address, birth date, phone, relationship status)
- Is patient condition related to employment, auto accident or other accident?
- Other coverage
- Date of claim
- Date if patient had same or similar illness before
- Referring physician or other source (Name, NPI)
- Dates patient unable to work in current occupation
- Hospitalization dates related to current services
- Place of treatment

- Outside lab
- Total charge ($)
- Amount paid ($)
- Balance due ($)
- Quality Improvement Organization (QIO) prior authorization number
- Diagnosis or nature of illness or injury (1-4 times)
- Service (1-6 times):
    - From date
    - To date
    - Place of service
    - Emergency?
    - CPT/HCPCS code of procedure, service or supply
    - Modifier (1-4 times)
    - Related diagnosis
    - Charges ($)
    - Days or units
    - Provider NPI

J400
J400 is the form used to bill for dental services.

Information available on each claim:

- Patient information
- Provider information
- Service (1-10 times):
    - Procedure date
    - Area of Oral Cavity
    - Tooth System
    - Tooth numbers or letters
    - Tooth surface
    - Procedure code
    - Description
    - Fee
- Missing teeth information
- Remarks
- Place of treatment (provider’s office, hospital, ECF, other)
- Number of enclosures (radiographs, oral images, models)
- Treatment for orthodontics?
- Replacement of prosthesis?
- Treatment resulting from occupational illness/injury, auto accident or other accident?

UB-04
Institutional claim filling form used for inpatient and long-term care. This form is used by:

- Ambulatory Surgery Centers (ASC)
- End-Stage Renal Disease (ESRD)
- Clinics
- Home Health Agencies (HHA)
- Hospice Providers
- Hospitals
- Long-Term Care (LTC) Facilities
- Rehabilitation Hospital Facilities

Information available on each claim:

- Patient information
- Payer information
- Insurer information
- Provider information
- Statement covering period
- Admission
- Condition codes + date occurrence from-through (1-10 times):
- Value codes and amount
- Revenue codes
    - Revenue code
    - Description
    - NDC Code
    - HCPCS/Rates
    - Service date
    - Units of service
    - Total charges
    - Non-Covered charges
- Treatment authorization codes
- Diagnosis/Procedure code qualifier
- Admitting Diagnosis Code
- Patient's Reason for Visit code
- PPS Code
- Externa Cause of Injury Code
- Principial Procedure Code + Date
- Other Procedure Code + Date (5 times)

Drug Claim Form
This form is used to bill for medicines. It can hold multiple claims for different patients at the same time.

Information available on each claim:

- Patient information (First name, last name, Medicaid Recipient Identification Number (RID))
- Prescribing provider (NPI-number)
- Was it used for an emergency? (yes/no)
- Is the patient pregnant and is the medication related to the pregnancy? (yes/no)
- Patient residence situation
  - Not specified
  - Skilled Nursing Facility
  - Nursing Facility
  - Inpatient Psychiatric Facility
  - ICF/MR
  - Hospice
- Prescription number
- Brand medically necessary indicator
  - No product selection indicated
  - Substitution Allowed – Patient requested product dispensed
  - Substitution Allowed – pharmacist selected product dispensed
  - Substitution Allowed – Generic drug not in stock
  - Substitution Allowed – Brand drug dispensed as a generic
  - Substitution Allowed – Generic drug not available in market
  - Substitution Allowed by prescriber but plan requests brand
- Number of times this prescription is refilled (0 for first time)
- Quantity dispensed
- Approximate number of days supply for the quantity of the drug dispensed
- Total amount charged for the prescription dispensed
- Date prescribed
- Date dispensed
- NDC number for the drug(s) dispensed
- Other coverage (type, amount paid, patient responsibility amount)
- Provider's info (name, address, NPI, type, signature)
- Date billed

Some of the terms in the source data require some extra explanation:

NPI stands for National Provider Identifier. A short description by CMS (Centers for Medicare & Medicaid Services, 2012) NPI is a Health Insurance Portability and Accountability Act (HIPAA) Administrative Simplification Standard. The NPI is a unique identification number for covered health care providers. Covered health care providers and all health plans and health care clearinghouses must use the NPIs in the administrative and financial transactions adopted under HIPAA.

When a provider registers it is able to give up to three taxonomy numbers describing its primary and other activities. As of date there are 828 different taxonomies (for example Ambulance, Nuclear Pharmacy, Oncology) (Washington Publishing Company, 2012). The list is maintained by the National Uniform Claim Committee Code Subcommittee.

HCPCS/CPT code refers to a standardized coding system for services provided. A short description by CMS (Centers for Medicare & Medicaid Services, 2012): HCPCS is divided into two principal subsystems, referred to as level I and level II of the HCPCS. Level I of the HCPCS is comprised of CPT (Current Procedural Terminology), a numeric coding system maintained by the American Medical Association (AMA). The CPT is a uniform coding system consisting of descriptive terms and identifying codes that are used primarily to identify medical services and procedures furnished by physicians and other health care professionals. Level II of the HCPCS is a standardized coding system that is used primarily to identify products, supplies, and services not included in the CPT codes, such as am (Centers for Medicare & Medicaid Services, 2012)bulance services and durable medical equipment, prosthetics, orthotics, and supplies (DMEPOS) when used outside a physician's office.

An HCPCS/CPT code can have modifiers that represent extra complications that happened while providing the service. There are 4 modifications allowed per HCPCS code. The price is based on the HCPCS and the adjustments.

NDC number is the number under which the drug is known by the Food and Drug Administration in their National Drug Code Directory (Food and Drug Administration, 2012).

Diagnoses follow the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) standard. A short description by CDC (Center for Disesae Control and Prevention, 2012): ICD-9-CM is based on the World Health Organization's Ninth Revision, International Classification of Diseases (ICD-9). ICD-9-CM is the official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States.

3.1.1.2 Granularity

To keep the data accessible from various claim forms four different claim types: inpatient, long term, pharmacy and professional/other will be introduced.

A general core that each claim exists of can be extracted among the different claim forms: a patient, a provider, lines with several diagnoses, procedures and amounts charged.

It has been decided that the fact table will represent a single line from a claim to offer the most flexibility to the user in his quest for finding fraud. As shown later, this will keep performance and quality of detail high. For each claim-line a type that defines in which table type-specific information can be found will be defined. This is called the header of a form.

The source systems do not only contain the claims. Whenever a mistake has been made an adjustment form will be filled and becomes part of the source system data. The added value of adjustments is very little and does not balance out the extra complexity and required space that it causes within the data warehouse. Thus whenever an adjustment will come in the effected row will be updated and the old information discarded. This is called a Type 1 update strategy by Kimball & Ross (2008).

3.1.2 Choose the dimensions that apply to each fact table row

The next step in the development of the data warehouse is choosing the dimension. How should the data be sliced? To define this dimensions are chosen so that the seven levels of fraud detection by Sparrow are supported. This is then combined with the proposed dimensions by Kimball & Ross (2008) and results in the following dimensions:

- Calendar date (claim filled, service rendered, claim paid)
- Provider (executing, referring)
- Patient
- Health plan
- Treatment
- Diagnosis
- Claim type
- Drug
- Type
- Outcome
- Location (office, clinic, outpatient facility, hospital)

Some dimensions will not be available for every claim. They are still included in the data warehouse as they can offer extra functionality when only on certain information in the data warehouse is focused. The outcome dimension is based on the possible outcomes (discharge status) which is only available in the UB-04 forms. The location dimension is given in the CMS1500 form, among others, in the form of a CPT-code.

All dimensions will be treated as slowly changing dimensions. To keep track of this in the data warehouse a Type 2 update strategy (Kimball & Ross, 2008) will be used. A Type 2 update strategy will add a new row to the dimension whenever an update is provided. All new facts will point to the new record. This way when an update is provided for a patient all claim lines that relate to that patient are pointing at the relevant record. When you want to have the full claim record of a patient one only has to filter the patient dimension by SSN (the primary key of a patient) and get all related dimension records.

3.1.3 Identify the numeric facts that will populate each table row

The next step in the development of the data warehouse is to define which numeric facts will populate each table row and what their nature is. Numeric facts are crucial according to Kimball & Ross (2008) because data warehouse applications almost never retrieve a single fact table row. Rather, they bring back hundreds, thousands, or even millions of fact rows at a time, and the most useful thing to do with so many rows is to add them up.

Within numerical facts Kimball & Ross (2008) distinguishes between additive, semiadditive and nonadditive facts. Additive facts can be summed when grouping several facts, independent of which dimensions are used for grouping. Semiadditive facts can only be grouped along side specific dimensions. Nonadditive facts cannot be summed but only support other methods like counts or averages.

A numeric fact can either be computed based on other facts/dimensions within the same row or come from one of the source systems. For performance it is advised by Kimball & Ross (2008) that the computed values are saved physically in the database.

Looking at the different data that is available from the source system the following numeric facts can be distinguished:

- Covered charges ($), *additive*
- Non-covered charges ($), *additive*
- Total charges ($), *computed additive*
- Units of service, *additive*
- Covered price per unit, *computed nonadditive*
- Total price per unit, *computed nonadditive*
- Treatment duration, *computed nonadditive*
- Number of days between service rendered and claim filled, *computed nonadditive*
- Number of days between claim filled and claim paid, *computed nonadditive*
- Number of days between service rendered and claim paid, *computed nonadditive*
- Distance between provider and patient in kilometers, *computed nonadditive*

The facts that are computed will be computed during the staging phase.

3.2 Data Model

To keep the complexity of the data model low and to increase performance of the database not everything will be normalized. This is for example why a provider will have three specialization attributes instead of normalizing this to an external table. This is in line with the advice from Kimball & Ross (2008) who states that normalized modeling is helpful for operational processing but will confuse users and database query optimizers.

Another advice from Kimball & Ross (2008) is to minimize the use of codes in the dimension tables and replace them with more verbose textual attributes. An example of this is in the provider table where the name of the taxonomy will be used instead of the code. This advice will help new users understand the system by reducing the required domain knowledge.

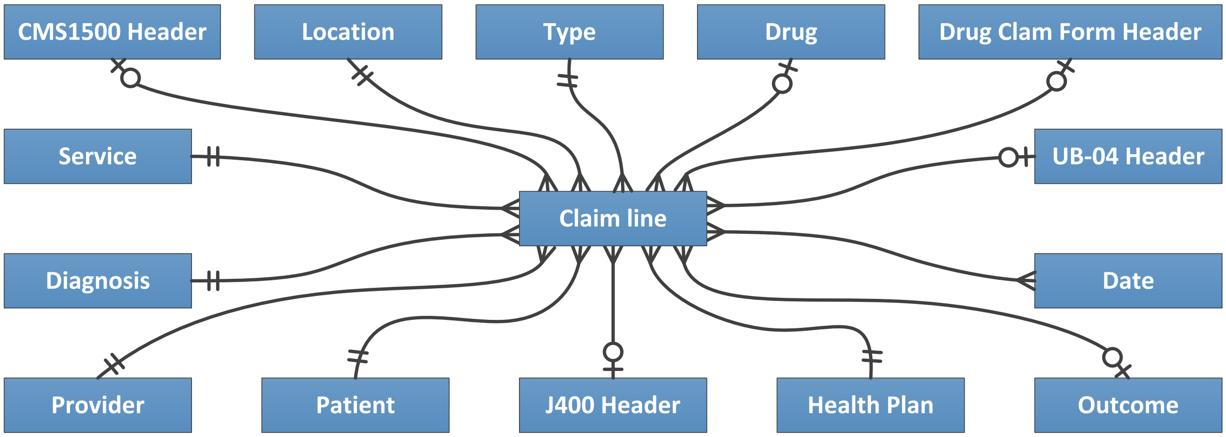


Figure 10 Proposed data warehouse model to support fraud detection

The data warehouse consists of one fact table “Claim line” which can be dissected on several dimensions to get necessary data. In the next sections each entity and its attributes will be described in detail.

3.2.1 Claim line

Claim line is a fact-table. Each row represents a single line on a claim.

Facts

Covered charges	Amount charged that is covered by the insurance in dollars.
Non-covered charges	Amount charged that is not covered by the insurance in dollars.
Total charges	Total amount charged in dollars
Units of service	Units of service that have been provided.
Covered price per unit	Price per unit in dollars.
Total price per unit	Price per unit in dollars.
Treatment duration	Amount of days the service took place.
Number of days service – claim filled	Amount of days between the service end date and the date the claim is filled.
Number of days claim filled – claim paid	Amount of days between the date the claim is filled and the claim is paid.
Number of days service – claim paid	Amount of days between the service end date and the date the claim is paid.
Distance between provider and patient in kilometers	The distance in kilometers between the location of the patient and the provider.

Dimensions

Date (Service Start)	The date the service started being provided to the patient.
Date (Service End)	The date the service stopped being provided to the patient.
Date (Claim filled)	The date the claim is filled.
Date (Claim Paid)	The date the claim is paid.
Diagnosis	The diagnosis that is done and lead to the provider provide the service to the patient.
Service	The medical treatment or supplies that are provided by the provider.

Patient	The patient being diagnosed and receiving the treatment.
Provider	The provider that delivered a service to the patient.
Referring/prescribing provider	The provider that prescribed the drug or referred the patient to the provider.
Drug	Drugs related to the claim, if any.
Outcome	The outcome of the service.
Location	The location the service took place.

3.2.2 Date dimension

Kimball & Ross (2008) prescribe to model dates via a dimension. The reason why is to be able to hold the calendar logic in the dimension table and not move it to application code.

The date dimension represents a unique row per day. Besides the date it contains extra information about that date.

Attributes

Date Primary Key	Primary key identifying this date record.
Date	The date.
Full Date Description	Full description of the date, e.g. “January 5, 2012”.
Day of Week	Day of Week of the date.
Day Number in Calendar Month	Day Number in Calendar Month of the date.
Day Number in Calendar Year	Day Number in Calendar Year of the date.
Last Day in Week Indicator	Last Day in Week Indicator of the date.
Last Day in Month Indicator	Last Day in Month Indicator of the date.
Calendar Week Ending Date	Calendar Week Ending Date of the date.
Calendar Week Number in Year	Calendar Week Number in Year of the date.
Calendar Month Name	Calendar Month Name of the date.
Calendar Year-Month	Calendar Year-Month of the date (YYYY-MM).
Calendar Quarter	Calendar Quarter of the date.
Calendar Year-Quarter	Calendar Year-Quarter of the date.
Calendar Half Year	Calendar Half Year of the date.
Calendar Year	Calendar Year of the date.
Weekday Indicator	Indicates if the date is during a weekday.

3.2.3 Patient dimension

A patient is a person that is receiving medical treatment or supplies.

Updates to this row will be done via a Type 2 update strategy so history will be preserved.

Attributes

Patient Primary Key	Primary key identifying this patient record.
Social Security Number	Social security number identifying the patient.
First Name	First name of the patient.
Last Name	Last name of the patient.
Middle Initial	Middle initials of the patient.
Sex	Gender of the patient.
Birthdate	Birthdate of the patient.

Death indicator	Indicator if patient is death.
Date of Death	Date of death of the patient.
Status	Status of the patient: • Single • Married • Other • Employed • Full-Time Student • Part-Time Student
Ethnicity	Ethnicity of the patient.
Location Address line 1	Address line 1 of the patient.
Location Address line 2	Address line 2 of the patient.
Location Zip code	Zip code of the patient.
Location City	City of the patient.
Location County	County of the patient.
Location State	State of the patient.
Location Country	Country of the patient.
Location latitude	Latitude of the location of the patient.
Location longitude	Longitude of the location of the patient.
Health Plan	The health plan the patient is enrolled to.
Health Plan ID	Identification number of the policy.
Health Plan enrollment date	The date the patient was enrolled.
Health Plan cancel date	The date the patient cancelled his health plan.

3.2.4 Provider dimension

A provider is providing the medical treatment and/or supplies to a patient. Data in this dimension is based on the NPI definition (Department of Health and Human Services, 2004).

Updates to this row will be done via a Type 2 update strategy so history will be preserved.

Attributes

Provider Primary Key	Primary key to identify the row
NPI	10-position all-numeric identification number assigned by the NPS to uniquely identify a health care provider
Entity type	One of the following: • Person • Non-person
Replacement National Provider Identifier	The most recent NPI issued by the NPS to this provider.
Previous National Provider Identifier.	The NPI that had previously been issued to this provider.
Provider SSN	The SSN assigned by the Social Security Administration (SSA) to the individual being identified.
Provider IRS Individual Taxpayer Identification Number	The taxpayer identifying number assigned by the IRS (to individuals who are not eligible to be assigned SSNs) to the individual being identified.
Provider Employer Identification	The Employer Identification Number (EIN), assigned by the IRS, of

Number	the provider being identified.
Provider last name or organization name	The last name of the provider (if an individual) or the name of the organization provider. If the provider is an individual, this is the legal name. If the provider is an organization.
Provider first name	The first name of the provider, if the provider is an individual.
Provider middle name	The middle name of the provider, if the provider is an individual.
Provider first line mailing address	The first line mailing address of the provider being identified.
Provider second line mailing address	The second line mailing address of the provider being identified.
Provider mailing address City name	The State or Province name in the mailing address of the provider being identified.
Provider mailing address County name	The County name in the mailing address of the provider being identified.
Provider mailing address State name	The State or Province name in the mailing address of the provider being identified.
Provider mailing address postal code	The postal ZIP or zone code in the mailing address of the provider being identified. NOTE: ZIP code plus 4-digit extension, if available.
Provider mailing address postal code without extension	The postal ZIP or zone code in the mailing address of the provider being identified without a 4-digit extension.
Provider mailing address country code	The country code in the mailing address of the provider being identified.
Provider mailing address latitude	The latitude of the mailing address of the provider being identified.
Provider mailing address longitude	The longitude of the mailing address of the provider being identified.
Provider first line location address	The first line location address of the provider being identified. For providers with more than one physical location, this is the primary location.
Provider second line location address	The second line location address of the provider being identified. For providers with more than one physical location, this is the primary location.
Provider location address city	The city name in the location address of the provider being identified.
Provider location address County name	The county name of the location address of the provider being identified.
Provider location address State name	The State or Province name in the location address of the provider being identified.
Provider location address postal code	The postal ZIP or zone code in the location address of the provider being identified. NOTE: ZIP code plus 4-digit extension, if available
Provider location address postal code without extension	The postal ZIP or zone code in the location address of the provider being identified without a 4-digit extension.
Provider location address country code	The country code in the location address of the provider being identified.
Provider location address latitude	The latitude of the location address of the provider being identified.
Provider location address longitude	The longitude of the location address of the provider being identified.

Provider primary taxonomy	The primary provider type, classification, and specialization. Taxonomies are from the Healthcare Provider Taxonomy code list.
Provider primary taxonomy license number	The license number for the primary taxonomy issued to the provider being identified.
Provider primary taxonomy license number State	The State that issued the license for the primary taxonomy.
Provider other taxonomy	The other provider type, classification, and specialization. Taxonomies are from the Healthcare Provider Taxonomy code list.
Provider other taxonomy license number	The license number for the other taxonomy issued to the provider being identified.
Provider other taxonomy license number State	The State that issued the license for the other taxonomy.
Provider other2 taxonomy	The other2 provider type, classification, and specialization. Taxonomies are from the Healthcare Provider Taxonomy code list.
Provider other2 taxonomy license number	The license number for the other2 taxonomy issued to the provider being identified.
Provider other2 taxonomy license number State	The State that issued the license for the other2 taxonomy.
Provider enumeration date	The date the provider was assigned a unique identifier (assigned an NPI).
Last update date	The date that a record was last updated or changed.
NPI deactivation reason	The reason that the provider's NPI was deactivated in the NPS: • Death of provider (personal providers only) • Disbandment (non-personal providers only) • Fraud • Other
NPI deactivation date	The date that the provider's NPI was deactivated in the NPS.
NPI reactivation date	The date that the provider's NPI was reactivated in the NPS.
Authorized official name	The name of the person authorized to submit the NPI application or to change NPS data for a health care provider.
Authorized official title or position	The title or position of the authorized official.

3.2.5 Health Plan dimension

A health plan that provides coverage to patients.

Updates to this row will be done via a Type 2 update strategy so history will be preserved.

Attributes

Health Plan Primary Key	Primary key to identify the health plan.
Name	Name of the health plan.
Organization	Organization that maintains the health plan.

3.2.6 Service dimension

This is a service that is provided by a provider to a patient. This can be either medical treatment or medical supplies. Services are modeled after the HCPCS and CPT standards as discussed in section 3.1.1.1 combined with Ambulatory Payment Classifications, another set of codes used for reporting services. Each service can have up to 4 modifiers applied to it. A modifier provides the means by which the reporting physician or provider can indicate that a service or procedure that has been performed has been altered by some specific circumstance but not changed in its definition or code.

Updates to this row will be done via a Type 2 update strategy so history will be preserved.

Attributes

Service Primary Key	Primary key to identify this row.
Code	Code to identify the service.
Name	Name of the service
Type	One of: • CPT • HCPCS • NDC
Modifier 1	First modifier, if applied.
Modifier 2	Second modifier, if applied.
Modifier 3	Third modifier, if applied.
Modifier 4	Fourth modifier, if applied.
Group name	Name of the group where this service belongs.
Pricing Indicator Code 1	Methodology for developing unique pricing amounts.
Pricing Indicator Code 2	Methodology for developing unique pricing amounts.
Pricing Indicator Code 3	Methodology for developing unique pricing amounts.
Pricing Indicator Code 4	Methodology for developing unique pricing amounts.
Berenson-Eggers Type Of Service	The Berenson-Eggers Type of Service (BETOS) for the procedure based on generally agreed upon clinically meaningful groupings of procedures and services.
Service added	The date the service was added to the system.
Service canceled	The date this specific service cannot be used anymore.
Can service be used indicator	Indicator if the service is not canceled.

3.2.7 Diagnosis dimension

This represents the diagnosis of the problem that has been made by a provider and which was the reason a treatment was given. The diagnosis dimension is based on the ICD-9-CM standard as discussed in section 3.1.1.1 combined with the Diagnosis Related Group codes. The diagnoses are a hierarchical tree where each branch goes into more details about what the diagnoses entailes.

Updates to this row will be done via a Type 2 update strategy so history will be preserved.

Attributes

Diagnosis Primary Key	Primary key to identify the row.
Code	Code to represent the diagnosis
Name	Name of the diagnosis
Parent diagnosis	Under which parent diagnosis does the diagnosis fall

3.2.8 Drug dimension

The drug dimension represents the involved drugs in the claim. This data is available on the Drug Claim Form and on the UB-04 forms. It is based on the NDC-standard.

Attributes

Drug Primary Key	Primary key to identify the row.
Drug active ingredient	Active ingredient of the drug.
Drug administration route	Route the drug has to be taken.
Drug dosage form	The form the drug has to be taken.
Drug labeler name	Company that labels the drugs.
Drug product type	Type of the drug.
Drug marketing category	The marketing category of the drug.
Drug size & type	Drug doses per container and container per box.

3.2.9 Location dimension

The location dimension represents were the claim line has taken place. This dimension is based on the locations within the CMS1500 forms. This dimension can be used to distinguish between the types of claims.

Attributes

Location Primary Key	Primary key to identify the row.
Name	Name of the location.
Code	The code that represents the location.

3.2.10 Outcome dimension

This dimension represents the outcome of the claim. This dimension is only available for UB-04 claims.

Attributes

Location Primary Key	Primary key to identify the row.
Description	Description of the outcome

3.2.10 CMS1500 Header dimension

This dimension provides the header for a line from a CMS1500 claim.

For claim information the adjustments available in the source system are directly applied to the claim-lines to keep things simple. A Type 1 update strategy will be used on this dimension. This means that no history will be preserved.

Attributes

CMS1500 Primary Key	The primary key of this row.
Claim ID	ID of the claim
Patient's condition relation	Patient's relation is related to: • Nothing • Employment • Auto Accident • Other accident
Date of first symptoms/accident	Date the patient noticed the first symptoms.
Date of same or similar condition	Date the patient had the same or similar condition.
Start date patient unable to work	The start date that the patient is unable to work from.
End date patient unable to work	The end date that the patient is unable to work from.
Start date hospitalization related	The start date that the patient got hospitalized because of this

to service	claim.
End date hospitalization related to service.	The end date that the patient got hospitalized because of this claim.
Outside lab indicator	Indicates if the service took place outside of the lab.
Prior authorization number	Number of the prior authorization that was given.

3.2.11 Drug Claim Form Header dimension

This dimension provides the header for a line from a Drug Claim Form claim.

For claim information the adjustments available in the source system are directly applied to the claim-lines to keep things simple. A Type 1 update strategy will be used on this dimension. This means that no history will be preserved.

Attributes

Drug Claim Form Primary Key	The primary key of this row.
Claim ID	ID of the claim
Emergency indicator	Indicates if it was an emergency.
Pregnancy indicator	Indicates if the patient is pregnant and the medication is related to the pregnancy.
Patient Residence	One of the following: • Not Specified • Skilled Nursing Facility • Nursing Facility • Inpatient Psychiatric Facility • ICF/MR Hospice
Prescription number	The number of the prescription.
Brand medically necessary indicator	Indicates if the brand is important. Can have following values: • No product selection indicated • Substitution Allowed – Patient requested product dispensed • Substitution Allowed – pharmacist selected product dispensed • Substitution Allowed – Generic drug not in stock • Substitution Allowed – Brand drug dispensed as a generic • Substitution Allowed – Generic drug not available in market Substitution Allowed by prescriber but plan requests brand
Refill number	For original prescription it is 0. Else indicates number of the refill.
Quantity dispensed	Indicate the quantity of the drug dispensed.
Days supply	The approximate number of days supply for the quantity of the drug dispensed.
Date prescribed	The date the prescription was written.

3.2.12 UB-04 Header dimension

This dimension provides the header for a line from a UB-04 claim.

For claim information the adjustments available in the source system are directly applied to the claim-lines to keep things simple. A Type 1 update strategy will be used on this dimension. This means that no history will be preserved.

Attributes

UB-04 Primary Key	The primary key of this row.
Claim ID	ID of the claim
Start claim date	Date the services billed on the claim started.
End claim date	Date the services billed on the claim ended.
Type of facility	• Hospital • Skilled Nursing • Home Health Facility • Religious Non-medical Health Care Institutions (RNHCI) – Hospital Inpatient • Reserved for National Assignment by the NUBC • Intermediate Care (not used for Medicare) • Clinic • Special Facility or ASC Surgery
Type of claim	If type of facility is a clinic: • Rural Health Clinic • Clinic – Hospital Based or Independent Renal Dialysis Center • Freestanding • ORF • CORF • CMHC • Federally Qualified Health Center (FQHC) If type of facility is a special facility: • Hospice (Non-hospital based) • Hospice (Hospital based) • Ambulatory Surgery Center • Freestanding Birthing Center • Critical Access Hospital • Residential Facility (Not used for Medicare) • Reserved for National Assignment by NUBC • Reserved for National Assignment by NUBC • Special Facility - Other (Not used for Medicare) Otherwise: • Inpatient (Including Medicare Part A) • Inpatient (Medicare Part B Only) (Includes HHA Visits Under a Part B Plan of Treatment) • Outpatient (Includes HHA Visits Under a Part A Plan of Treatment Including DME Under Part A) • Laboratory Services Provided to Non-Patients, or Home Health Not Under a Plan of Treatment • Intermediate Care Level 1

	• Intermediate Care Level II • Swing Beds
Frequency of the claim	• Nonpayment / Zero Claim • Admit through Discharge Claim • Interim – First Claim • Interim – Continuing Claim (Not valid for Medicare PPS Claims) • Interim – Last Claim (Not valid for Medicare Inpatient Hospital PPS Claims) • Late Charges Only Claim • Reserved for National Assignment by NUBC • Replacement of Prior Claim • Void / Cancel of a Prior Claim • Final Claim for a Home Health PPS Episode
Admission date	Date the care begins.
Admission hour	Hour the care begins.
Discharge hour	Hour patient is discharged.
Priority of visit	• Emergency • Urgent • Elective • Newborn • Trauma Center • Information Not Available
Source of referral for admission or visit	If priority of visit is Newborn: • Born inside this hospital • Born outside this hospital Otherwise: • Nonhealthcare Facility Point of Origin • Clinic or Physician's Office • Reserved for assignment by the NUBC • Transfer From a Hospital (Different Facility) • Transfer from a Skilled Nursing Facility or Intermediate Care Facility or Assisted Living Facility • Transfer from Another Health Care Facility • Court/Law Enforcement • Information Not Available • Transfer from One Distinct Unit of the Hospital to Another Distinct Unit of the Same Hospital Resulting in a Separate Claim to the Payer • Transfer from Ambulatory Surgery Center • Transfer from Hospice Facility
Condition 1	Conditions related to the claim that may effect the processing of it.
Condition 2	Conditions related to the claim that may effect the processing of it.
Condition 3	Conditions related to the claim that may effect the processing of it.
Condition 4	Conditions related to the claim that may effect the processing of it.
Condition 5	Conditions related to the claim that may effect the processing of it.

Condition 6	Conditions related to the claim that may effect the processing of it.
Condition 7	Conditions related to the claim that may effect the processing of it.
Condition 8	Conditions related to the claim that may effect the processing of it.
Condition 9	Conditions related to the claim that may effect the processing of it.
Condition 10	Conditions related to the claim that may effect the processing of it.
Occurrence 1	Significant event 1 associated with the claim that affects processing by the payer (e.g., accident, employment related, etc.).
Occurrence 1 date	Date of significant event 1.
Occurrence 2	Significant event 1 associated with the claim that affects processing by the payer (e.g., accident, employment related, etc.).
Occurrence 2 date	Date of significant event 2.
Occurrence 3	Significant event 3 associated with the claim that affects processing by the payer (e.g., accident, employment related, etc.).
Occurrence 3 date	Date of significant event 3.
Occurrence 4	Significant event 4 associated with the claim that affects processing by the payer (e.g., accident, employment related, etc.).
Occurrence 4 date	Date of significant event 4.
Treatment authorization code 1	Authorization code if treatment is pre-authorized.
Treatment authorization code 2	Authorization code if treatment is pre-authorized.
Treatment authorization code 3	Authorization code if treatment is pre-authorized.
Extra diagnosis	
Extra diagnosis 1	Extra diagnosis if applicable.
Extra diagnosis 2	Extra diagnosis if applicable.
Extra diagnosis 3	Extra diagnosis if applicable.
Extra diagnosis 4	Extra diagnosis if applicable.
Extra diagnosis 5	Extra diagnosis if applicable.
Extra diagnosis 6	Extra diagnosis if applicable.
Extra diagnosis 7	Extra diagnosis if applicable.
Extra diagnosis 8	Extra diagnosis if applicable.
Extra diagnosis 9	Extra diagnosis if applicable.
Extra diagnosis 10	Extra diagnosis if applicable.
Extra diagnosis 11	Extra diagnosis if applicable.
Extra diagnosis 12	Extra diagnosis if applicable.
Extra diagnosis 13	Extra diagnosis if applicable.
Extra diagnosis 14	Extra diagnosis if applicable.
Extra diagnosis 15	Extra diagnosis if applicable.
Extra diagnosis 16	Extra diagnosis if applicable.
Extra diagnosis 17	Extra diagnosis if applicable.
Extra diagnosis 18	Extra diagnosis if applicable.
Admitting diagnosis	Diagnosis of the patient at the time of admission.
Reason to visit 1	Patient’s reason for an outpatient visit (from ICD-9-CM standard)
Reason to visit 2	Patient’s reason for an outpatient visit (from ICD-9-CM standard)
Reason to visit 3	Patient’s reason for an outpatient visit (from ICD-9-CM standard)
External Cause of Injury	In the case of external causes of injuries, poisonings, or adverse affects, the appropriate ICD-9- CM diagnosis is reported in this field.

Attending provider	Provider responsible for the care.
Operating provider	Provider responsible for performing surgical procedures.

3.2.13 J400 Header dimension

This dimension provides the header for a line from a J400 claim.

For claim information the adjustments available in the source system are directly applied to the claim-lines to keep things simple. A Type 1 update strategy will be used on this dimension. This means that no history will be preserved.

Attributes

J400 Primary Key	The primary key of this row.
Claim ID	ID of the claim
Missing tooth 1	Indicates if tooth 1 is missing
Missing tooth 2	Indicates if tooth 2 is missing
Missing tooth 3	Indicates if tooth 3 is missing
Missing tooth 4	Indicates if tooth 4 is missing
Missing tooth 5	Indicates if tooth 5 is missing
Missing tooth 6	Indicates if tooth 6 is missing
Missing tooth 7	Indicates if tooth 7 is missing
Missing tooth 8	Indicates if tooth 8 is missing
Missing tooth 9	Indicates if tooth 9 is missing
Missing tooth 10	Indicates if tooth 10 is missing
Missing tooth 11	Indicates if tooth 11 is missing
Missing tooth 12	Indicates if tooth 12 is missing
Missing tooth 13	Indicates if tooth 13 is missing
Missing tooth 14	Indicates if tooth 14 is missing
Missing tooth 15	Indicates if tooth 15 is missing
Missing tooth 16	Indicates if tooth 16 is missing
Missing tooth 17	Indicates if tooth 17 is missing
Missing tooth 18	Indicates if tooth 18 is missing
Missing tooth 19	Indicates if tooth 19 is missing
Missing tooth 20	Indicates if tooth 20 is missing
Missing tooth 21	Indicates if tooth 21 is missing
Missing tooth 22	Indicates if tooth 22 is missing
Missing tooth 23	Indicates if tooth 23 is missing
Missing tooth 24	Indicates if tooth 24 is missing
Missing tooth 25	Indicates if tooth 25 is missing
Missing tooth 26	Indicates if tooth 26 is missing
Missing tooth 27	Indicates if tooth 27 is missing
Missing tooth 28	Indicates if tooth 28 is missing
Missing tooth 29	Indicates if tooth 29 is missing
Missing tooth 30	Indicates if tooth 30 is missing
Missing tooth 31	Indicates if tooth 31 is missing
Missing tooth 32	Indicates if tooth 32 is missing
Orthodontics treatment indicator	Indicates if the treatment was for orthodontics.

Date appliance placed	Date the orthodontic appliance is placed.
Months of orthodontic treatments remaining	Approximate amount of months of orthodontic treatment left.
Replacement of prosthesis indicator	Indicates if the treatment replaced a prosthesis.
Date of placement prior prosthesis	Date the prior prosthesis was placed.

3.2.14 Type dimension

This dimension defines the type of the claim. The type is based on the source claim form and some attributes. Because this type does not contain any attributes besides the name it is possible to collapse this type into the claim line to reduce complexity of the final schema.

The different types are:

Inpatient claims

Inpatient claims is when the patient stays at a hospital during treatment. Treatments that take less then 24 hours are considered an outpatient claim.

A subset of the UB04 forms are this type.

Long-term care claims

These are the claims that take care of people who have chronic illness or disability. Most long-term care is to assist people with their daily activities like getting dressed or going to the toilet. The claims in this section are not incident based but are happening regularly. Long-term care can take place in various types as listed by CMS on the Medicare website (Centers for Medicare & Medicaid Services, 2012) and shown in Table 2.

Long-term care type	Help with activities of daily living	Help with additional services	Help with care needs	Range of costs
Community-Based Services	Yes	Yes	No	Low to medium
Home Health Care	Yes	Yes	Yes	Low to high
In-Law Apartments	Yes	Yes	Yes	Low to high
Housing for Aging and Disabled Individuals	Yes	Yes	No	Low to high
Board and Care Homes	Yes	Yes	Yes	Low to high
Assisted Living	Yes	Yes	Yes	Medium to high
Continuing Care Retirement Communities	Yes	Yes	Yes	High
Nursing Homes	Yes	Yes	Yes	High

Table 2 Types of long-term care

The claims in this type will be distinguished by filtering the CMS-1500 forms on EPSDT/Family Planning (24h) attribute equaling 10 (Long Term Care Resident) or Place of Service (24b) equaling Nursing Facility, Hospice or similar.

Pharmacy claims

This are all the claims that are pharmacy related. All drug claim forms are claims of this type.

Professional/other claims

These are all outpatient claims. This means that the patient didn't had to stay over at a location for more than 24 hours. This is made up from all J400 claims and part of the CMS1500 claims that are not longterm.

Attributes

Type Primary Key	The primary key of this row.
Name	The name of the type.

4 Fraud Detection using the Data Warehouse

Major and Riedinger (2002) have developed an electronic fraud detection workflow as seen in Figure 11. The workflow describes the process of employees doing fraud detection using a fraud detection system. This workflow is used as a basis to create the workflow to use the data warehouse to find fraud. In the following sections each step will be explained in more detail.

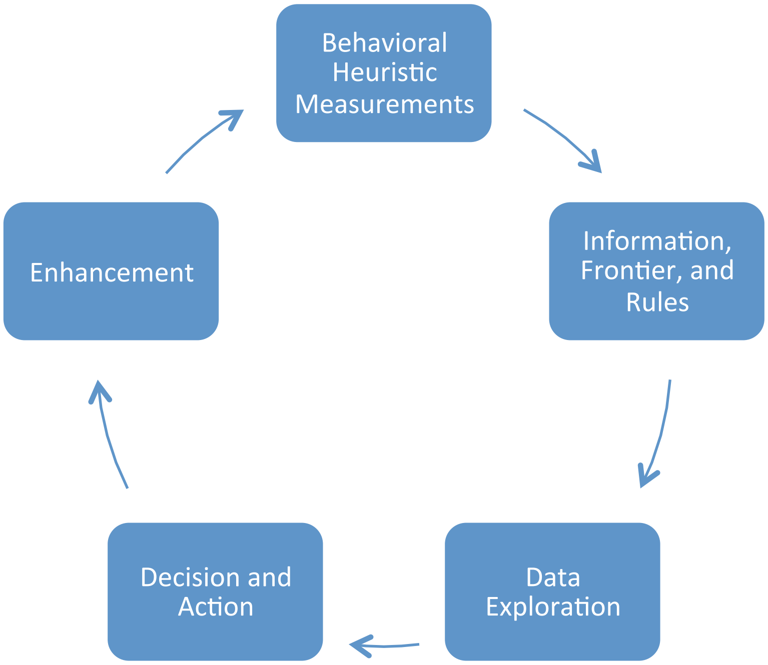

Behavioral Heuristic Measurements
Information, Frontier, and Rules
Data Exploration
Decision and Action
Enhancement

Figure 11 Electronic Fraud Detection workflow (Major & Riedinger, 2002)

4.1 Behavioral Heuristic Measurements / Data Preprocessing

Major and Riedinger use the term behavioral heuristic measurements to enrich the data set with certain attributes to aid in the detection of fraud.

It is not enough for the raw claim data to be just enriched. It also needs to be preprocessed and formatted to fit in the designed data warehouse. The biggest challenge in building fraud detection systems is to automate this process so fraud detection can include recent data. This step can take up to 80% of the total time in fraud detection according to Li, et al. (2008).

Li, et al. (2008) have developed a flow chart of the key steps in the data preprocessing for fraud detection within the health care as shown in Figure 12. This will be used to get the data into the system.

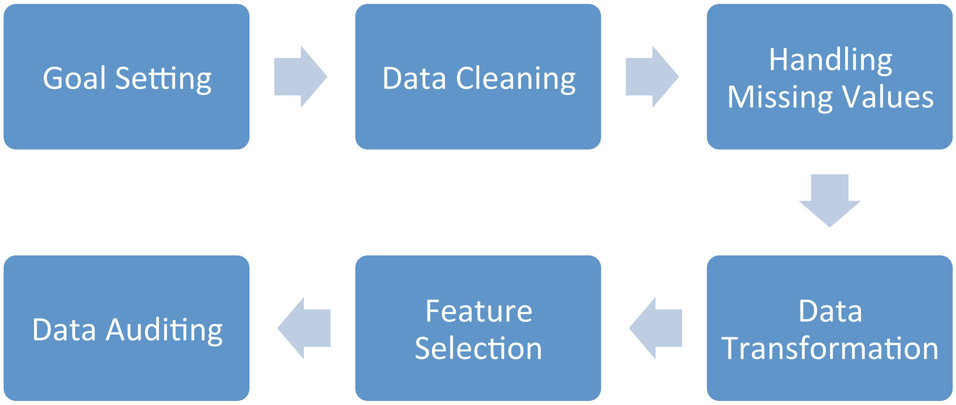


Figure 12 Data preprocessing steps (Li, Huang, Jin, & Shi, 2008)

4.1.1 Goal setting

The first step decides on which fraud detection will be focused. In this case it will be the fraud levels as set forth by Sparrow.

4.1.2 Data cleaning

To prepare the data for entering the system the first thing that has to happen is that it gets cleaned. There are a lot of different data sources that will be combined to fill the data in the system. One of the issues that resulted from analyses of data sets by Sokol, et al. (2001) was that different digit formats were used to represent the same physician's unique identification number. Cleaning the data would, among other things, make sure that all data sources use the same identification to refer to patients, providers, diagnoses, services and policies.

4.1.3 Handling missing values

Missing values are a common occurrence in health care data according to Li, et al. (2008): some data elements are not collected due to omission, irrelevance, excess risk, or inapplicability in a specific clinical context. A complete data set is required for most statistical methods for fraud detection so methods are required to fill in the gaps. Regression analysis could be used for this.

4.1.4 Data transformation

This is the actual step were the raw data is transformed to match the data models in the designed data warehouse. This will for example convert claims and their mutations into single claims and create aggregated rows per patient, provider and service.

This is also the step were data enrichment is done. For example extra information will be added on providers, drugs, services and taxonomies and linked to the claims.

Examples of enrichment that is done is the calculation of the distance between the provider and the patient. Another enrichment is classifying the claims into different types. This will be done based on a combination of source form and available attributes.

4.1.5 Feature selection

In this step the features are selected that should be available on the aggregated data. The platform already prescribes a few features that will be used for fraud detection but more can always be added as deemed necessary.

4.1.6 Data auditing

This step is to get familiar with the data. This can be done by doing basic statistical analysis, visualization and reporting. This step will be skipped in this approach as this is done in the next step of finding fraud. This will be explained in the next section.

4.2 Information, Frontier, and Rules

This step will let the system do statistical analyses on the data to find and flag anomalies. The data will be approached based on Sparrow's seven levels of health care fraud control. As stated before its first level, the claim-level, will be skipped.

Per level the involved entities from the data model will be listed and how one would approach them to find fraud. Examples will be given of fraud that can be found on each level. For each type a base query will be given that will give the information for that level. Some algorithms can be executed solely by executing an SQL query. Most queries will need some extra processing in an analytics tool.

Output of lower levels can be also be used as input for higher levels. One might for example flag suspicious patients on level 2 or 3 and use it to group patients when searching for fraud at level 6.

When writing algorithms to find fraud it is important for the indicators to be as accurate as possible. If you write an algorithm that generates a lot of false positives it will make the task to decide which cases to pursue more difficult. The other side of the medal is that if you are too strict you might have false negatives and leave out a lot of fraud. Ortega et al. (2006) suggests using a training set of data with known fraud to tweak the indicators on the algorithms.

4.2.1 Level 2: Patient/Provider relationship

The second level focuses on the relationship between one patient and one provider.

Approach: Look at the relationship between 1 provider and 1 patient and the related claims.

Base query:

```
SELECT
        *
FROM
        tbl_claim_line
INNER JOIN
        tbl_patient ON tbl_claim_line.patient_fk = tbl_patient.patient_pk
INNER JOIN
        tbl_provider ON tbl_claim_line.provider_fk = tbl_provider.provider_pk
```

```
WHERE
        tbl_patient.ssn = X AND tbl_provider.npi = Y
```

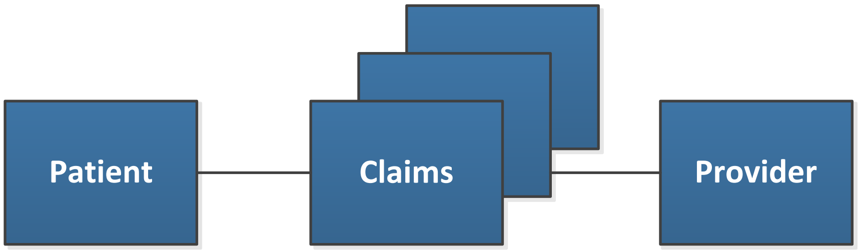


Figure 13 Involved entities at Sparrows level 2

Examples of checks that can be done:

- Duplicate claim/has the claim been claimed on this patient before (Sparrow, 2000, pp. p163-164).
- How often and in what order are certain services provided to the patient and does it fit within the reasonable norms for the providers specialty and for the patient's diagnosis (Sparrow, 2000, p. p233):
  - Categorize all claims, for example hospital claims.
  - Check the frequency and the order the services have been performed:
    - Was the x-ray performed before the splint or partial cast has been applied?
    - Have there been three splints applied within a short period?

4.2.2 Level 3: Patient and Provider levels

The third level shows all claims of either the patient or the provider. One could look for how a patient is treated by providers or if a provider only provides the most expensive treatments to its patients.

A. Patient level
Approach: For every patient, look at all of his/her claims and related providers.

Base query:

```
SELECT
        *
FROM
        tbl_claim_line
INNER JOIN
        tbl_patient ON tbl_claim_line.patient_fk = tbl_patient.patient_pk
INNER JOIN
        tbl_provider ON tbl_claim_line.provider_fk = tbl_provider.provider_pk
WHERE
        tbl_patient.ssn = $input_SSN
```

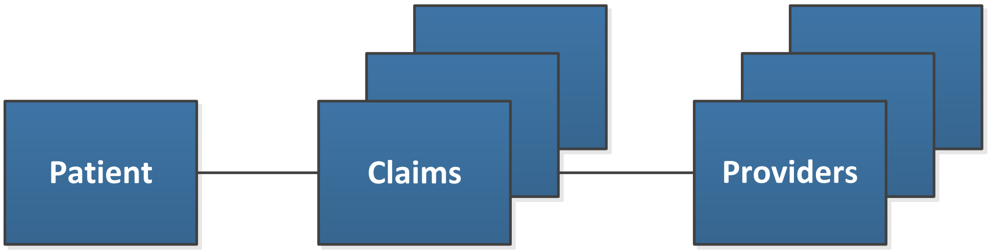


Figure 14 Involved entities at Sparrows level 3A

Examples of checks that can be done:

- Compare the amount of providers a patient visited to the average.
- Does the distance between the locations of the providers and the patients make sense:
    - Get a batch of claims over a period of time
    - Calculate the distance between the locations of sequential claims and see if frequent or unreasonable big jumps are made:
        - An example could be for a resident of San Diego to receive a claim based in San Diego on Tuesday, a claim based in San Francisco on Wednesday and another claim based in San Diego on Thursday again. That a person would visit San Francisco for 1 day from San Diego is very unlikely. This doesn't indicate that it is fraud per se but it is a suspicious act that could lead to a review of the case.
- How often is a certain service provided to the patient and does it fit within the reasonable norms for the providers' specialty and for the patient's diagnosis (Sparrow, 2000, p. p233).

<u>B. Provider level</u>
Approach: For every provider, look at all of its claims and related patients.
Base query:

```
SELECT
        *
FROM
        tbl_claim_line
INNER JOIN
        tbl_patient ON tbl_claim_line.patient_fk = tbl_patient.patient_pk
INNER JOIN
        tbl_provider ON tbl_claim_line.provider_fk = tbl_provider.provider_pk
WHERE
        tbl_provider.npi = $input_NPI
```

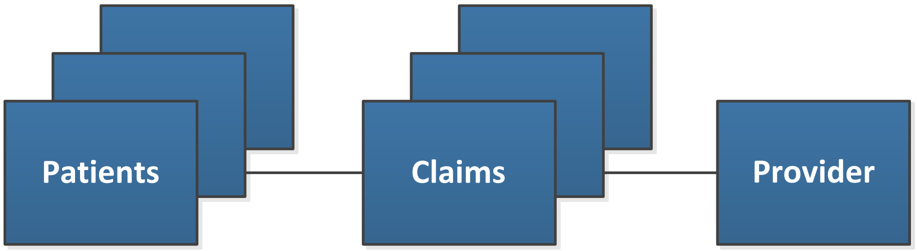


Figure 15 Involved entities at Sparrows level 3B

Examples of checks that can be done:

- If there are claims that are duplicates or very similar among different patients.
- How far the patients and provider are apart. This check could involve the results of the the similar check at level 3A to see which providers have a high amount of suspicious patients.
- If the provider has done more expensive or rare treatments than its competition:
    - Compare the distribution of the performed services to an average distribution of that industry. For the average distribution one could take into account the region and its demographic properties. For example the country side should not be compared to

4.2.3 Level 4: Patient Group and Provider Group levels

A. Patient group/Provider

Level 4 A expects providers to focus patients that have specific policies.

Approach: Per policy, look at the related claims, patients and providers.

Base query:

```
SELECT
        *
FROM
        tbl_claim_line
INNER JOIN
        tbl_patient ON tbl_claim_line.patient_fk = tbl_patient.patient_pk
INNER JOIN
        tbl_provider ON tbl_claim_line.provider_fk = tbl_provider.provider_pk
WHERE
        tbl_claim_line.health_plan_fk = $input_Health_Plan
```

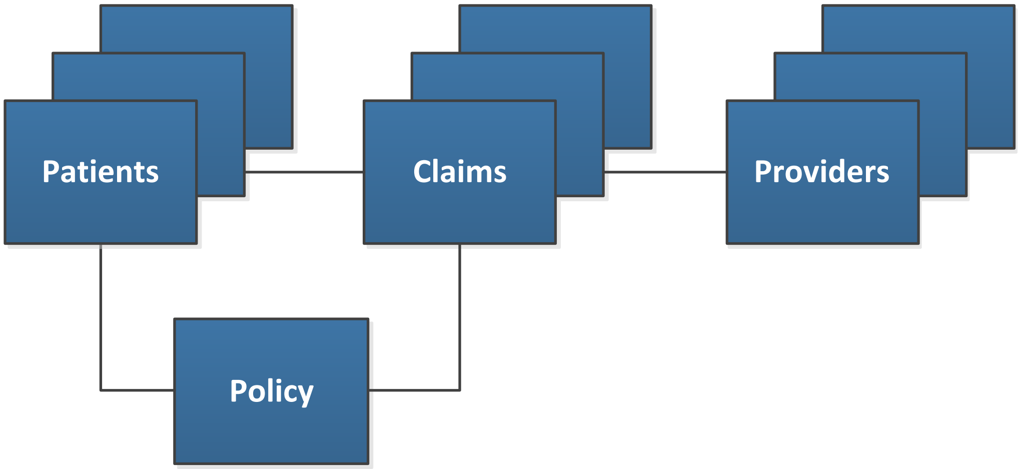


Figure 16 Involved entities at Sparrows level 4A

Examples of checks that can be done:

- Look for patients related to the same policy receiving more services from certain providers.
    - A provider might target people in Medicaid as they do not pay premiums and thus are not financially triggered to find fraud.
    - A provider might target patients with policies with extensive dental coverage.

B. Patient/Practice (clinic)

Level 4 B assumes that clinic or hospital administrators might be the ones committing fraud and spread the fraud over all of their providers. To detect fraud on this level all claims of the providers that work for 1 practice or hospital are grouped. Suspicious behavior will be looked for related to those groups.

Approach: Create provider groups based on providers within the same clinic or hospital. Per group look at the related claims and patients.

Base query:

```
SELECT
        *
FROM
        tbl_claim_line
INNER JOIN
        tbl_patient ON tbl_claim_line.patient_fk = tbl_patient.patient_pk
INNER JOIN
        tbl_provider ON tbl_claim_line.provider_fk = tbl_provider.provider_pk
WHERE
        tbl_provider.npi IN ($input_list_of_NPIs_related_to_1_hostpial)
```

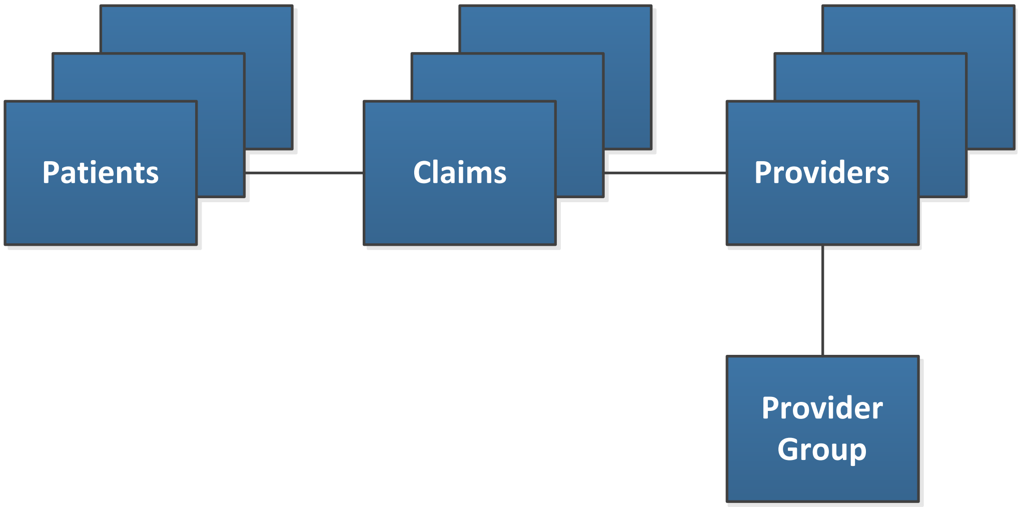


Figure 17 Involved entities at Sparrows level 4B

Examples of checks that can be done:

- Look if providers within the same hospital provide a certain service more than their competition.
    - One way to do this is to compare the distribution of blood tests done with other hospitals in similar regions.

4.2.4 Level 5: Policy/Practice relationship

Level 5 is a combination of level 4 A and B. All claims of the providers that work for 1 practice are grouped. Those groups are then used to see if suspicious behavior can be found when looking at single policies.

Approach: Create provider groups based on providers within the same clinic or hospital. Per group and per policy look at the related claims and patients.

Base query:

```
SELECT
        *
FROM
        tbl_claim_line
INNER JOIN
        tbl_patient ON tbl_claim_line.patient_fk = tbl_patient.patient_pk
INNER JOIN
        tbl_provider ON tbl_claim_line.provider_fk = tbl_provider.provider_pk
WHERE
        tbl_claim_line.health_plan_fk = $input_Health_Plan AND
        tbl_provider.npi IN ($input_list_of_NPIs_related_to_1_hospital)
```

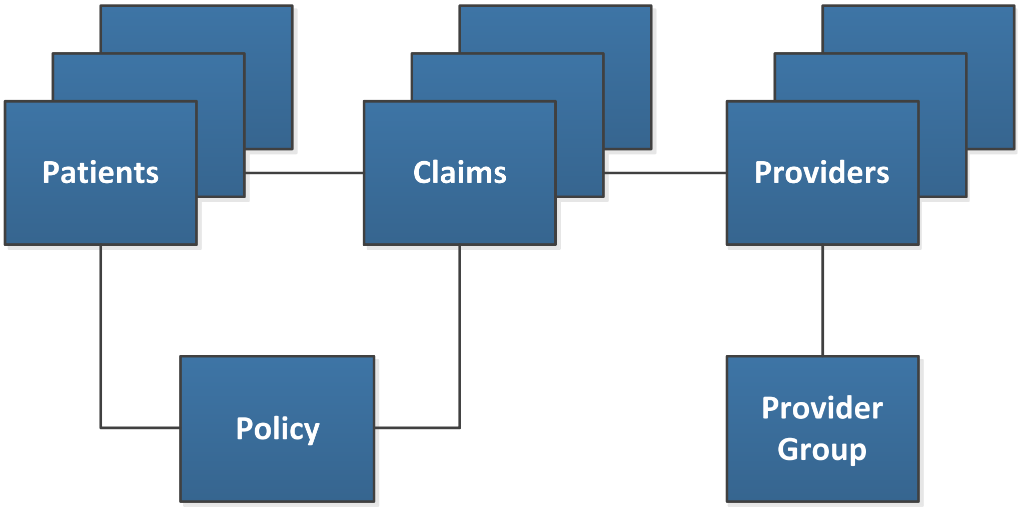


Figure 18 Involved entities at Sparrows level 5

Examples of checks that can be done:

- Look if providers within the same hospital provide a certain service more than their competition to patients with a certain policy.
    - Hospitals could target certain policies with extra coverage and provide a variety of unnecessary tests to patients.

4.2.5 Level 6: Defined groups of patients, Practices and Clinics

A. Defined group of patients

In level 6 A the focus is on patient groups. One would compare behavior of the groups to the average behavior on patients when acted upon by the same diagnosis and specialty.

Approach: Create patient groups based on factors that would allow fraud perpetrators to manipulate the accounts for the group. For example this could be a family covered by the same policy or residents of the same nursing home. Per group look at the related claims, patients and providers.

Base query:

```
SELECT
        *
FROM
        tbl_claim_line
INNER JOIN
        tbl_patient ON tbl_claim_line.patient_fk = tbl_patient.patient_pk
INNER JOIN
        tbl_provider ON tbl_claim_line.provider_fk = tbl_provider.provider_pk
WHERE
        tbl_patient.ssn IN ($input_list_of_SSNs_in_the_patient_group)
```

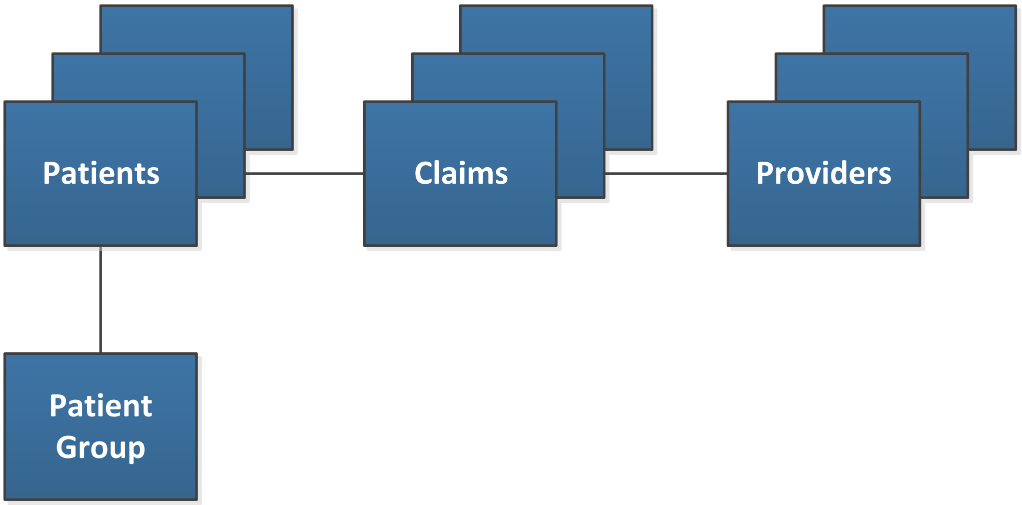


Figure 19 Involved entities at Sparrows level 6A

Examples of fraud that can be found with this level:

- People in the same nursing house receiving twice as many diapers than the average patient with the same condition and age.
    - Example of fraud happening at this level: One group optometry practice sent salespeople to the director of nursing or social worker at different nursing facilities to offer routine eye examinations for all the patients, at no cost. Such examinations are not covered by Medicare; the nursing staff provided access to the patients' records, enabling the optometry practice to use them as the basis for billings to Medicare. Medicare was billed for all kinds of services that were never provided. (Sparrow, 2000, p. 24)
- Patients who all share the same postbox as their address might have been a victim to a fraud perpetrator not only having access to his insurance record but also control it so EOMB's wont arrive. Real world example: Investigators at major Medicare contractors described how providers were altering beneficiaries' addresses on their claim forms and switching them to post office boxes under their own control, knowing that the claims-processing system would automatically update the beneficiaries' address before dispatching the EOMB. (Sparrow, 2000, pp. 133-134)
    - To find this fraud look for addresses that are shared by a an above average amount of patients.

B. Practices or Clinics

Level 6 B focuses on the connections between providers within provider groups and see if suspicious behavior is found. This approach is very similar to level 4 B where groups are formed based on practices or hospitals except that now other characteristics for grouping are used.

Approach: Create provider groups based on characteristics that relate them together. This could be for example location or referrals to each other.

Base query:

```
SELECT
        *
FROM
        tbl_claim_line
INNER JOIN
        tbl_patient ON tbl_claim_line.patient_fk = tbl_patient.patient_pk
INNER JOIN
        tbl_provider ON tbl_claim_line.provider_fk = tbl_provider.provider_pk
WHERE
        tbl_provider.npi IN ($input_list_of_NPIs_in_the_provider_group)
```

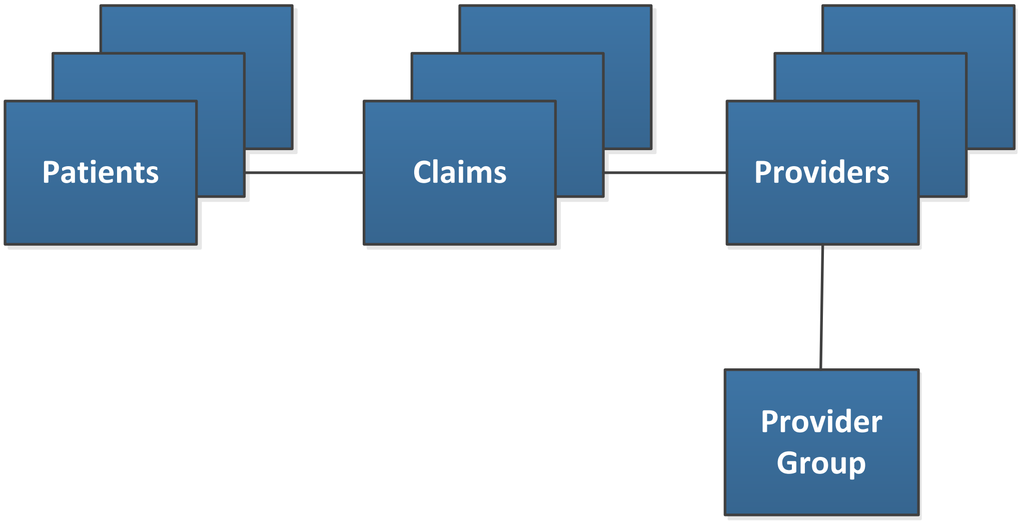


Figure 20 Involved entities at Sparrows level 6B

An example of this level by Sparrow (2000, p. p234) concerns “Medicaid Mills”, where several providers set up within a provider group and continually refer patients to-and-fro among themselves for needless tests and services. To find this kind of fraud one would:

- Select providers that have a high amount of referrals to each other.
- See if they are certain kind of services that are performed more regularly compared to the average.
- This will result in a list of suspicious providers. Second opinions will be required to see if the original diagnosis was biased.

4.2.6 Level 7: Multiparty, criminal conspiracies

Level 7 concerns all fraud that is part of criminal networks which involve many different beneficiaries and/or providers.

According to Sparrow (2000, p. p234) the “art of detection at this level involves watching for broad patterns of coincidence or connection between hundreds or thousands of otherwise innocuous transactions”.

Approach: A good approach for this level is not available.

4.3 Data Exploration

The fraud investigator will look into the anomalies produced in the previous step. He will look up extra information that he or she deems necessary.

4.4 Decision and Action

The fraud investigator has collected information on anomalies within the system. It will input the information into a decision support system to decide which providers will have to be investigated further.

4.5 Enhancement

By using the system fraud investigators will see what shortcomings there are in the system. It might decide that some aggregations are necessary to find certain types of fraud or data it had to collect from the system manually in the data exploration step are necessary to be auto generated in the future.

6 Validation

To further validate the platform this chapter will show how it can be used to detect the most prevalent fraud schemes, detect recent cases and walk through the platform with an expert.

6.1 Most prevalent health care fraud schemes

The platform will be tested against the most prevalent fraud schemes within the health care industry (as presented in chapter 2). Is it possible for the presented approach to reveal suspicious behavior hinting that these fraud schemes are going on?

While reading the results one might notice that a lot of these fraud detection approaches are using Sparrows level 2 and level 3. This does not mean that the other levels of Sparrow are useless. They are just not applicable to find the most prevalent fraud schemes because they look at the data from a different perspective.

6.1.1 Billing for services not rendered

There are multiple approaches to try to detect this type of fraud. The easiest way to detect this type of fraud would be to ask the patients. This happens by insurance companies by sending patients Explanation of Medical Benefits (EOMB). This is an automatically generated overview of the bills that have been paid for services that the patient received. Patients not always read the statements and if they do, have a tough time understanding the medical terms. In Medicaid there is also an extra problem: because the patient is not paying any premium there is no incentive to check the EOMB for fraudulent behavior. According to Sparrow (2000, pp. 133-134) there have been cases reported were DME suppliers in Florida were giving patients $5 for each unopened EOMB envelop or they altered the address of the patient so sent EOMB would not arrive.

It is difficult to find this kind of fraud. The best bet to find this fraud is if the fraud perpetrator is greedy or careless and bills too many fake treatments. When the amounts of fake bills increase there are two ways that they will reveal themselves.

First way would be through an analysis of the claim history of the patient. By analyzing all the claims on a patient it is possible to look for the order and frequency of treatments. If the fake treatments are in conflict with real treatments this might show up as an anomaly. This kind of analysis can be done using the platform using Sparrows level 3a combined with an external source that provides knowledge on sensible frequency and order of treatments.

The second approach would be to analyze the provider and all of its claims using Sparrows level 3b. An unusual growth of patients served, an unusual increase of services per patient or an unusual distribution of services rendered might indicate that something suspicious is going on.

6.1.2 Upcoding of services and items

Again, approaching patients to notice this would be a good indicator. But since patients lack medical knowledge it is very difficult to spot this kind of fraud.

To detect this type of fraud one would have to look at level 3 B, the provider level analysis. Although exact fraud of this kind cannot be revealed by an analysis one could flag providers as suspicious when they offer on average more expensive services than other providers. To do this one would compare the distributions of the services or items sold.

6.1.3 Duplicate claims

There is a lot of literature on finding duplicates. Researches like Naumann and Herschel (2010) show that it is a difficult problem to tackle – but not impossible. It involves the searching for duplicates that are not exact duplicates, so called fuzzy matching, to get the job done. The problem is that many duplicate claims are legit and are the result of doctors treating patients with the same problem. Because of this there will be a lot of false positives as result of this algorithm.

To detect this fraud scheme one would create an algorithm to operate on the data at Sparrows level 3b, the provider level. At this level duplicate or very similar claims by the same provider can be found using fuzzy matching. A distribution of services rendered could be used to see if the provider has an unusual high amount of claims of a certain type that also contain a high amount of duplicate or highly similar claims.

To retrieve the data used for the comparison of the distribution of services rendered two queries can be executed. The first one is to get the service distribution from the providers with the same primary taxonomy:

```
SELECT
        tbl_service.name,
        COUNT (*) as count_claims,
        SUM(tbl_claim_line.covered_charges) as sum_covered_charges,
        SUM(tbl_claim_line.non_covered_charges) as sum_non_covered_charges,
        SUM(tbl_claim_line.total_charges) as sum_total_charges,
        AVG(tbl_claim_line.units_of_service) as avg_units_of_service,
        AVG(tbl_claim_line.covered_price_per_unit) as avg_covered_price_per_unit,
        AVG(tbl_claim_line.price_per_unit) as avg_price_per_unit,
        AVG(tbl_claim_line.treatment_duration) as avg_treatment_duration,
        COUNT(DISTINCT tbl_patient.ssn) as count_patients,
        ACG(tbl_claim_line.distance_between_provider_and_patient_in_km) as avg_distance
FROM
        tbl_claim_line
INNER JOIN
        tbl_date ON tbl_claim_line.service_rendered_date_fk = tbl_date.date_fk
INNER JOIN
        tbl_service ON tbl_claim_line.service_fk = tbl_service.service_pk
INNER JOIN
        tbl_provider ON tbl_claim_line.provider_fk = tbl_provider.provider_pk
WHERE
        tbl_date.quarter_year = “2010-5”  AND
        tbl_provider.primary_taxonomy = “Chiropractor”
GROUP BY
        tbl_service.name
ORDER BY
        tbl_service.name
```

The second query will be able to get the same information but then only on one provider. The NPI of the provider (unique identifier) will be used as $input_NPI in the following query:

```
SELECT
        tbl_service.name,
        COUNT (*) as count_claims,
        SUM(tbl_claim_line.covered_charges) as sum_covered_charges,
        SUM(tbl_claim_line.non_covered_charges) as sum_non_covered_charges,
        SUM(tbl_claim_line.total_charges) as sum_total_charges,
        AVG(tbl_claim_line.units_of_service) as avg_units_of_service,
        AVG(tbl_claim_line.covered_price_per_unit) as avg_covered_price_per_unit,
        AVG(tbl_claim_line.price_per_unit) as avg_price_per_unit,
        AVG(tbl_claim_line.treatment_duration) as avg_treatment_duration,
        COUNT(DISTINCT tbl_patient.ssn) as count_patients,
        AVG(tbl_claim_line.distance_between_provider_and_patient_in_km) as avg_distance
FROM
        tbl_claim_line
INNER JOIN
        tbl_date ON tbl_claim_line.service_rendered_date_fk = tbl_date.date_fk
INNER JOIN
        tbl_service ON tbl_claim_line.service_fk = tbl_service.service_pk
INNER JOIN
        tbl_provider ON tbl_claim_line.provider_fk = tbl_provider.provider_pk
WHERE
        tbl_date.quarter_year = “2010-5”  AND
        tbl_provider.npi = $input_NPI
GROUP BY
        tbl_service.name
ORDER BY
        tbl_service.name ASC
```

This data can then be fed into analytical tools to see if the provider is far off from the average. The output from the second query can also be used to look at the distribution of the claims of one provider each quarter.

6.1.4 Unbundling

To detect unbundling all the claims that a provider has billed on a patient have to be analyzed. This data is available in Sparrows level 2. When this data is combined with an external source that contains information on which services could be bundled an algorithm can be written to detect this fraudulent behavior. It would look for claims within a certain time range that could have been bundled.

As it is always possible for unbundling to happen by accident this algorithm could be extended to operate on Sparrows level 3b, the provider level. It would then be able to check which providers have an unusual high amount of cases that could have been bundled compared to other providers.

6.1.5 Medically unnecessary or excessive services

Detection of this kind of behavior is similar to detection of services not rendered. It will only be able to detect if the provider is too careless or greedy and does a lot of excessive services. This will give the provider a boost in patients and treatments done per patient. If he does not manage to make the growth or the distribution of treatments done look natural an algorithm might be able to pick it up.

Algorithms would look at Sparrows level 3b to the growth of the provider over a period of time compared to other providers in the region. External information on the growth of the region could be included to make the algorithm more precise.

Queries to retrieve this information is the same like the queries proposed to detect duplicate claims.

6.1.6 Kickbacks

Kickbacks are very difficult to find. Referrals are a normal business for doctors if a patient has a problem that they are unable to solve. To look for kickbacks providers that referred to each other are to be analyzed to see if suspicious behavior is going on.

To do this an algorithm can be written in two steps.

First step is to look at Sparrows level 3b, the provider level. The amount of patients, the reason that a provider is referring and the reason that he is referred to are to be analyzed. When a normal distribution of this analysis would differ from the average the provider is marked as suspicious.

To get a list of all references a provider has done the following query can be used to get a breakdown to which providers was referenced for what reason:

```
SELECT
        tbl_provider.npi,
        tbl_claim_line.diagnosis,
        COUNT(*) as sum_referrals
FROM
        tbl_claim_line
INNER JOIN
        tbl_date ON tbl_claim_line.service_rendered_date_fk = tbl_date.date_pk
INNER JOIN
        tbl_provider ON tbl_claim_line.provider_fk = tbl_provider.provider_pk
INNER JOIN
        tbl_provider tbl_ref_provider ON tbl_claim_line.referring_or_prescribing_provider_fk =
        tbl_ref_provider.provider_pk
INNER JOIN
        tbl_diagnosis ON tbl_claim_line.diagnosis_fk = tbl_diagnosis.diagnosis_pk
WHERE
        tbl_ref_provider.npi = $input_NPI AND
        tbl_date.calendar_year_quarter = $input_YEAR_QUARTER
GROUP BY
        tbl_provider.npi,
        tbl_diagnosis.name
```

The next piece of is the amount of referrals to a specific type of provider. This information is retrieved with the following query:

```
SELECT
        tbl_provider.primary_taxonomy,
        tbl_claim_line.diagnosis,
        COUNT(*) as count_referrals,
FROM
        tbl_claim_line
INNER JOIN
        tbl_date ON tbl_claim_line.service_rendered_date_fk = tbl_date.date_pk
INNER JOIN
        tbl_provider ON tbl_claim_line.provider_fk = tbl_provider.provider_pk
INNER JOIN
        tbl_provider tbl_ref_provider ON tbl_claim_line.referring_or_prescribing_provider_fk =
        tbl_ref_provider.provider_pk
INNER JOIN
        tbl_diagnosis ON tbl_claim_line.diagnosis_fk = tbl_diagnosis.diagnosis_pk
WHERE
        tbl_ref_provider.npi = $input_NPI AND
        tbl_date.calendar_year_quarter = $input_YEAR_QUARTER
GROUP BY
        tbl_provider.primary_taxonomy,
        tbl_diagnosis.name
```

These distributions should then be compared to the average to see if something suspicious is going on, for example a high amount of referrals on a specific diagnosis.

The second step would operate on Sparrows level 6b that looks on data related to provider groups. Provider groups based on the fact that providers referred a high number of times to each other are to be created. Then look at groups that have a high amount of suspicious providers from the previous step.

To get persons that have more than a set minimum amount of referrals to each other the following query can be used:

```
SELECT
        tbl_provider.npi,
        COUNT(*) as sum_referrals
FROM
        tbl_claim_line
INNER JOIN
        tbl_date ON tbl_claim_line.service_rendered_date_fk = tbl_date.date_pk
INNER JOIN
        tbl_provider ON tbl_claim_line.provider_fk = tbl_provider.provider_pk
INNER JOIN
        tbl_provider tbl_ref_provider ON tbl_claim_line.referring_or_prescribing_provider_fk =
        tbl_ref_provider.provider_pk
WHERE
        tbl_ref_provider.npi = $input_NPI AND
        tbl_date.calendar_year_quarter = $input_YEAR_QUARTER
GROUP BY
        tbl_provider.npi
HAVING
        COUNT(*) > $input_MINIMUM_REFERRALS
```

To see if certain groups referred a lot to each other the referral data has to be feeded into a network analysis tool.

6.2 Recent cases of health care fraud

In this section the fraud detection system will be tested to see if it could have been used to detect recent health care cases. A variety of cases are selected for which the structure of the fraud scheme is analyzed and if the fraud cases could have been detected.

6.2.1 HIV injection and infusion Medicare fraud scheme

This case was published by the United States department of Justice on April 14, 2011 (United States Department of Justice, 2011). Relevant excerpts are printed below:

Miami-area physician Rene De Los Rios was convicted of five felony counts today by a federal jury for his role in a $23 million dollar HIV injection and infusion Medicare fraud scheme, the Departments of Justice and Health and Human Services (HHS) announced.

In 2003, Metro Med began operating as an HIV infusion clinic that purportedly provided injection and infusion therapies to HIV positive Medicare beneficiaries. In fact, the injection and infusion therapies were medically unnecessary and not provided. Metro Med paid cash kickback payments to patients at the Metro Med clinic in exchange for those patients allowing Metro Med to use their Medicare numbers to bill the Medicare program.

Evidence at trial established that as part of the scheme, Metro Med hired the De Los Rios to order unnecessary tests, sign medical analysis and diagnosis forms, and authorize treatments to make it *appear that legitimate medical services, including injection and infusion therapies, were being provided to patients who were Medicare beneficiaries. The defendant also signed patient charts, often without seeing the patient, indicating that injection and infusion treatments were medically necessary, when, in fact, he knew they were not. Evidence at trial also established that the defendant diagnosed almost all of the patients at Metro Med with the same rare blood disorders, which the patients did not in fact have, in order to ensure maximum reimbursement from Medicare. Moreover, the evidence at trial showed that the defendant prescribed expensive medications, including Winrho, Procrit and Neupogen, to patients for the sole purpose of receiving reimbursement from the Medicare program. The evidence showed that Metro Med paid the defendant $3,000 per week for his involvement in the HIV infusion scheme.*

The following interactions are happening:

- Patients are paid to be falsely diagnosed with a rare blood disorder and be billed for treatments that they do not receive.
- The physician is paid to provide the false diagnose. He receives a kickback for this from Metro Med. Based on the diagnoses the patient is in need of injection and infusion therapies.
- Metro Med bills Medicare for the medical unnecessary therapies that are not provided.

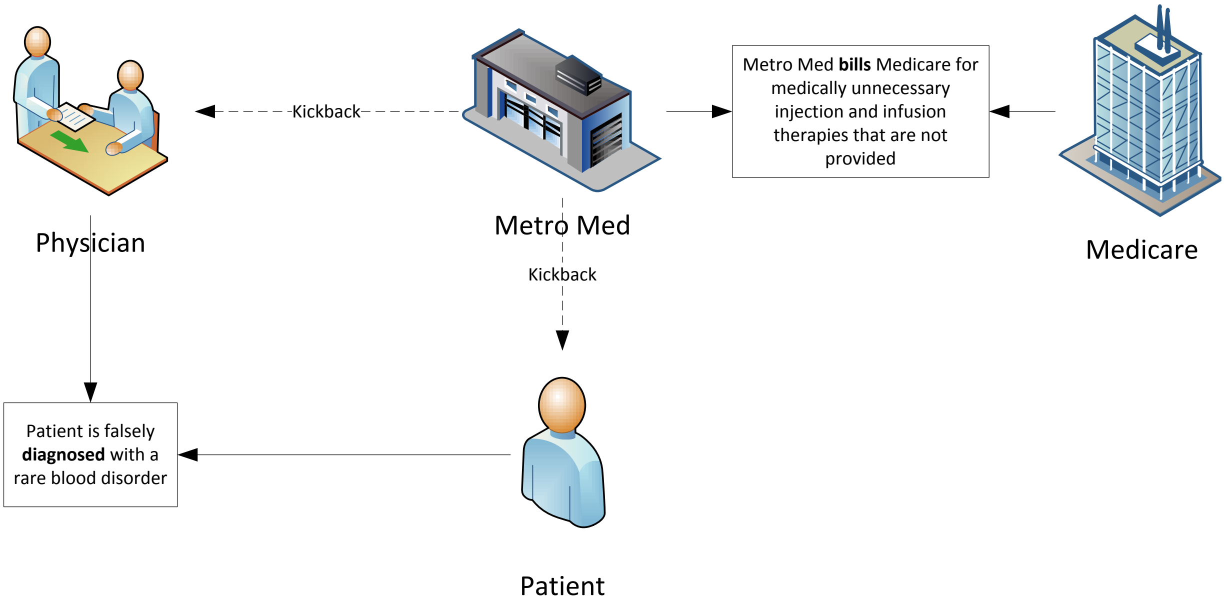

Kickback
Metro Med bills Medicare for medically unnecessary injection and infusion therapies that are not provided
Physician
Metro Med
Medicare
Kickback
Patient is falsely diagnosed with a rare blood disorder
Patient

Figure 23 Visualization of the interactions in the first case

These interactions indicate that a Pill Mill scheme is being used. The behavior that would give this scheme away is the fact that the physician chose to diagnose almost all patients with the same rare blood disorders.

To detect this and similar schemes an algorithm would have to look at the physician and all the patients that it served. One can use Sparrows level 3b for this. Use an algorithm that creates a distribution of the diagnoses and see how it relates to the average distribution. In such a comparison the high amount of diagnoses indicating a rare blood disorder will be flagged as suspicious.

6.2.2 False claims to Medicare for durable medical equipment

This case was published by the United States department of Justice on May 2, 2012 (United States Attorney's Office, Central District of California, 2012). Relevant excerpts are printed below:

According to court documents filed in the Central District of California, two Orange County doctors and two of their co-schemers were charged for allegedly submitting nearly $5.7 million in false claims to Medicare for durable medical equipment (DME). Specifically, the defendants billed Medicare for enteral nutrition, a liquid nutritional supplement. Medicare will only pay for enteral nutrition if a patient has a feeding tube. According to the indictment, Dr. Augustus Ohemeng, 62, of Buena Park, and Dr. George Tarryk, 72, of Seal Beach, wrote fraudulent prescriptions for enteral nutrition for patients who did not have feeding tubes. Co-defendant George Samuel Laing, 41, of Sylmar, who managed the clinic where Tarryk and Ohemeng practiced, allegedly received kickbacks in exchange for referring the prescriptions to Ivy Medical Supply, owned by co-defendant Emmanuel Chidueme, 59, of Mira Loma. Ivy then fraudulently billed Medicare for the enteral nutrition, even though it was not medically necessary and was not delivered to patients in the quantities billed to Medicare. Ohemeng, Tarryk, Laing and Chidueme were arrested this morning and are scheduled to make their initial appearances before a U.S. Magistrate Judge this afternoon.

The following interactions happening:

- Physicians wrote fraudulent prescriptions for enteral nutrition for patients who did not have feeding tubes.
- Prescriptions are referred to Ivy Medical Supply in exchange for kickbacks.
- Prescriptions are fulfilled by Ivy Medical Supply who did not (fully) deliver the quantities billed to Medicare to the patients.

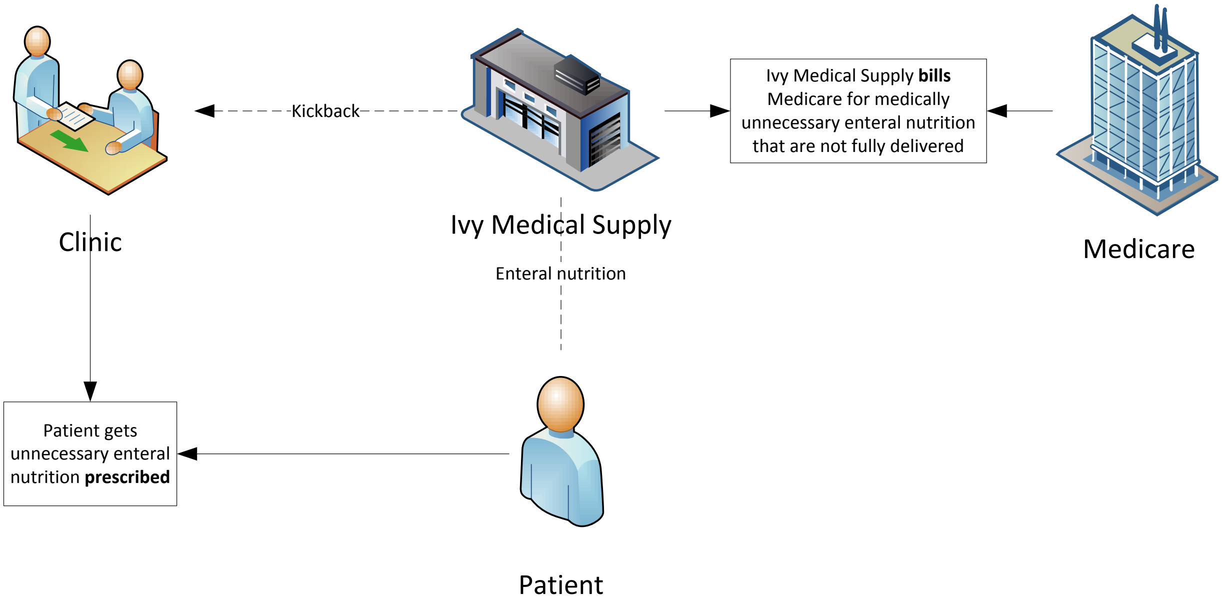

Kickback
Ivy Medical Supply **bills** Medicare for medically unnecessary enteral nutrition that are not fully delivered
Clinic
Ivy Medical Supply
Medicare
Enteral nutrition
Patient gets unnecessary enteral nutrition **prescribed**
Patient

Figure 24 Visualization of the interactions of the second case

As the enteral nutrition is only to be prescribed to patients that have a feeding tube this fraud could have been detected by analyzing the patients records at Sparrows level 3a. By analyzing the order of claims one could find out that the patient was never provided a feeding tube.

It could be the case that the patient has received a feeding tube while enrolled in a different health insurance and thus could be flagged as a false positive. To get a better indicator it is in this case better to look at the percentage of suspicious patients (those receiving enteral nutrition without getting a feeding tube) that a provider is treating.

The following query will give us a percentage of users that a provider is treating that did not receive a feeding tube. To do this first look up the services that deliver a feeding tube to a patient and which provide enteral nutrition. Looking at the HCPCS-standard the following codes are found:

Code	Description
B4034	ENTERAL FEEDING SUPPLY KIT; SYRINGE, PER DAY
B4035	ENTERAL FEEDING SUPPLY KIT; PUMP FED, PER DAY
B4036	ENTERAL FEEDING SUPPLY KIT; GRAVITY FED, PER DAY
B4081	NASOGASTRIC TUBING WITH STYLET
B4082	NASOGASTRIC TUBING WITHOUT STYLET
B4083	STOMACH TUBE - LEVINE TYPE
B9000	ENTERAL NUTRITION INFUSION PUMP - WITHOUT ALARM
B9002	ENTERAL NUTRITION INFUSION PUMP - WITH ALARM

Table 3 HCPCS-codes for tube providing services

Code	Description
B4150	ENTERAL FORMULAE; CATEGORY I; SEMI-SYNTHETIC INTACT PROTEIN/PROTEIN ISOLATES, ADMINISTERED THROUGH AN ENTERAL FEEDING TUBE, 100 CALORIES = 1 UNIT
B4151	ENTERAL FORMULAE; CATEGORY I; NATURAL INTACT PROTEIN/PROTEIN ISOLATES, ADMINISTERED THROUGH AN ENTERAL FEEDING TUBE, 100 CALORIES = 1 UNIT
B4152	ENTERAL FORMULAE; CATEGORY II; INTACT PROTEIN/PROTEIN ISOLATES (CALORICALLY DENSE), ADMINISTERED THROUGH AN ENTERAL FEEDING TUBE, 100 CALORIES = 1 UNIT
B4153	ENTERAL FORMULAE; CATEGORY III; HYDROLIZED PROTEIN/AMINO ACIDS, ADMINISTERED THROUGH AN ENTERAL FEEDING TUBE, 100 CALORIES = 1 UNIT
B4154	ENTERAL FORMULAE; CATEGORY IV; DEFINED FORMULA FOR SPECIAL METABOLIC NEED, ADMINISTERED THROUGH AN ENTERAL FEEDING TUBE, 100 CALORIES = 1 UNIT
B4155	ENTERAL FORMULAE; CATEGORY V; MODULAR COMPONENTS, ADMINISTERED THROUGH AN ENTERAL FEEDING TUBE, 100 CALORIES = 1 UNIT
B4156	ENTERAL FORMULAE; CATEGORY VI; STANDARDIZED NUTRIENTS, ADMINISTERED THROUGH AN ENTERAL FEEDING TUBE, 100 CALORIES = 1 UNIT

Table 4 HCPCS-codes for enteral nutrition

For each provider count the patients that received enteral nutrition without getting a feeding tube before. These patients can be selected with the following query $QUERY-PATIENT-WITH-NUTRITION-PRIOR-TUBE:

```
SELECT
        COUNT(*)
FROM
        tbl_claim_line
INNER JOIN
        tbl_date ON tbl_claim_line.service_rendered_date_fk = tbl_date.date_pk
INNER JOIN
        tbl_service ON tbl_claim_line.service_fk = tbl_service.service_pk
INNER JOIN
        tbl_provider ON tbl_claim_line.provider_fk = tbl_provider.provider_pk
INNER JOIN
        tbl_patient parent_tbl_patient ON
                tbl_claim_line.patient_fk = parent_tbl_patient.patient_pk
WHERE
        tbl_service.type = “HCPCS” AND
        tbl.service.code IN
                ("B4150","B4151","B4152","B4153","B4154","B4155","B4156") AND
        tbl_provider.npi = $input_NPI
GROUP BY
        parent_tbl_patient.ssn
HAVING
        MIN(tbl_date) <
        NVL(($QUERY-PATIENT-FOR-FIRST-FEEDING-TUBE), TO_DATE('9999', 'yyyy'))
```

This query uses a subquery to determine the first date that a feeding tube was received by the patient. If the patient never received a feeding tube the subquery will return NULL. NVL will then replace the value by TO_DATE(‘9999’,’yyyy’) which represents the future date January 1, 9999 and thus will return the patient aswell.

The query to get the first nutrition tube delivery date for a patient $QUERY-PATIENT-FOR-FIRST-FEEDING-TUBE:

```
SELECT
        MIN(tbl_date.date)
FROM
        tbl_claim_line
INNER JOIN
        tbl_date ON tbl_claim_line.service_rendered_date_fk = tbl_date.date_pk
INNER JOIN
        tbl_service ON tbl_claim_line.service_fk = tbl_service.service_pk
INNER JOIN
        tbl_patient ON tbl_claim_line.patient_fk = tbl_patient.patient_pk
WHERE
        tbl_service.type = “HCPCS” AND
```

```
    tbl.service.code IN
    ("B4034","B4035","B4036","B4081","B4082","B4083","B9000","B9002") AND
    tbl_patient.ssn = $input_SSN
```

In this case set $input_SSN to *parent_tbl_patient.ssn* so it only checks the services related to the specific patient being processed.

To be able to calculate how much percentage of the patients of a provider are suspicious it is needed to query the total amount of patients that get enteral nutrition from the provider. This can be done with the following query $QUERY-TOTAL-PATIENTS-GETTING-NUTRITION:

```
SELECT
    COUNT(DISTINCT tbl_patient.ssn)
FROM
    tbl_claim_line
INNER JOIN
    tbl_service ON tbl_claim_line.service_fk = tbl_service.service_pk
INNER JOIN
    tbl_provider ON tbl_claim_line.provider_fk = tbl_provider.provider_pk
INNER JOIN
    tbl_patient tbl_patient ON tbl_claim_line.patient_fk = tbl_patient.patient_pk
WHERE
    tbl_service.type = “HCPCS” AND
    tbl.service.code IN
        ("B4150","B4151","B4152","B4153","B4154","B4155","B4156") AND
    tbl_provider.npi = $input_NPI
```

These two queries combined can be used to query all providers that serve a high percentage of their patients enteral nutrition before receiving a feeding tube:

```
SELECT
    tbl_mainprovider.*,
    ($QUERY-TOTAL-PATIENTS-GETTING-NUTRITION) as total_patients,
    ($QUERY-PATIENT-FOR-FIRST-FEEDING-TUBE) as suspicious_patients,
    suspicious_patients / total_patients as suspicious_score
FROM
    tbl_provider as tbl_mainprovider
ORDER BY
    suspicious_patients DESC
```

For both subqueries set $input_NPI to tbl_mainprovider.npi. This will get the provider information sorted by suspicious score, highest first.

6.2.3 Billing Medicare for unnecessary expensive, high-end power wheelchairs and orthotics

This case was published by the United States department of Justice on February 27, 2012 (United States Department of Justice, 2012). Relevant excerpts are printed below:

A former Los Angeles church pastor, who owned and operated several fraudulent durable medical equipment (DME) supply companies with her husband, was sentenced today to serve 36 months in prison for her role in a $14.2 million Medicare fraud scheme, the Department of Justice, FBI and Department of Health and Human Services (HHS) announced.

According to the trial evidence, Ikpoh, Iruke, Marroquin and their co-conspirators used fraudulent prescriptions and documents that Ikpoh and Iruke purchased from a number of illicit sources to bill Medicare for expensive, high-end power wheelchairs and orthotics that were medically unnecessary or never provided. Each power wheelchairs cost approximately $900 per wholesale, but were billed to Medicare at a rate of approximately $6,000 per wheelchair. Witness testimony established that Ikpoh and Iruke hid the money they used to pay for these fraudulent prescriptions by writing checks to a company called "Direct Supply," a fictitious company that Iruke created in the name of an Arms of Grace church member. Iruke cashed the checks that he and Ikpoh wrote to Direct Supply and used the money to purchase the fraudulent prescriptions.

Witnesses who sold the fraudulent prescriptions and documents that Ikpoh, Iruke and their co-conspirators used to defraud Medicare testified that they and others paid cash kickbacks to street-level marketers to offer Medicare beneficiaries free power wheelchairs and other DME in exchange for the beneficiaries' Medicare card numbers and personal information. These witnesses testified that they and their associates used this information to create fraudulent prescriptions and medical documents, which they sold to Iruke and the operators of other fraudulent DME supply companies for $1,100 to $1,500 per prescription.

The following interactions are happening:

- Medicare beneficiaries get offered free power wheelchairs and other DME in exchange for Medicare card number and personal information.
- Medicare card numbers and personal information are used to create fraudulent prescriptions and medical documents.
- Wheelchairs were bought for approximately $900 and billed to Medicare at a rate of approximately $6.000 per wheelchair. The wheelchairs were in some cases never provided to the patients.

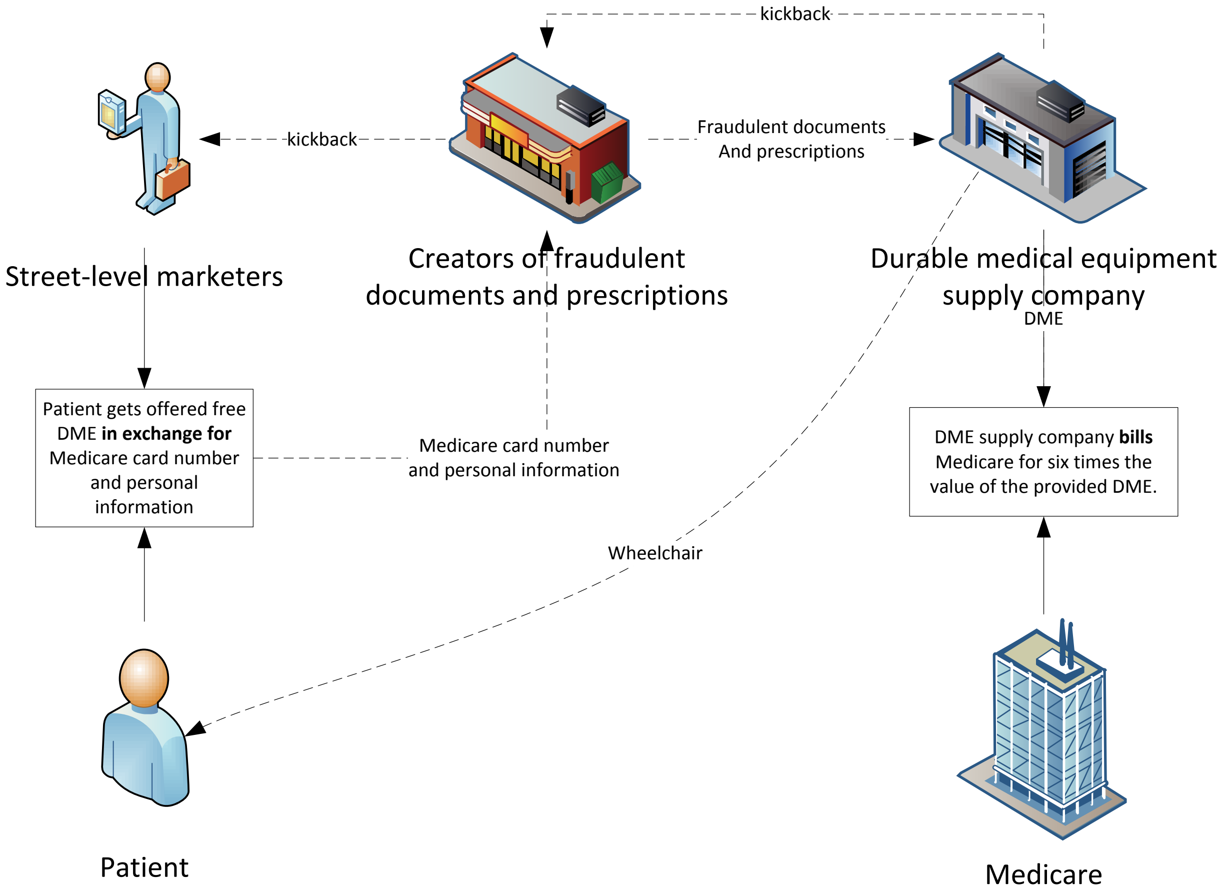

kickback
kickback
Fraudulent documents
And prescriptions
Street-level marketers
Creators of fraudulent
documents and prescriptions
Durable medical equipment
supply company
DME
Patient gets offered free DME **in exchange for** Medicare card number and personal information
Medicare card number
and personal information
DME supply company **bills** Medicare for six times the value of the provided DME.
Wheelchair
Patient
Medicare

Figure 25 Visualization of the interactions in the third case

This fraud scheme is very difficult to detect because there are no clear indicators that something is going on.

Looking at possible indicators the first one that comes to mind is the fact that they only sell expensive wheelchairs. But if this company was a reseller of a specific brand it might be that they only sell expensive wheelchairs.

Another indicator might be that the company is showing an unnatural growth: their business has been rising a lot in a very short period of time. Growth numbers can be derived by analyzing the number of claims billed and the number of patients treated over time.

A third indicator might be the amount of patients served with wheelchairs that did not had mobility-related problems before. This could be detected by checking all patients served by a certain provider and analyze their records. The definition of mobility-related problems is very vague and will therefore be difficult to implement.

It will be very difficult or close to impossible to detect this kind of fraud as there are no clear indicators that this fraud is happening. Next to that this fraud is very difficult to detect it is also very easy to enter the DME business as there is no required medical education necessary. Because of these reasons fraud with DME is happening the most according to the FBI (2009).

6.2.4 Conclusion of the cases

Three recent health care cases have been discussed. The first one can probably be detected easily because it uses a rare blood disorder as diagnoses to maximize reimbursements. The second one needed an algorithm that looks if services are provided to a patient in logical order. This requires medical knowledge on which services are related and can be therefore quite complex but is doable. The third case is almost impossible to detect as there are no clear indicators. This is a limitation that applies to all fraud detection mechanisms.

6.3 Interview with an expert

In October I sat down twice with fraud detection expert Christopher Bunnell. A transcript of the interviews can be found in the appendix.

History

Bunnell is currently employed by CNI Inc working on fraud detection within Medicaid and Medicare. He has been active within the fraud detection for twelve years. Before that he was working as a hospital administrator for eight years.

Workflow

The way Bunnell would detect fraud is as follows:

1. Get an extract from the COBALT-system that the state uses for the Medicaid-system and insert it into Oracle. Add extra reference data like procedure codes, DRG-codes.
2. Brainstorm about ideas for algorithms with the use of a decision support system
3. Test the data
4. Present the data to the client if fraud was found

The biggest bottleneck in the process of recovering the money is the bureaucracy. The employees at the states differ in commitment. You have the ones that are very committed but also people that prefer if no fraud is found because they will get the blame for the claim being paid in the first place.

In his experience the contracts that get paid per service instead of a percentage of the fraud that was found would perform better.

This workflow does not show any major differences with the electronic fraud detection workflow by Major and Riedinger (2002). The difference between the systems is the extra step by Major and Riedinger to propose enhancements to the data-set based on previous fraud findings. It is very likely that this was also done during Bunnell's work but that he didn't mention it.

Creating a standardized data format

Bunnell doesn't believe in the attempts of the federal government to homogenize the data and create one format that will work for all states. When looking at the data you will see a lot of elements having no value because the information is not saved in the source system.

The reason for that is that every state is unique. What works for California – a big state, a desert state and huge population – doesn't work necessarily for Rhode Island. A lot of things do but not everything. Every state has its own laws, rules and regulations. Those are reflected by the systems to a small extent.

The data warehouse assumes that all the used dimensions and attributes are available in the source systems – this might not be the case. So when an insurance company plans to implement a data warehouse for fraud detection it should adopt the used dimensions to the available data in the source systems.

Evaluation of the proposed model

Companies have been doing this for a long time according to Bunnell and he states that anything that makes the analysis better is useful. Bunnell judged every dimension to see if it was useful for fraud detection. This is shown in Table 5.

Dimension	Useful for fraud detection
Diagnosis	Yes
Patient	Yes
Service	Yes
Provider	Yes
Drug	Yes
Location	Yes
Date	Yes
Outcome	Never seen that data being available
Health Plan	Not applicable in Medicaid

Table 5 Usefulness of the dimensions for fraud detection according to Bunnell

Bunnell doesn't know of any algorithms that would look at an episode of care (using the patient dimension). It is not a bad idea according to him but as far as he knows the federal government does not use any.

The outcome data that Bunnell didn't see is the patient discharge status that is recorded by the UB-04 forms.

Bunnell states that a smaller data warehouse is being used by states as a decision support system (DSS). The DSS is used to choose which areas are going to be focused on for finding fraud. They are not using the DSS to find the fraud.

7 Conclusion

This thesis tried to answer the research question:

"What is an efficient way of finding fraud within Medicaid?"

The information system research framework by Hevner, et al. (2004) was used as approach to answer this question:

- Describe the current knowledge base.
- Develop an artifact to solve this question.
- Evaluate the developed artifact.

The knowledge base was explored and domain knowledge was gathered on fraud within the health care and how Medicaid and health care works in the United States of America. It was found that fraud is a big problem for Medicaid as there is little incentive to report fraud by the patients, especially if they are involved in the scheme. Different fraud detection approaches are elaborated upon during the study.

The first part of the artifact that was created is a data warehouse that will help in the search to detect fraud. This was done according to the specifications by Kimball & Ross (2008). This is done based on the information that is inserted into the system: the claim forms used by the providers. This information will be enriched by extra reference information describing extra information on drugs, services and providers. By reviewing the available information it was decided that a line describing the service rendered will be the fact in the data warehouse. This means that it is the finest amount of detail that fraud finding employees want to be able to view and filter upon when searching for fraud. Different dimension were defined that can be used to filter and group the fact-table.

A workflow for using the data warehouse for electronic fraud detection was proposed as the second artifact. As basis the electronic fraud detection workflow of Major & Riedinger (2002) was used. It was expanded using the data cleansing methods of Li, et al. (2008) and the levels of fraud detection as specified by Sparrow (2000). This resulted in a structured approach to fetch different a wide spectrum of fraud.

For evaluation a prototype of the data warehouse was built to test the structure of the data warehouse and the workflow. This proved that the data warehouse could be used to find fraud and provided feedback on which extra attributes can be included in the data warehouse to support fraud detection.

An attempt was made to tackle most prevalent fraud schemes according to the FBI (2009). It was shown how some fraud can be detected and attempts can be made to find others. Some fraud schemes are too much disguised as innocent claims that they cannot be directly detected.

Through an interview with fraud detection expert Christopher Bunnell the data warehouse was further validated. Concluded was that the structure is well-formed but should be adapted to the structure of the source-system used by the state. The reason for this is that states have different laws related to

Medicaid and thus have slightly different implementations. Trying to map this to a standard that covers all states will receive in a lot states offering incomplete data according to Bunnell.

All in all, it can be concluded that the creation of a single data warehouse to do fraud detection within Medicaid across all states is a practical unachievable goal in the current regulatory setting. Instead the focus should be on a per state basis.

8 Discussion & Future Research

8.1 Research

There is not much published about fraud detection within the health care for various reasons. Finding information about Medicaid is not as straightforward as one might expect from a government-funded project. This made it difficult to bring all pieces together and still leaves some holes in the knowledge base about the healthcare fraud detection space. Extra research could be done to get an overview of the algorithms that states currently use and which ones are successful and why.

This thesis describes a data warehouse that can be used as a platform to do fraud detection and shows some examples of fraud detection that could be done. Future research can be done on exploring the different kind of algorithms that can be done on the data. Research could explore how fraud can be found by analyzing the behavior of a patient or provider over time, look for specific behavioral patterns or do geographic analyses. More research can also be done on the provider and patient groups used by Sparrows seven levels of fraud detection. Which groups of patients are more probable to be targeted by fraud perpetrators, how can they be identified within the data and is there any extra external data needed to do this?

According to the NCHAA (year unknown) there were over 4 billion health care claims processed in the year 2007. These are too many claims to let humans alone do all the inspection. Future research could dig into finding a way to do unsupervised fraud detection. This would mimic the ability of humans to detect unknown patterns into a computer algorithm. Without this only known fraud can be detected because the system is trained to look for them.

Some fraud is detected by looking for anomalies by comparing behavior to the average behavior. This means that when all providers are doing fraud it would not show up as fraudulent behavior. The chance that there are more dishonest than honest providers is slim so it is probably not an issue.

8.2 Data warehouse

The described data warehouse in this thesis is based on the information that is inserted into the source system at the different states. Since states have different regulations and laws different data is collected by each state. When you want to create a data warehouse specific for a single state the design should not be based upon the data that is used as input for the source system but based on the data layout of the source system. The source system might discard certain information and enrich other data with other available information.

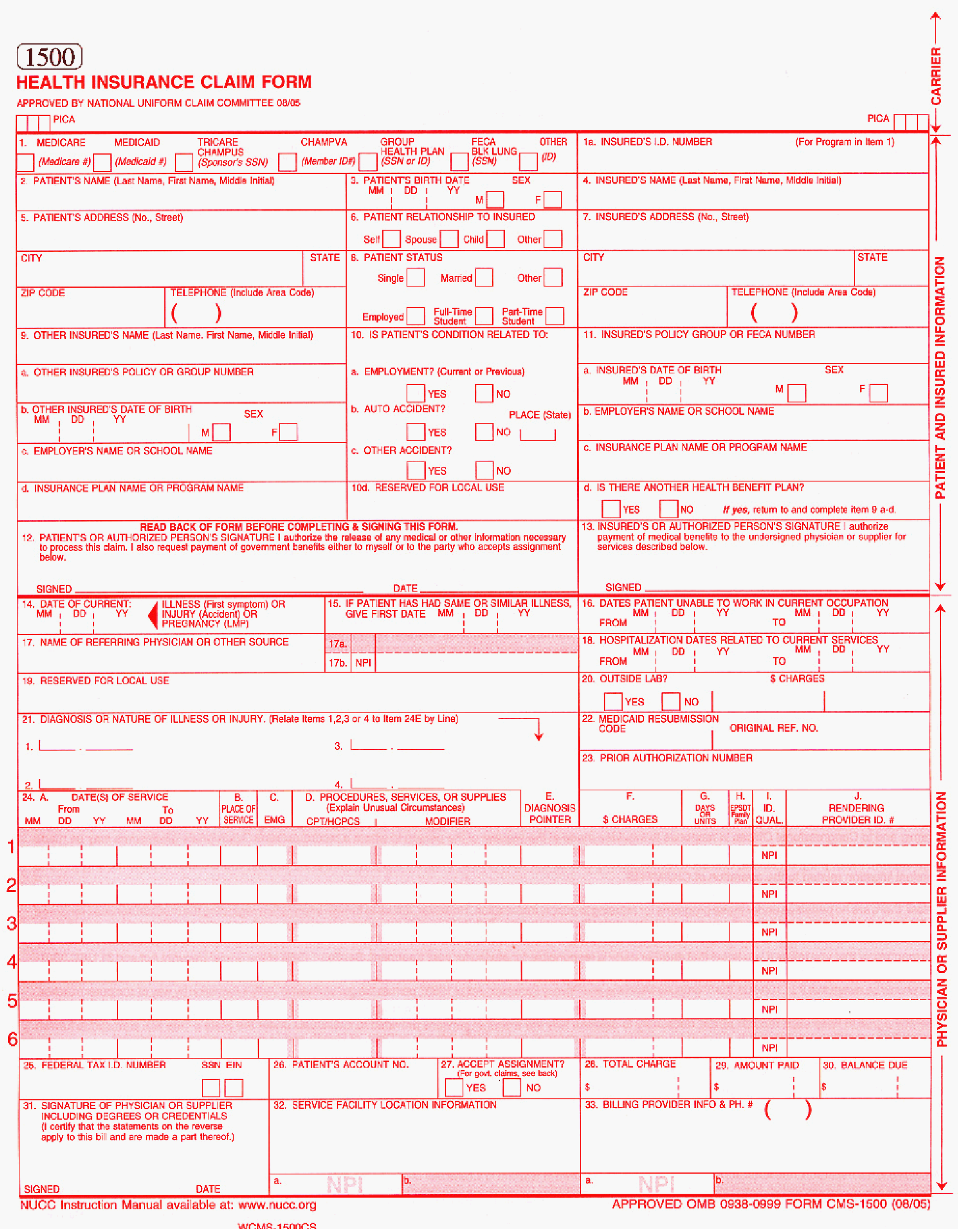
Getting the data from the source systems combined with extra reference data into the structure of the data warehouse is a big and time-consuming challenge. This is partly caused by the large number of fields that is contained within the data warehouse. This means that the creation of the process to convert the data from the different sources into the data warehouse can be a lengthy process. This process could be speeded up if fewer fields are required. Further research could be done on which fields are useful to find fraud and which fields are never used and thus should not be included in the data warehouse. Research can also focus on which external data can be integrated into the data warehouse to support the fraud detection efforts.

8.3 Fraud detection in the big picture

The perfect fraud-finding tool is not the whole solution. It is merely a part of the larger process of prosecuting fraud perpetrators and recovering the money. Even if the tool would be able to solve all the technical and organizational challenges that are there to find fraud within the data, it will not be able to solve the bureaucratic environment that is handling the found fraud.

As shown in the thesis there is a big problem in handling all the fraud claims at health insurances. When data shows that fraud might be going on the health insurance will have to get into contact with the fraud perpetrator to retrieve extra documentation. This is a lengthy and human resource intensive process. Bringing fraud to court is also a human resource and capital intensive activity that might not result in getting the money back. If money is recovered the money is split between the federal and the state government as to who paid which part of the original bill. This happens while the fraud detection efforts are paid by either one of them. Also it is not always possible to get fraud perpetrators out of business after they are proven guilty. You cannot just close the only hospital in a region. Future research can look into how states are currently organized and how they work. The goal of that research could be to define steps how the described approach in this thesis can be integrated into their current business processes of handling healthcare fraud. Such research could also be made broader to involve not only the states but also the other different stakeholders (providers, federal government, patients).

Appendix A: HCFA-1500 (CMS1500)



1500

HEALTH INSURANCE CLAIM FORM

APPROVED BY NATIONAL UNIFORM CLAIM COMMITTEE 08/05

□□□ PICA

PICA □□□

CARRIER

PATIENT AND INSURED INFORMATION

1. MEDICARE □ (Medicare #) MEDICAID □ (Medicaid #) TRICARE CHAMPUS □ (Sponsor's SSN) CHAMPVA □ (Member ID#) GROUP HEALTH PLAN □ (SSN or ID) FECA BLK LUNG □ (SSN) OTHER □ (ID)

1a. INSURED'S I.D. NUMBER (For Program in Item 1)

2. PATIENT'S NAME (Last Name, First Name, Middle Initial)

3. PATIENT'S BIRTH DATE MM DD YY SEX M □ F □

4. INSURED'S NAME (Last Name, First Name, Middle Initial)

5. PATIENT'S ADDRESS (No., Street)

6. PATIENT RELATIONSHIP TO INSURED Self □ Spouse □ Child □ Other □

7. INSURED'S ADDRESS (No., Street)

CITY STATE

8. PATIENT STATUS Single □ Married □ Other □

CITY STATE

ZIP CODE TELEPHONE (Include Area Code) ( )

Employed □ Full-Time Student □ Part-Time Student □

ZIP CODE TELEPHONE (Include Area Code) ( )

9. OTHER INSURED'S NAME (Last Name, First Name, Middle Initial)

10. IS PATIENT'S CONDITION RELATED TO:

11. INSURED'S POLICY GROUP OR FECA NUMBER

a. OTHER INSURED'S POLICY OR GROUP NUMBER

a. EMPLOYMENT? (Current or Previous) □ YES □ NO

a. INSURED'S DATE OF BIRTH MM DD YY SEX M □ F □

b. OTHER INSURED'S DATE OF BIRTH MM DD YY SEX M □ F □

b. AUTO ACCIDENT? □ YES □ NO PLACE (State)

b. EMPLOYER'S NAME OR SCHOOL NAME

c. EMPLOYER'S NAME OR SCHOOL NAME

c. OTHER ACCIDENT? □ YES □ NO

c. INSURANCE PLAN NAME OR PROGRAM NAME

d. INSURANCE PLAN NAME OR PROGRAM NAME

10d. RESERVED FOR LOCAL USE

d. IS THERE ANOTHER HEALTH BENEFIT PLAN? □ YES □ NO *If yes*, return to and complete item 9 a-d.

READ BACK OF FORM BEFORE COMPLETING & SIGNING THIS FORM.

12. PATIENT'S OR AUTHORIZED PERSON'S SIGNATURE I authorize the release of any medical or other information necessary to process this claim. I also request payment of government benefits either to myself or to the party who accepts assignment below.

SIGNED ____________ DATE ____________

13. INSURED'S OR AUTHORIZED PERSON'S SIGNATURE I authorize payment of medical benefits to the undersigned physician or supplier for services described below.

SIGNED ____________

PHYSICIAN OR SUPPLIER INFORMATION

14. DATE OF CURRENT: MM DD YY ◀ ILLNESS (First symptom) OR INJURY (Accident) OR PREGNANCY (LMP)

15. IF PATIENT HAS HAD SAME OR SIMILAR ILLNESS, GIVE FIRST DATE MM DD YY

16. DATES PATIENT UNABLE TO WORK IN CURRENT OCCUPATION FROM MM DD YY TO MM DD YY

17. NAME OF REFERRING PHYSICIAN OR OTHER SOURCE

17a.

17b. NPI

18. HOSPITALIZATION DATES RELATED TO CURRENT SERVICES FROM MM DD YY TO MM DD YY

19. RESERVED FOR LOCAL USE

20. OUTSIDE LAB? □ YES □ NO $ CHARGES

21. DIAGNOSIS OR NATURE OF ILLNESS OR INJURY. (Relate Items 1,2,3 or 4 to Item 24E by Line)

1. ___ . ___

3. ___ . ___

2. ___ . ___

4. ___ . ___

22. MEDICAID RESUBMISSION CODE ORIGINAL REF. NO.

23. PRIOR AUTHORIZATION NUMBER

24. A. DATE(S) OF SERVICE						B.	C.	D. PROCEDURES, SERVICES, OR SUPPLIES (Explain Unusual Circumstances)		E.	F.	G.	H.	I.	J.
From			To												
MM	DD	YY	MM	DD	YY	PLACE OF SERVICE	EMG	CPT/HCPCS	MODIFIER	DIAGNOSIS POINTER	$ CHARGES	DAYS OR UNITS	EPSDT Family Plan	ID. QUAL.	RENDERING PROVIDER ID. #
1														NPI	
2														NPI	
3														NPI	
4														NPI	
5														NPI	
6														NPI	

25. FEDERAL TAX I.D. NUMBER SSN EIN □ □

26. PATIENT'S ACCOUNT NO.

27. ACCEPT ASSIGNMENT? (For govt. claims, see back) □ YES □ NO

28. TOTAL CHARGE $

29. AMOUNT PAID $

30. BALANCE DUE $

31. SIGNATURE OF PHYSICIAN OR SUPPLIER INCLUDING DEGREES OR CREDENTIALS (I certify that the statements on the reverse apply to this bill and are made a part thereof.)

SIGNED DATE

32. SERVICE FACILITY LOCATION INFORMATION

a. NPI b.

33. BILLING PROVIDER INFO & PH. # ( )

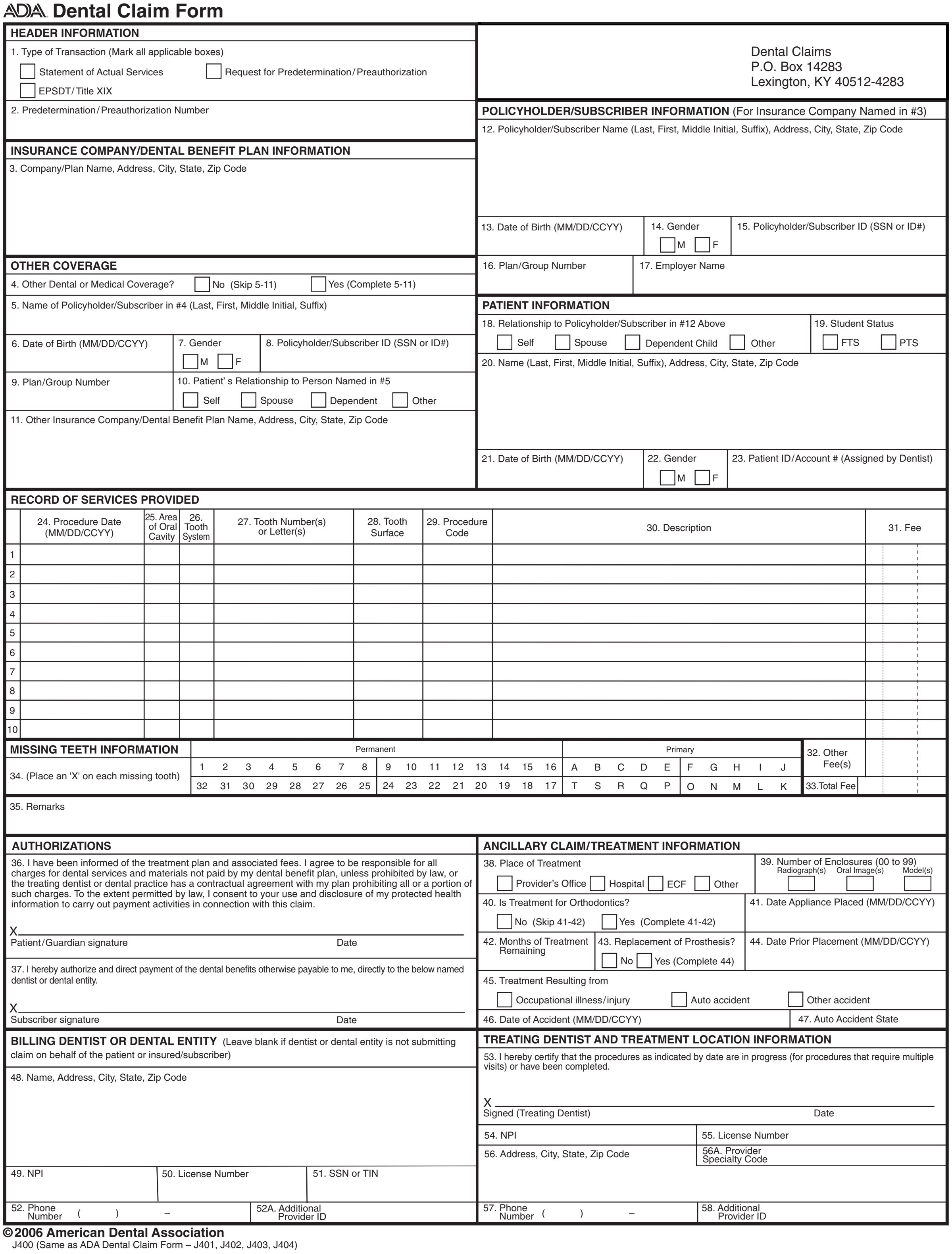
a. NPI b.

NUCC Instruction Manual available at: www.nucc.org

APPROVED OMB 0938-0999 FORM CMS-1500 (08/05)

WCMS-1500CS

Appendix B: J400 Dental Claim form



ADA Dental Claim Form

HEADER INFORMATION

1. Type of Transaction (Mark all applicable boxes)

□ Statement of Actual Services □ Request for Predetermination / Preauthorization

□ EPSDT / Title XIX

2. Predetermination / Preauthorization Number

Dental Claims
P.O. Box 14283
Lexington, KY 40512-4283

INSURANCE COMPANY/DENTAL BENEFIT PLAN INFORMATION

3. Company/Plan Name, Address, City, State, Zip Code

OTHER COVERAGE

4. Other Dental or Medical Coverage? □ No (Skip 5-11) □ Yes (Complete 5-11)

5. Name of Policyholder/Subscriber in #4 (Last, First, Middle Initial, Suffix)

6. Date of Birth (MM/DD/CCYY)

7. Gender □ M □ F

8. Policyholder/Subscriber ID (SSN or ID#)

9. Plan/Group Number

10. Patient' s Relationship to Person Named in #5

□ Self □ Spouse □ Dependent □ Other

11. Other Insurance Company/Dental Benefit Plan Name, Address, City, State, Zip Code

POLICYHOLDER/SUBSCRIBER INFORMATION (For Insurance Company Named in #3)

12. Policyholder/Subscriber Name (Last, First, Middle Initial, Suffix), Address, City, State, Zip Code

13. Date of Birth (MM/DD/CCYY)

14. Gender □ M □ F

15. Policyholder/Subscriber ID (SSN or ID#)

16. Plan/Group Number

17. Employer Name

PATIENT INFORMATION

18. Relationship to Policyholder/Subscriber in #12 Above

□ Self □ Spouse □ Dependent Child □ Other

19. Student Status

□ FTS □ PTS

20. Name (Last, First, Middle Initial, Suffix), Address, City, State, Zip Code

21. Date of Birth (MM/DD/CCYY)

22. Gender □ M □ F

23. Patient ID/Account # (Assigned by Dentist)

fold

fold

RECORD OF SERVICES PROVIDED

	24. Procedure Date (MM/DD/CCYY)	25. Area of Oral Cavity	26. Tooth System	27. Tooth Number(s) or Letter(s)	28. Tooth Surface	29. Procedure Code	30. Description	31. Fee
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								

MISSING TEETH INFORMATION

34. (Place an 'X' on each missing tooth)

Permanent																Primary									
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	A	B	C	D	E	F	G	H	I	J
32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	T	S	R	Q	P	O	N	M	L	K

32. Other Fee(s)

33.Total Fee

fold

fold

35. Remarks

AUTHORIZATIONS

36. I have been informed of the treatment plan and associated fees. I agree to be responsible for all charges for dental services and materials not paid by my dental benefit plan, unless prohibited by law, or the treating dentist or dental practice has a contractual agreement with my plan prohibiting all or a portion of such charges. To the extent permitted by law, I consent to your use and disclosure of my protected health information to carry out payment activities in connection with this claim.

X

Patient / Guardian signature Date

37. I hereby authorize and direct payment of the dental benefits otherwise payable to me, directly to the below named dentist or dental entity.

X

Subscriber signature Date

ANCILLARY CLAIM/TREATMENT INFORMATION

38. Place of Treatment

□ Provider's Office □ Hospital □ ECF □ Other

39. Number of Enclosures (00 to 99)

Radiograph(s) □ Oral Image(s) □ Model(s) □

40. Is Treatment for Orthodontics?

□ No (Skip 41-42) □ Yes (Complete 41-42)

41. Date Appliance Placed (MM/DD/CCYY)

42. Months of Treatment Remaining

43. Replacement of Prosthesis?

□ No □ Yes (Complete 44)

44. Date Prior Placement (MM/DD/CCYY)

45. Treatment Resulting from

□ Occupational illness / injury □ Auto accident □ Other accident

46. Date of Accident (MM/DD/CCYY)

47. Auto Accident State

BILLING DENTIST OR DENTAL ENTITY (Leave blank if dentist or dental entity is not submitting claim on behalf of the patient or insured/subscriber)

48. Name, Address, City, State, Zip Code

49. NPI

50. License Number

51. SSN or TIN

52. Phone Number ( ) –

52A. Additional Provider ID

TREATING DENTIST AND TREATMENT LOCATION INFORMATION

53. I hereby certify that the procedures as indicated by date are in progress (for procedures that require multiple visits) or have been completed.

X

Signed (Treating Dentist) Date

54. NPI

55. License Number

56. Address, City, State, Zip Code

56A. Provider Specialty Code

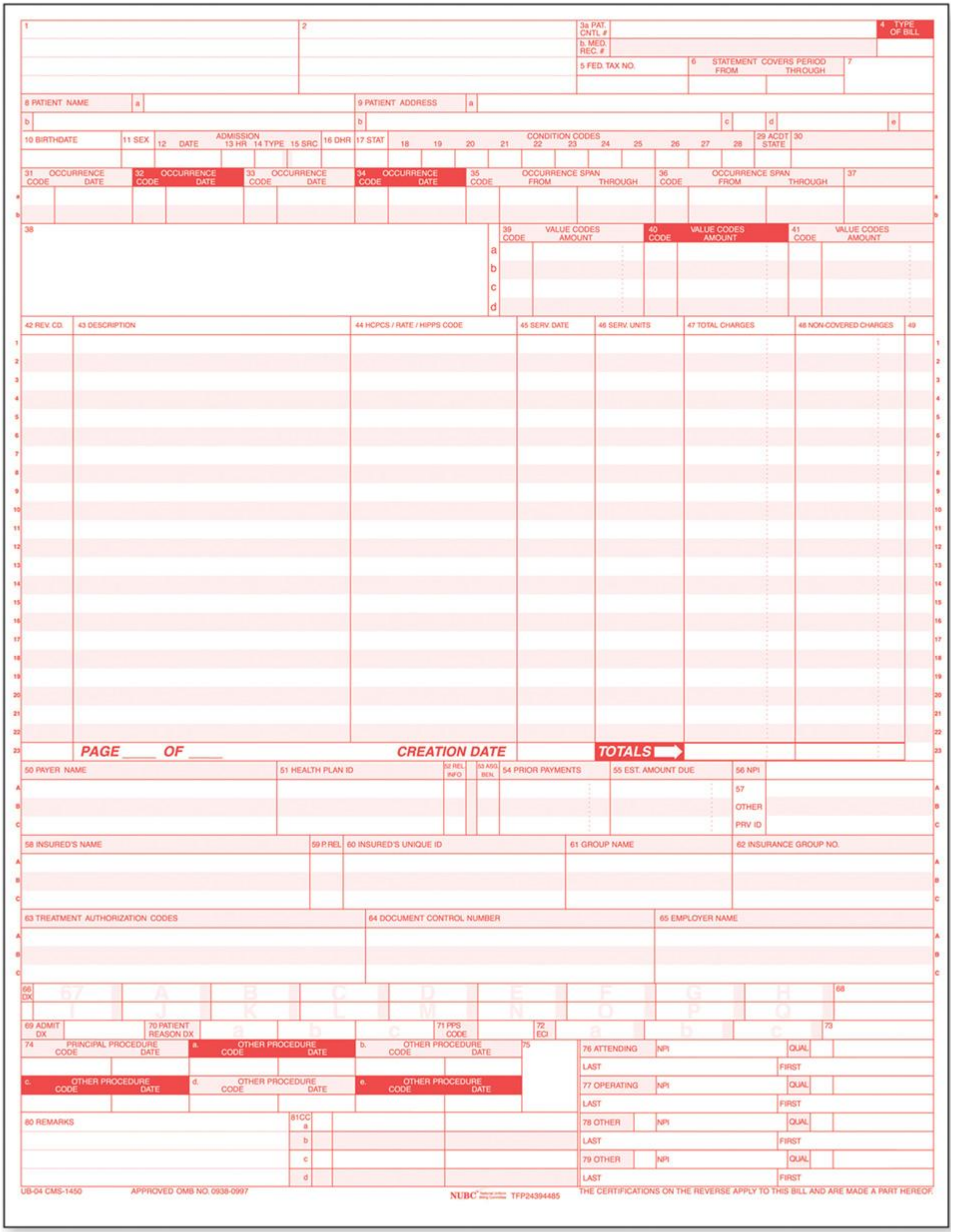
57. Phone Number ( ) –

58. Additional Provider ID

©2006 American Dental Association

J400 (Same as ADA Dental Claim Form – J401, J402, J403, J404)

Appendix C: UB-04



1 | 2 | 3a PAT. CNTL # | 4 TYPE OF BILL
b. MED. REC. #
5 FED. TAX NO. | 6 STATEMENT COVERS PERIOD FROM THROUGH | 7

8 PATIENT NAME a | 9 PATIENT ADDRESS a
b | b c d e

10 BIRTHDATE | 11 SEX | ADMISSION 12 DATE 13 HR 14 TYPE 15 SRC | 16 DHR | 17 STAT | CONDITION CODES 18 19 20 21 22 23 24 25 26 27 28 | 29 ACDT STATE | 30

31 OCCURRENCE CODE DATE | 32 OCCURRENCE CODE DATE | 33 OCCURRENCE CODE DATE | 34 OCCURRENCE CODE DATE | 35 OCCURRENCE SPAN CODE FROM THROUGH | 36 OCCURRENCE SPAN CODE FROM THROUGH | 37
a
b

38 | 39 VALUE CODES CODE AMOUNT | 40 VALUE CODES CODE AMOUNT | 41 VALUE CODES CODE AMOUNT
a
b
c
d

42 REV. CD. | 43 DESCRIPTION | 44 HCPCS / RATE / HIPPS CODE | 45 SERV. DATE | 46 SERV. UNITS | 47 TOTAL CHARGES | 48 NON-COVERED CHARGES | 49

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

PAGE ____ OF ____ | CREATION DATE | TOTALS

50 PAYER NAME | 51 HEALTH PLAN ID | 52 REL INFO | 53 ASG. BEN. | 54 PRIOR PAYMENTS | 55 EST. AMOUNT DUE | 56 NPI
A | 57
B | OTHER
C | PRV ID

58 INSURED'S NAME | 59 P.REL | 60 INSURED'S UNIQUE ID | 61 GROUP NAME | 62 INSURANCE GROUP NO.
A
B
C

63 TREATMENT AUTHORIZATION CODES | 64 DOCUMENT CONTROL NUMBER | 65 EMPLOYER NAME
A
B
C

66 DX | 67 A B C D E F G H | 68
I J K L M N O P Q

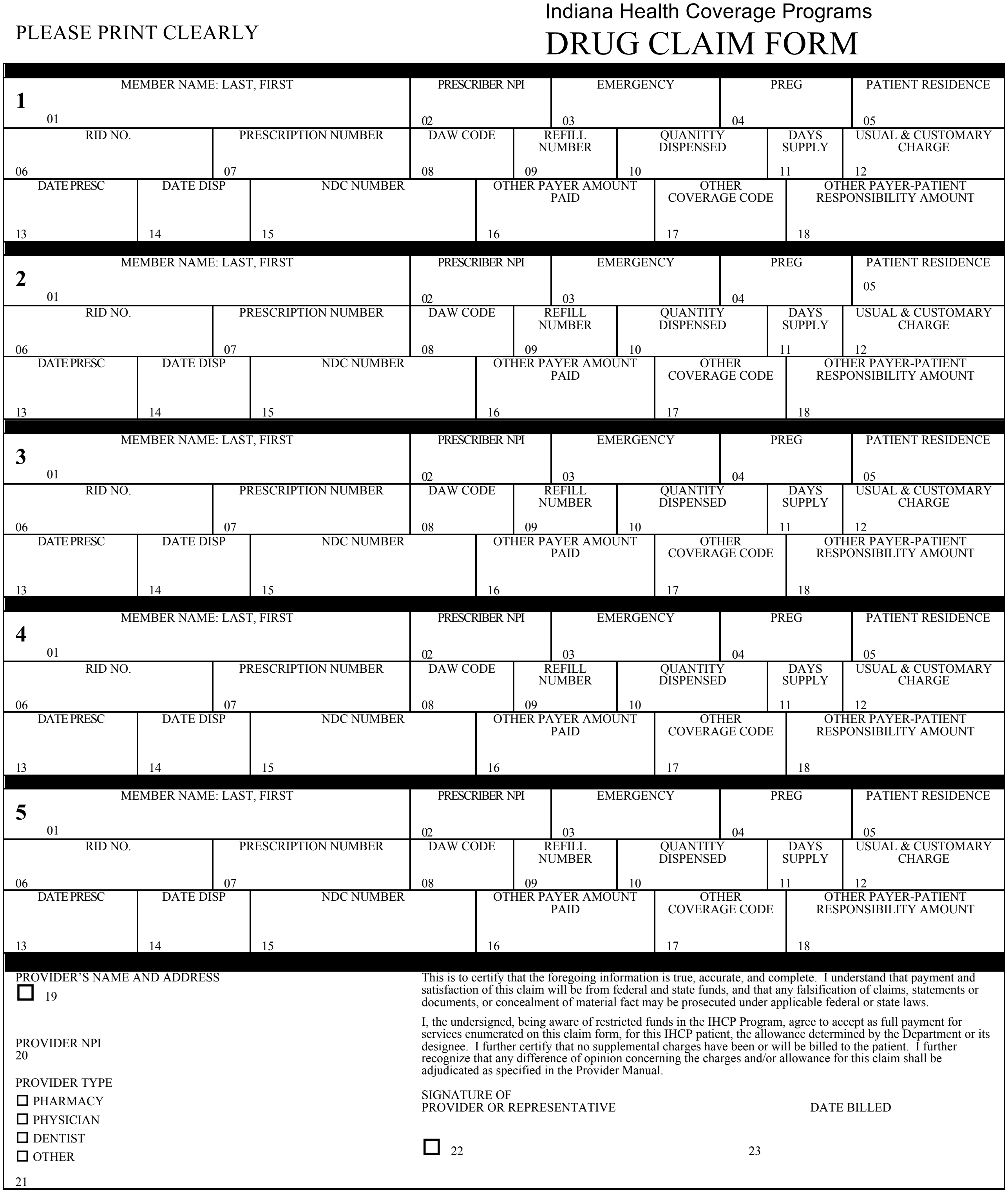
69 ADMIT DX | 70 PATIENT REASON DX a b c | 71 PPS CODE | 72 ECI a b c | 73

74 PRINCIPAL PROCEDURE CODE DATE | a. OTHER PROCEDURE CODE DATE | b. OTHER PROCEDURE CODE DATE | 75 | 76 ATTENDING NPI QUAL
LAST FIRST
c. OTHER PROCEDURE CODE DATE | d. OTHER PROCEDURE CODE DATE | e. OTHER PROCEDURE CODE DATE | 77 OPERATING NPI QUAL
LAST FIRST

80 REMARKS | 81CC a | 78 OTHER NPI QUAL
b | LAST FIRST
c | 79 OTHER NPI QUAL
d | LAST FIRST

UB-04 CMS-1450 APPROVED OMB NO. 0938-0997 NUBC TFP24394485 THE CERTIFICATIONS ON THE REVERSE APPLY TO THIS BILL AND ARE MADE A PART HEREOF.

Appendix D: Drug Claim Form



PLEASE PRINT CLEARLY

Indiana Health Coverage Programs

DRUG CLAIM FORM

1	MEMBER NAME: LAST, FIRST 01	PRESCRIBER NPI 02	EMERGENCY 03	PREG 04	PATIENT RESIDENCE 05	
RID NO. 06	PRESCRIPTION NUMBER 07	DAW CODE 08	REFILL NUMBER 09	QUANITTY DISPENSED 10	DAYS SUPPLY 11	USUAL & CUSTOMARY CHARGE 12
DATE PRESC 13	DATE DISP 14	NDC NUMBER 15	OTHER PAYER AMOUNT PAID 16	OTHER COVERAGE CODE 17	OTHER PAYER-PATIENT RESPONSIBILITY AMOUNT 18	

2	MEMBER NAME: LAST, FIRST 01	PRESCRIBER NPI 02	EMERGENCY 03	PREG 04	PATIENT RESIDENCE 05	
RID NO. 06	PRESCRIPTION NUMBER 07	DAW CODE 08	REFILL NUMBER 09	QUANITTY DISPENSED 10	DAYS SUPPLY 11	USUAL & CUSTOMARY CHARGE 12
DATE PRESC 13	DATE DISP 14	NDC NUMBER 15	OTHER PAYER AMOUNT PAID 16	OTHER COVERAGE CODE 17	OTHER PAYER-PATIENT RESPONSIBILITY AMOUNT 18	

3	MEMBER NAME: LAST, FIRST 01	PRESCRIBER NPI 02	EMERGENCY 03	PREG 04	PATIENT RESIDENCE 05	
RID NO. 06	PRESCRIPTION NUMBER 07	DAW CODE 08	REFILL NUMBER 09	QUANITTY DISPENSED 10	DAYS SUPPLY 11	USUAL & CUSTOMARY CHARGE 12
DATE PRESC 13	DATE DISP 14	NDC NUMBER 15	OTHER PAYER AMOUNT PAID 16	OTHER COVERAGE CODE 17	OTHER PAYER-PATIENT RESPONSIBILITY AMOUNT 18	

4	MEMBER NAME: LAST, FIRST 01	PRESCRIBER NPI 02	EMERGENCY 03	PREG 04	PATIENT RESIDENCE 05	
RID NO. 06	PRESCRIPTION NUMBER 07	DAW CODE 08	REFILL NUMBER 09	QUANITTY DISPENSED 10	DAYS SUPPLY 11	USUAL & CUSTOMARY CHARGE 12
DATE PRESC 13	DATE DISP 14	NDC NUMBER 15	OTHER PAYER AMOUNT PAID 16	OTHER COVERAGE CODE 17	OTHER PAYER-PATIENT RESPONSIBILITY AMOUNT 18	

5	MEMBER NAME: LAST, FIRST 01	PRESCRIBER NPI 02	EMERGENCY 03	PREG 04	PATIENT RESIDENCE 05	
RID NO. 06	PRESCRIPTION NUMBER 07	DAW CODE 08	REFILL NUMBER 09	QUANITTY DISPENSED 10	DAYS SUPPLY 11	USUAL & CUSTOMARY CHARGE 12
DATE PRESC 13	DATE DISP 14	NDC NUMBER 15	OTHER PAYER AMOUNT PAID 16	OTHER COVERAGE CODE 17	OTHER PAYER-PATIENT RESPONSIBILITY AMOUNT 18	

PROVIDER'S NAME AND ADDRESS
□ 19

PROVIDER NPI
20

PROVIDER TYPE
□ PHARMACY
□ PHYSICIAN
□ DENTIST
□ OTHER
21

This is to certify that the foregoing information is true, accurate, and complete. I understand that payment and satisfaction of this claim will be from federal and state funds, and that any falsification of claims, statements or documents, or concealment of material fact may be prosecuted under applicable federal or state laws.

I, the undersigned, being aware of restricted funds in the IHCP Program, agree to accept as full payment for services enumerated on this claim form, for this IHCP patient, the allowance determined by the Department or its designee. I further certify that no supplemental charges have been or will be billed to the patient. I further recognize that any difference of opinion concerning the charges and/or allowance for this claim shall be adjudicated as specified in the Provider Manual.

SIGNATURE OF
PROVIDER OR REPRESENTATIVE

□ 22

DATE BILLED

23

Appendix E: Interview with Chris Bunnell

Interview – October 16, 2012

What is your background in fraud detection?

Earlier on in my career I was in hospital administration. I was involved in operations. I ran a small hospital where I was in charge of everything except nursing and finance. I have sort of this broad stroke health care background.

In 2000 one of my friends was running the state of Maine Medicaid IT services and he wanted to support the data Program Integrity efforts so he went out and hired a company called Sapient from Boston. There was a million dollars paid for the program, the software and the hardware. There was $15 million recovered from the providers in the first year. Sapient saw that there were opportunities to do this kind of work for other states. And they founded a company that would do that for state governments. The company was called HWT and I was one of the first employees there.

We received contracts from several states. First there was Kentucky, Washington, West Virginia, Colorado, Rhode Island, Oklahoma and there were lots of other smaller projects that we did.

With the states we would work with an extract from their MMIS data that we would take from their COBOL-based mainframe. We put that into Oracle, relationonized it and essentially put it in a format that allowed us to go and review the claims electronically. With paper claims you could only review 100 at a time. We could now review millions or hundreds of millions of claims. The Washington database had some three billion lines of claim data. We helped state Program Integrity departments recover $100 millino dollars. The company was eventually sold to UnitedHealthcare. This was ten years ago, so it was a lot more money than it is now. It was pretty significant at that time seeing how no one had really much faith that we could plug holes that were in the Medicaid program.

After HWT was sold I became one of the founding members of Predicted Solutions, which was a similar project. We developed many products and services and offered consulting services. We gained a little bit of notoriety and Urix purchased the intellectual property. Urix spun off Predicted Solutions into a company called Verisk.

What kind of cases would you work on?

We started with basic algorithms by conducting brainstorming sessions with the people in our company that had healthcare experience. We would get into a room and would try to develop several pharmacy algorithms. Next we would convey the idea to a developer who would write code to test the data. If we found there was evidence of overpayments in the data we would develop a report and bring it to our client to get approval to test against their live dataset.

We later developed specialty algorithms. One of the most successful was called Drug Rebate, another was called JCode Rebate. These specialty algorithms help recover rebate dollars from drug manufacturers based upon pre-negotiated contracts.

The rebate process is complicated and requires the state to jump through many hoops in order to recover the rebate amount. We automated the process and recovered millions of dollars for each state that asked us to run the analysis. A few years back we were in talks with California and they had nearly a billion dollars owed to them through drug rebate.

You said that you would get the extracts from COBOL and get them in to a relational database. What would the relational database look like?

Back then out of the fifty programs in the United States all were COBOL-based system. We would obtain an extract of data load it into Oracle. We would break it apart into tables and connect it so you could run any type of query.

Do you know any details on how you would break it down? What kind of tables?

It is pretty similar to what we are developing now. I'm not sure if you saw my response to the work on Kansas. It is the same idea, not much has changed.

(work on Kansas is four different tables, one for each claim type (IP, LP, RX, Other))

All the data that we ran up against was MMIS. We never used some kind of odd datasource like MSIS. The idea of not using source data just doesn't seem to make sense to me.

(MMIS is a data format that differs per state. MSIS is a data format that is the same for all states.)

So you mean that they are odd datasource because they transform the MMIS data into something called MSIS before handing it out?

Any system that seeks to take states with what is unique about them and their system and homogenize the data into a standard format is always going to have limitations. The biggest limitation of all is the inability to get all the states to write an extract that delivers exactly the same type of data.

So homogenizing the data either through MSIS or T-MSIS is not helpful in the aid of finding fraud?

I wouldn't say it is not helpful. I would say it is not optimal.

Because MSIS might require certain fields that might not be in the source systems so you get fragmented data?

Yes.

Besides the MMIS data that you guys used, did you guys use any external data for fraud detection?

We did. We would buy data from companies like Ingenix. They had procedure codes, DRG-codes and descriptions. First Databank is well known for having a really good data file for drugs. In our data we would have the NDC-code. We needed to be able to link out and get all the details of that particular drug and that was contained in these files that you buy.

What would be the biggest bottleneck that you would encounter when trying to find fraud or recover the fraud money?

Incentivizing all the stakeholders.

Would you say that the bureaucracy among states and the federal government is a bottleneck in recovering overpayments?

Yes. Here is an example: let's say we are in a state. And we are working alongside people, some of whom were extremely committed the finding and recovering overpayments. And some who are embarrassed by the fact that they had paid these claims to begin with. Think of it from this perspective: their organization already paid this money while they shouldn't have. And we're going out to recover it. This can make it look like they did a really bad job paying claims in the first place. So from that perspective it is an embarrassment for some people in the program.

So you are working along some people who are really committed to recovery yet others care less. And other people who really don't want you to recover because it makes other people in the program look bad or even them themselves look bad because they paid the overpayments in the first place.

So you have various levels of commitment to things. Our experience of Washington State was that everyone was really on the same page. Mostly because they had wonderful leadership from Heidi Robbins Brown, Katty Ott and other people out there that just get it. They have a great perspective and it really helps them do a great job.

Did you take a look at my model?

I am not that familiar. Maybe if you could walk me through it.

(I brilliantly describe my thesis in a nutshell.)

So a star schema is not anything new. People have been doing that for a long time. I guess I can comment on it this way: anything that makes the analysis better is useful.

The thing is that there is a roadblock right now with people being unable to do the analysis. Speed of the system is not really a big deal. Whether the results of fifty million dollars of fraud that they found comes today, tomorrow or next month. It doesn't make that big a difference. As long as it comes.

What is much more important is the quality of what you are doing. The understanding of what is being done by the analyst.

It is not only the speed, it is also the way the data is structured. It allows easier ways of slicing through the data. So it would support for example faster algorithm development.

What is the single big improvement that the structure you are suggesting has compared to what we are currently doing?

It is very easy to put BI tools on top of it to get reports of the spending for example. And you can easily build profiles of certain taxonomies of providers or certain providers. It is very efficient in doing that.

I think it is an advantage to work like this. There is a lot of research to the way a data warehouse is structured.

Interview – October 23, 2012

I want to evaluate with you the dimensions of my data warehouse. I would like you to specify for each dimension if it would be used for fraud detection.

Diagnosis

Yes

Patient

The specific patient? I don't know of any algorithms that would look at an episode of care. It is not a bad idea but I know that the feds don't have any.

It is an important field, no doubt about it. It is a key field.

Service

Yes

Provider

Yes, absolutely

Drug

Yes

Location

Yes, that's useful

Date

Absolutely

Outcome

The outcome, that's an interesting question. We don't have any outcome data.

I was looking at the UB-04 claim form. And that one specifies an outcome.

I have never seen that in the data anywhere. Maybe they only recently started collecting it. I have never run across it in my career.

The question is, how do they get the outcome? Say, you have had surgery on your neck. They discharge you and you go home and you recuperate. They would have to do a follow up.

Health plan

I think you could find a reason to use it. It is not used very much. Especially in Medicaid it is pretty much standard stuff. If you are eligible you are covered for everything on the program. That's in most cases.

If you would develop, look at the state source data that they have in the system. Would you say they have to develop a system based specifically on 1 state. Like a data warehouse. Or say, try to, get 1 standard used for all states?

You can't really make 1 standard for all states. You could do a system like that but it would need to have a way to set the specifications. So you could have a certain core group of things that are standard across all systems and then there would be items in the system that you could adjust. And would act differently depending upon the system. The reason for that is that every state is unique. What works for California – a big state, a desert state, huge population – doesn't work necessarily for Rhode Island. A lot of things would work but not everything. Every state has its own laws, its own rules, its own regulations. Those are reflected by the systems to some extent. Not a huge extent but a small extent.

In our e-mail conversation you stated that every state has a decision support system (DSS)?

Almost every state has a DSS.

Is the DSS available for usage by the people that are looking for fraud?

Yes, absolutely. In fact, in most cases, those are the people that push for that type of a tool.

Would you use the DSS to see at which locations you want to go look for fraud?

Yes.

Would you know to what level of detail the DSS would be able to provide information?

A smaller data warehouse containing the most common fields is generally used by most users.

Bibliography

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

Center for Disesae Control and Prevention. (2012, 10 04). *ICD - ICD-9-CM - International Classification of Diseases, Ninth Revision, Clinical Modification*. Retrieved 10 08, 2012, from Center for Disesae Control and Prevention: http://www.cdc.gov/nchs/icd/icd9cm.htm

Centers for Medicare & Medicaid Services. (2012, 07 24). *HCPCS - General Information | Centers for Medicare & Medicaid Services*. Retrieved 08 10, 2012, from Centers for Medicare & Medicaid Services: http://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html

Centers for Medicare & Medicaid Services. (2012, August 3). *Medicare.gov - Types of Long-Term Care*. Retrieved October 9, 2012, from Medicare.gov: the official U.S. government site for Medicare: http://www.medicare.gov/longtermcare/static/TypesOverview.asp

Centers for Medicare & Medicaid Services. (2012, 03 25). *National Provider Identifier Standard (NPI) | Centers for Medicare & Medicaid Services*. Retrieved 10 08, 2012, from Centers for Medicare & Medicaid Services: http://www.cms.gov/Regulations-and-Guidance/HIPAA-Administrative-Simplification/NationalProvIdentStand/index.html

Centers for Medicare and Medicaid Services. (2010, August). Medicaid and Chip Statistical Information System - File Specification and Data Dictionary. (Release 3.1). Retrieved from http://www.cms.gov/MSIS/Downloads/msisdd2010.pdf

Copeland, L., Edberg, D., & Wendel, J. (2011). Applying Business Intelligence Concepts to Medicaid Claim Fraud Detection.

Department of Health and Human Services. (2004, January 23). HIPAA Administrative Simplification: Standard Unique Health Identifier for Health Care Providers; Final Rule. *Federal Register , 69* (15), pp. 3434-3469.

Department of Justice. (2006, June 29). *Department of Justice*. Retrieved August 27, 2012, from Tenet Healthcare Corporation to Pay U.S. more than $900 Million to Resolve False Claims Act Allegations: http://www.justice.gov/opa/pr/2006/June/06_civ_406.html

Federal Bureau of Investigation. (2009). *2009 Financial Crimes Report*. Retrieved February 6, 2012, from http://www.fbi.gov/stats-services/publications/financial-crimes-report-2009/financial-crimes-report-2009#health

Food and Drug Administration. (2012, 10 1). *Drug Approvals and Databases > National Drug Code Directory*. Retrieved 10 8, 2012, from U.S. Food and Drug Administration: http://www.fda.gov/Drugs/InformationOnDrugs/ucm142438.htm

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., et al. (1997). Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab and Sub-Totals. *Data Mining and Knowledge Discovery* (1), 29-53.

Health Care Cost Institute. (2012). *Health Care Cost and Utilization Report: 2010.*

Hevner, A. R., Ram, S., March, S. T., & Park, J. (2004). Design Science in Information System Research. *MIS Quaerterly , 28* (1), 75-105.

Institute of Medicine. (2010). *The Healthcare Imperative: Lowering Costs and Improving Outcomes - Workshop Series Summary.* Washington, D.C.: The National Academies Press.

Kahn, J. G., Kronick, R., Kreger, M., & Gans, D. N. (2005). The Cost of Health Insurance Administration In California: Estimates For Insurers, Physicians, And Hospitals. *Health Affairs* (6), 1629-1639.

Kelley, R. (2009). *Where can $700 Billion in Waste be cut annually from the US Healthcare System?* Ann Arbor, MI: Thomson Reuters.

Kimball, R., & Ross, M. (2008). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (2nd Edition, Kindle Edition ed.). John Wiley and Sons.

Li, J., Huang, K.-Y., Jin, J., & Shi, J. (2008). A survey on statistical methods for health care fraud detection. *Heal Care Manage Sci* (11), 275-287.

Major, J. A., & Riedinger, D. R. (2002). EFD: A Hybrid Knowledge/Statistical-Based System for the Detection of Fraud. *The Journal of Risk and Insurance , 69* (3), 309-324.

Medicaid. (n.d.). Retrieved February 10, 2012, from http://www.medicaid.gov/

Morris, L. (2009). Combating Fraud In Health Care: An Essential Component Of Any Cost Containment Strategy. *Health Affairs , 28* (5), 1351-1356.

National Health Care Anti-Fraud Association. (year unknown). *The Problem of Health Care Fraud*. Retrieved February 2, 2012, from National Health Care Anti-Fraud Association: http://www.nhcaa.org/eweb/DynamicPage.aspx?webcode=anti_fraud_resource_centr&wpscode=TheProblemOfHCFraud

Naumann, F., & Herschel, M. (2010). An Introduction to Duplicate Detection. *Synthesis Lectures on Data Management , 2* (1), 1-87.

Office of Inspector General. (2012). *Inappropriate and Questionable Billing by Medicare Home Health Agencies.* Department of Health and Human Services.

Ortega, P., Figueroa, C., & Ruz, G. (2006). A medical claim fraud/abuse detection system based on data mining: a case study in Chile. Las Vegas, Nevada, USA.

Potetz, L., Cubanski, J., & Neuman, T. (2011). *Medicare Spending and Financing.* The Henry J. Kaiser Family Foundation.

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

Sokol, L., Garcia, B., West, M., Rodriguez, J., & Johnson, K. (2001). Precursory Steps to Mining HCFA Health Care Claims. *Proceedings of the 34th Hawaii International Conference on System Sciences.*

Sparrow, M. (2000). *License to Steal: How Fraud bleeds america's health care system.* Boulder: Westview Press.

Taxpayers Against Fraud Education Fund. (n.d.). *Top 20 False Claim Act Cases*. Retrieved August 27, 2012, from The False Claims Act Legal Center: http://www.taf.org/top20.htm

Travaille, P., Müller, R. M., Thornton, D., & Hillegersberg, J. v. (2011). Electronic Fraud Detection in the U.S. Medicaid Healthcare Program: Lessons Learned from other Industries. *AMCIS 2011 Proceedings - All Submissions.* Paper 328.

United States Attorney's Office, Central District of California. (2012, May 2). *USDOJ: US Attorney's Office - CENTRAL DISTRICT OF CALIFORNIA - 055*. Retrieved September 4, 2012, from United States Department of Justice: http://www.justice.gov/usao/cac/Pressroom/2012/055.html

United States Department of Health & Human Services. (2009, September 2). *Justice Department Announces Largest Health Care Fraud Settlement in its History*. Retrieved August 27, 2012, from United States Department of Health & Human Services: http://www.hhs.gov/news/press/2009pres/09/20090902a.html

United States Department of Justice. (2011, 04 12). *Miami Doctor Convicted in $23 Million Medicare Fraud Scheme*. Retrieved 09 04, 2012, from The United States Department of Justice: http://www.justice.gov/opa/pr/2011/April/11-crm-482.html

United States Department of Justice. (2012, February 27). *Welcome to the United States Department of Justice*. Retrieved September 4, 2012, from United States Department of Justice: http://www.justice.gov/opa/pr/2012/February/12-crm-256.html

United States General Accounting Office. (2000). *Health Care Fraud: Schemes to Defraud Medicare, Medicaid, and Private Health Care Insurers.*

Washington Publishing Company. (2012). *WPC References*. Retrieved 10 08, 2012, from Washington Publishing Company: http://www.wpc-edi.com/reference/

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