Improving the Medicaid eligibility determination process using big data

UNIVERSITY OF TWENTE.

CLEMSON
COMPUTING AND
INFORMATION TECHNOLOGY

Michel
Brinkhuis

s1011839

University of Twente
Clemson University

Supervisors:
Chintan Amrit
Robin Aly
Dallas Thornton

COLOPHON

M.E. BRINKHUIS
Business Information Technology
MASTER THESIS

DATE
August 3, 2015

VERSION
1.0

STATUS
Final version

PROJECT
IMPROVING THE MEDICAID ELIGIBILITY DETERMINATION PROCESS USING BIG DATA

AUTHOR
M.E. Brinkhuis

PHONE
+316 389 77 674

E-MAIL
mail@michelbrinkhuis.nl

SUPERVISORS
Dr. Chintan Amrit, University of Twente
Dr. Robin Aly, University of Twente
Dallas Thornton MBA, Clemson University

PREFACE

When I was in high school I saw advertisements from the University of Twente, displaying the slogan: 'Do you want to become the next Bill Gates? Choose Business & IT'. A message that sounded tempting. Now, 7 years later, I must conclude that I only partially succeeded. I made it to the United States, and learned a lot about the field of business and IT. However: we have our differences. Bill Gates dropped out of a university study, whereas I just try to finish it. And that's why this thesis is written: hopefully the last few miles to a new milestone on the road of success.

When starting with this thesis, I expected the opportunity to apply the knowledge I gathered during the years of studying Business Information Technology at the University of Twente. Surprisingly, next to applying knowledge, working on this thesis turned out to be an opportunity to learn much more. From getting to know the workings of mainframe database systems to applying machine learning algorithms on large datasets. From creating dashboards in Qlikview to getting to know the playing field of managing government IT systems. And to finish it all: getting to know the United States of America.

I got the chance to write my thesis at Clemson Computing and Information Technology. For making this possible I want to thank Dallas Thornton. However, next to having a place to do research, also people who monitor the scientific quality are needed. Therefore I want to thank Chintan Amrit and Robin Aly, who provided me with lots of feedback on how to improve this thesis.

Hopefully this thesis brings valuable insights to you.

Michel Brinkhuis

ABSTRACT

The United States spend a lot of money on their healthcare systems. The program Medicaid is part of the healthcare system. In this research the Medicaid application process is analyzed. A literature review on the state of art of this process is conducted, as well as a review on healthcare fraud. The topics are selected based on a classification of healthcare waste by (Hackbarth 2012). There is little literature available on the Medicaid application process and implications itself. The available literature mentions the language, method of application and evidence that citizens need to provide as the factors influencing this process. A 24 factor classification of healthcare fraud is provided based on existing literature. The application data from the Medicaid program in the state of South Carolina is analyzed using machine learning technologies in order to find stages in the application process where waste occurs. Based on this analysis the author recommends to use more external data sources for the validation of information, and to create better tools to monitor the Medicaid application process.

Keywords:

Medicaid, Machine learning, eligibility determination, health insurance, healthcare fraud, process analysis

MANAGEMENT SUMMARY

Context

Health care spending contributes to a significant part of the U.S. government spending. One of the government-run programs is Medicaid, on which the spending in 2012 was 415 billion U.S. Dollars on a national level. Since this is taxpayer's money being spent, taxpayers demand from the government that it is spent well. So, the question arises: is it spend well? Do the people who need Medicaid really have access to it? In this thesis the focus is on the eligibility determination process, which from a citizen perspective is also known as the Medicaid application process. This research is conducted at Clemson Computing and Information Technology (CCIT), which manages the IT systems for South Carolina Department of Health and Human Services (SCDHHS). SCDHHS manages the Medicaid program for the state of South Carolina.

Objectives

The objective of this research is to improve the Medicaid application process. CCIT/SCDHHS know that often applications take a long time to process, and for some applications the eligibility is determined the same day as the application date. In this research the objective is to find out why these differences exist, based on a dataset of Medicaid applications. The central research question that this research answers is:

How can the Medicaid eligibility determination process be optimized?

Methods

Based on a classification of waste in healthcare by (Hackbarth 2012) two important areas are identified: administrative complexity and fraud. The research starts by performing a literature study into those two topics. Secondly three machine learning technologies (Support Vector Machines, Random Forests and Neural Networks) are applied on a dataset containing Medicaid application information. Based on the knowledge from literature, and the findings of the data mining approach, conclusions will be drawn.

Results

The literature study on the Medicaid application process shows three areas of interest: the language, the method of application and the amount of proof people need to provide alongside

with their application. 24 Areas of healthcare fraud are identified, of which two are related to the Medicaid application process: identity fraud and misrepresenting one's eligibility.

Of all 'hard case' applications processed by case workers, online applications take the longest time to complete. 'Hard cases' are defined as follows:

- An application submitted online, that needs a caseworker to be completed
- Applications submitted on paper, by phone and in person, that need at least one evidence item to be verified by a caseworker.

	Total	'Hard cases'	On time	Too late
Online	28493	2137	482	1655
Paper	10408	9513	8634	879
Phone	40	37	30	7
In person	453	424	421	3

The online process is further analyzed. Of all applications that are handled too late, the majority is processed after 75 days after submission. A set of characteristics of an application is selected (independent variables) and correlations between those characteristics and the processing time are calculated. No strong correlations are found.

Secondly three machine learning algorithms are trained on the dataset. The selected algorithms are Support Vector Machines, Random Forests and a Neural Network. All executed tests had unsatisfying accuracies, all below 50% accuracy. Of the excluded variables the Method of Application seems to have the biggest influence: the accuracy is lowest when this variable is excluded from the dataset. The best performing data mining method are neural networks. Even though accuracies are low in general in this case, neural networks tend to give the 'highest' accuracy. However, the difference between the three algorithms was small in most cases. This algorithm is able to predict with an accuracy of 47,92 percent if an application would be handled on time. However, based on this result it turns out that even the best performing method used in this research does not provide sufficient performance to be usable.

Conclusions

There is little literature available on the Medicaid application process. More literature is available on the topic of healthcare fraud, but only a few papers touch topics that are relevant to the

Medicaid application process. Data analysis shows that a high percentage of online applications that are handled by caseworkers are completed after 45 days: over 77 percent.

The data analysis does not show specific characteristics of an application that contribute to an increased processing time. Two recommendations for practice are given. First, by integrating more extensive automated validation using external data sources more online applications can be automatically processed, without the need for a caseworker looking at the application. This might reduce employee workload, and it is expected to reduce processing times. Furthermore, the process can be monitored by generating real-time insight into the application handling process, using business intelligence or dashboards. These insights might help to identify applications that are waiting to be completed for a long time, and need priority handling.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ADR	Action Design Research
BENDEX	Beneficiary and Earnings Data Exchange
BI	Business Intelligence
CCIT	Clemson University Computing & Information Technology
DM	Data Mining
DMV	Department of Motor Vehicles
DRA	Deficit Reduction Act
DSM	Design Science Methodology
DSRM	Design Science Research Methodology
FPL	Federal Poverty Line
KD	Knowledge Discovery
KDD	Knowledge Discovery in Databases
KDP	Knowledge Discovery Process
MAGI	Modified Adjusted Gross Income
MCO	Managed Care Organization
SCDHHS	South Carolina Department of Health and Human Services
SC	South Carolina
SSN	Social Security Number
USCIS	United States Citizenship and Immigration Services
U.S.	United States

1 INTRODUCTION

Health care spending contributes to a significant part of the U.S. government spending. One of the government-run programs is Medicaid, on which the spending in 2012 was 415 billion U.S. Dollar on a national level. Since this is taxpayer's money being spent, taxpayers demand from the government it is spent well. The question arises: is it spend well? Do the people who need Medicaid really have access to it?

Approaches for monitoring and analyzing the processes that accompany the Medicaid program are needed. This research aims at improving the methods. In this chapter the context of this research will be introduced, as well as the research objectives and questions. Furthermore, an outline is provided for the rest of this document.

1.1 Context

Medicaid is a social health care program for U.S. citizens with low income and limited resources. The program is managed by individual states, and funded by a combination of government and state funding. All U.S. Citizens having an income up to 133% of the poverty line qualify for enrollment in the Medicaid program. On the state level the program is managed by the South Carolina Department of Health and Human Services (SCDHHS). SCDHHS has outsourced their IT operations to CCIT: The Clemson University Computing & Information Technology department.

In the fiscal year 2012 the total spending on the Medicaid program summed up to 415 billion U.S. dollar. 4,8 Billion U.S. dollar of this total was spent in the state of South Carolina (The Henry J. Kaiser Family Foundation 2015). Since this is taxpayer's money people demand from the government that it is spent well. However, even low estimates show that in six categories of waste 20% of the total health care expenditures is lost (Hackbarth 2012). The six categories of waste are shown in Table 1.

Managing a healthcare program involves handling a lot of data. This data is stored in several systems, such as systems for determining eligibility of new applicants, processing and paying claims, and reporting to several governmental bodies. CCIT plans on creating a data warehouse to integrate this data. This will provide BI capabilities, to analyze both the administrative processes as well to detect fraud and abuse.

Waste category	Description
Failures of care delivery	The waste that comes with poor execution or lack of widespread adoption of known best care processes.
Failures of care coordination	Faulty coordination of care provided, resulting in e.g. complications, increased dependency and declines in functional status.
Overtreatment	Waste that comes from providing care to patients of which according to sound science and the patients' own preferences cannot possibly help them.
Administrative complexity	Waste created by the inefficient or misguided rules.
Pricing failures	Waste resulted from non-market prices
Fraud and abuse	The waste that comes as fraudsters issue fake bills and run scams, and also from the blunt procedures of inspection and regulation that everyone faces because of the misbehaviors of a very few.

TABLE 1 - SIX CATEGORIES OF WASTE (HACKBARTH 2012)

The enrollment process is one of the administrative processes. Citizens can apply for Medicaid using several methods: there is an online portal, a phone number, people can apply by paper forms or visit an office. Applications are processed by a computer system, called an eligibility determination system. If the system is not able to process the application, for example because some information cannot be automatically validated, a SCDHHS employee will look into the case and resolve the issues. These employees are called *case workers*.

On one side South Carolina has very strict rules compared to some other states when it comes to eligibility, while on the other side the state tries to actively find people who are eligible but are not enrolled yet (Galewitz 2013). Therefore, having an optimized enrollment process is important. However: the processing time for applications submitted by citizens varies significantly. For one citizen it might take just 2 days before his or her eligibility has been determined, while another citizen that applies on the same day might have to wait 50 days before a final eligibility decision is made.

Not only do some people have to wait a long time before they know the outcome of their application, the current process might also influence the pressure on the case workers. If there is a large backlog of cases to be processed, it might influence the quality of their work. This might

make it easier for people to get falsely enrolled. These people can wrongly declare money from the Medicaid program. This is an example of 'Fraud and abuse' in Hackbarth's classification, where the application process in general might suffer from 'Administrative complexity'.

1.2 Research motivation

The healthcare domain is a complex setting. There are many different actors involved, and the number of entities is enormous. Data can be noisy due to the fact that all information is entered by humans and processed by humans. Furthermore Medicaid data is stored fragmented in several systems, as explained before.

This introduces the problem this research focuses on. How can we find ways to optimize the enrollment process? A lot of money is spent on the Medicaid program, and it is important its spent well. Linking the classification of (Hackbarth 2012) to the Medicaid enrollment process creates a presumption there is room for improvement. Bottlenecks in the enrollment process can be identified and removed. And if there are ways to optimize the enrollment process, the processing time for applications will go down. Case workers can then spend more time on individual cases that really need attention, resulting in a smaller chance of fraud taking place and taking more time to help the people that really need the Medicaid program.

1.3 Research questions & objectives

Based on the identified problem a research goal can be formulated. Since the problem can be addressed by designing a solution, this research is structured according to the Design Science Methodology (DSM) approach (Wieringa 2014). Wieringa provides a template for design problems, which helps to clarify the relation between the problem context, artifact, requirements and goals. The research goal can be structured according to the following template:

> **Improve** the problem context
> **By treating it** with a (re)designed artifact
> **Such that** artifact requirements
> **In order to** stakeholder goals

This template shows that four elements need to be identified in order to arrive at a research goal. In DSM the artifact is the element that has to be designed by the researcher. In this case this is the method of improving the enrollment process. Several machine learning technologies will be

applied and compared. This artifact interacts with a certain context, which is in this case the enrollment process itself. The interaction of the artifact with the context introduces the requirements. The requirement is to reduce the processing time of Medicaid applications, in order to reduce waste in the Medicaid application process.

Together these elements result in the following research goal:

Improving the Medicaid application process
By developing an approach to identify bottlenecks
Such that the application processing time can be reduced
In order to reduce waste in the application process

This research goal can be translated to the following research question:

How can the Medicaid eligibility determination process be optimized?

A number of sub questions will help answering the research question. The research starts with getting insights in literature on the Medicaid enrollment process and healthcare fraud in general. Medicaid enrollment is program-specific. Fraud however is a broader topic, and therefore healthcare fraud in general will be discussed. This answers the first sub question:

SQ1. What is the state-of-art in literature regarding Medicaid enrollment processes?
SQ2. How can Medicaid enrollment fraud types be classified?

The available systems contain a lot of data, stored in extended structures. There are multiple systems in use, and each of them contains over 300 tables. A selection has to be made to these data sets in order to create a data model only including the right and relevant information to analyze the Medicaid application process. The second sub question reads as:

SQ3. What data is needed to analyze the Medicaid application process?

A data structure is not the only factor needed to find analyze the application process. When the data is structured in the right way, knowledge has to be gathered from the dataset. Therefore data mining will be applied on the data. The right data mining method can be found by answering the fourth sub question:

SQ4. What is the best suitable data mining approach to analyze the Medicaid application process?

Finally, the performance of the selected data mining method on the created dataset will be evaluated:

SQ5. What do the results of the data analysis tell about the Medicaid application process?

1.4 Relevance of this study

The research will be conducted at CCIT. This introduces a need for both a scientific relevance as well relevance for CCIT.

1.4.1 Scientific relevance

This research contributes on multiple aspects to the scientific field. Firstly it will provide a classification of health care fraud types. Using an extensive literature study an overview of the field will be created. Second contribution is the application of machine learning methods on the Medicaid enrollment dataset, providing insights in the Medicaid enrollment process. Third, the literature study on Medicaid enrollment processes will provide an overview of the most discussed items related to the enrollment process.

1.4.2 Relevance for CCIT

CCIT wants to provide SCDHHS with insights in the Medicaid application process from a data perspective. Currently there are no real BI implementations in use. This research introduces the first steps in the field of data mining aimed at process optimization. This research contributes to this project by both developing the data model to analyze the application process, as well as by providing information based on that analysis.

The outcomes of this research can furthermore help other organizations who deal with application processes, by describing an approach on how to deal with this problem. These organizations can be organizations responsible for managing Medicaid in other states for example.

1.5 Stakeholder analysis

It might sound a bit confusing: Medicaid is managed by individual states. In the case of South Carolina it is managed by SCDHHS. The IT operations are managed by Clemson University's CCIT. To start with a good understanding of the playing field a stakeholder analysis is conducted. The approach designed by (Clements and Bass 2010) is followed. They identify several groups of stakeholders related to a company. The company in this case is CCIT, and its relationship with other entities is depicted in Figure 1.

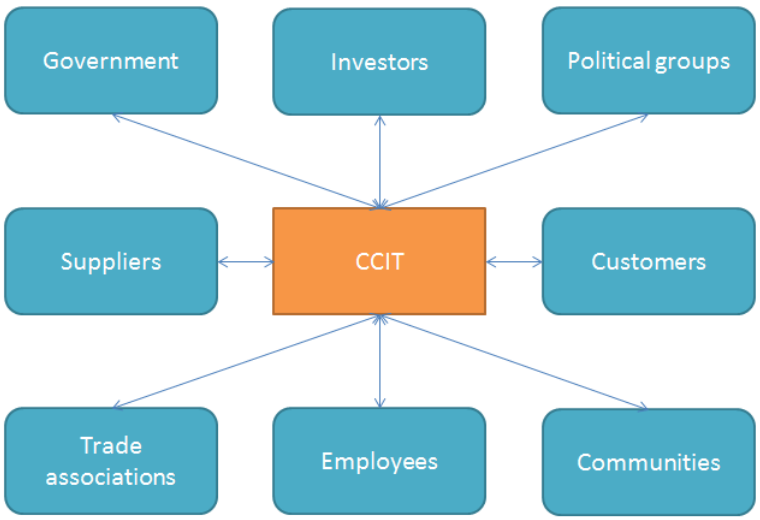


FIGURE 1 - STAKEHOLDER CATEGORIES RELATED TO CCIT

Each of the groups will be discussed. Some of them turn out not to be relevant to this project, while other groups might consists of several stakeholders, each having their own interests. Trade associations and communities are not relevant in this particular project, from the perspective of CCIT, and they will therefore not be discussed.

Government — CCIT is maintaining the system for the government. This makes the government a stakeholder. The government is also a stakeholder in a non-customer way. When working with personal identifiable data (PII) and personal healthcare information (PHI), legislation exists. An example is HIPAA: the Health Insurance Portability and Accountability Act which prescribes how sensible data should be treated.

Political groups — Political groups may influence the decision-making process. The project deals with taxpayer's money, and handles government data, political

	groups that come up for privacy and spending taxpayers' money well might have interests that need to be taken into account. There is however no direct relation between CCIT and political groups.
Suppliers	Software from several vendors is used. Mainframe systems use IBM software, linked with database software from CA (Computer Associates). Mainframes are migrated to IBM Cúram software, which uses Oracle database technology to store data.
Customers	CCIT manages the IT systems for SCDHHS. SCDHHS is therefore the customer. When taking end users also into account more customers exist: citizens, caseworkers and SCDHHS management use the IT applications to apply for Medicaid and manage the Medicaid program.
Employees	Employees are a stakeholder because they develop the system, and they maintain the relationship with the customer. Business analysts work with SCDHHS to design a system which fits the needs of SCDHHS. Programmers develop the system, and the quality assurance department verifies the work.
Investors	CCIT works on the Medicaid project having one customer: SCDHHS. SCDHHS is therefore the one and only investor in the project. Based on the money SCDHHS invests in the Medicaid project, CCIT can hire more employees and invest more time in the project.

1.6 Structure of this report

This thesis starts with an introduction to the problem, which in fact is the current chapter. Chapter 2 will introduce the Medicaid program, and point out the differences between Medicaid and Medicare. In chapter 3 an overview of the related topics will be presented, in the form of a structured literature to the topics of Medicaid enrollment, health insurance fraud and data mining methods for the detection of these fraud types. Chapter 4 introduces the research method that will be followed to design and structure the research. In chapter 5 the artifact design will be explained, and the results of applying the artifact can be found in chapter 6. This will help answer the research questions, which will be answered in chapter 7. In chapter 8 recommendations based on the research findings will be given.

2 AN INTRODUCTION TO MEDICAID

This chapter introduces the Medicaid and Medicare programs and explains the differences between them (2.1). Furthermore, in anticipation of the structured literature study of chapter 3, section 2.2 describes the current vision on Medicaid fraud from a government perspective.

2.1 Medicaid versus Medicare

In the United States two government-run health insurance programs are available for specific groups of citizens: Medicare and Medicaid (Department of Health and Human Services 2015). Medicare is health insurance for people ages 65 or older, people under 65 with certain disabilities and people of any age with End-Stage Renal Disease (permanent kidney failure requiring dialysis or a kidney transplant).

The Medicaid program is a health insurance program for people U.S. citizens with low income and limited resources or people with disabilities. The program is funded by both the federal government and the individual states. Medicaid is managed on a state-level. The precise requirements for citizens to be eligible differ on state-level too, because federal guidelines can be implemented in different ways. Approaches for handling the Medicaid program are thus unique to all states. The specifics of both Medicare and Medicaid are shown in Table 2. Related to the Medicaid program is CHIP: Children's Health Insurance Program. This provides health coverage to children, both under Medicaid and separate CHIP programs (Medicaid.gov 2015). The enrollment systems used in South Carolina are used for both Medicaid and CHIP eligibility determination.

Medicaid	Medicare
Aged 65 or older	Aged 65 or older
A child under 19	People under 65 with certain disabilities
Pregnant	People of any age with End-Stage Renal Disease (permanent kidney failure requiring dialysis or a kidney transplant)
Living with a disability	
A parent or adult caring for a child	
An adult without dependent children (state-specific)	
An eligible immigrant	

TABLE 2 - MEDICAID AND MEDICARE ELIGIBILITY REQUIREMENTS

In the state of South Carolina Medicaid is managed by SCDHHS. The IT-systems are outsourced to CCIT, as explained earlier on. The IT operations include several parts of the Medicaid program, of which the two most important are the systems for eligibility determination and claim handling. Claim handling consists of processing the claims and paying the claims. All systems also generates reports, that are used within SCDHHS but also sent to other government instances. The Centers for Medicare and Medicaid Services (CMS) for example require reports from SCDHHS.

2.2 Healthcare fraud schemes

As stated in chapter 1 one of the focus areas of this research is healthcare fraud. This can be analyzed not only from a scientific perspective, but also by analyzing what the government has published about this topic. In this sub chapter this will be done. The sources analyzed include publications from the U.S. government, which is fighting health care fraud, and the National Medical Fraud and Abuse Initiative (NMFAI), which helps organizations 'in the field' combat fraud and abuse.

First we will discuss the ways of fraud as identified by the U.S. government and NMFAI, followed by the literature study. The National Medicaid Fraud and Abuse Initiative makes a distinction between fraud and abuse (National Medicaid Fraud and Abuse Initiative 2000). They identify the terms as follows. *Fraud* is an intentional deception or misrepresentation made by a person with the knowledge that the deception could result in some unauthorized benefit to himself or some other person. *Abuse* on the other hand is provider practices that are inconsistent with sound fiscal, business or medical practices and result in unnecessary cost to the Medicaid program, or in reimbursement for services that are not medically necessary, or that fail to meet professionally recognized standards for health care. It also includes beneficiary practices that result in unnecessary costs to the Medicaid program. The Initiative identifies six fraud and abuse risk areas. These are shown in Figure 2.

FIGURE 2 – MANAGED CARE FRAUD TYPES (NATIONAL MEDICAID FRAUD AND ABUSE INITIATIVE 2000)

The Office of the Inspector General of DHHS (OIG) publishes a yearly compendium (OIG US DHHS 2014a), in which they give recommendations that are unimplemented and would most positively impact Health and Human services programs. Medicaid is one of these programs. They identify the problem of citizens who have been wrongly determined eligible too. With respect to Medicaid and Medicare the OIG identifies three groups of people that shouldn't be eligible:

- Incarcerated beneficiaries
- Unlawfully present beneficiaries
- Deceased beneficiaries

The OIG also mentions 'Identify and recover improper payments' as one of the top challenges for Medicaid in 2014 (OIG US DHHS 2014b). That improper payment rate was 5.8% in 2013, and that "payments made on behalf of individuals who should have not been enrolled in the program were the main source of error."

3 THEORETICAL FRAMEWORK

A structured literature review is performed to identify the concepts related to Medicaid enrollment, health insurance fraud and the detection of these fraud types. By following a structured approach the quality of the work can be safeguarded. We follow the approach described by (Webster and Watson 2002). Section 3.1 will explain the methodology, whereas in section 3.2 the literature on Medicaid enrollment process will be discussed. Section 3.3 describes the results in the field of health insurance fraud. Section 3.4 focusses on the use of machine learning for the detection of fraud. Conclusions will be drawn in 3.5.

3.1 Methodology

(Webster and Watson 2002) describe the concept matrix as one of the approaches to a literature study. In this review this matrix is used to identify fraud types, and generate an overview of which authors discuss which types of health insurance fraud. To get to know the current state of literature on Medicaid enrollment processes the selected articles have been read, and the relevant concepts found in the papers are discussed.

For the general identification of papers, a five-step approach has been taken:

1. Identify the set of keywords
2. Refine keywords based on result set
3. Filter results by title, filter out non-supported languages and clearly non-related articles
4. Filter results by reading abstracts
5. Identify concepts by reading resulting articles in more detail

The search for work related to Medicaid enrollment (Q1) the query in Figure 3 is used. The aim is to find literature related to the Medicaid application process, which is an enrollment process. After executing the query, the result set included a lot of irrelevant results which contained the keyword 'enrollees'. Therefore articles containing that keyword in their title are not included.

```
Q1: TITLE ( Medicaid AND (eligible OR eligibility OR application OR enroll*) AND NOT enrollees
```

FIGURE 3 - SEARCH QUERY TO FIND MEDICAID ENROLLMENT LITERATURE

Due to the relation between health care fraud, and ways of detecting it, it is not inconceivable that some papers describe concepts related to both topics. One can imagine an article

introducing some type of health insurance fraud, and subsequently describing an approach how to detect this type of fraud using a specific data mining approach.

Both concepts can be captured in one search query, which is shown in Figure 4 (Q2). A paper discussing ways to detect fraud is likely to use the word fraud in its title, as so is a paper purely describing health fraud types. Fraud related to or the health insurance programs, or the Medicaid program specifically is interesting for this research. The terms "health", "healthcare", "medical insurance" cover the whole spectrum of potential interesting articles. Even though the focus of this research is on enrollment fraud, searching for fraud in a broader scope helps to put this type of fraud in the right perspective. The broad search provides insight in how different types of fraud might relate together.

The combination of eligibility and the healthcare related topics results in enormous result sets, containing a variety of unrelated topics. In the other way, the combination of fraud and eligibility returns a result set that has to be narrowed down with one of the healthcare related keywords that is in the query below already. Thus the query in Figure 4 covers also the topic of eligibility in relation to fraud.

```
Q2: TITLE ( ( Medicaid  OR  "health" OR "healthcare"  OR  "medical insurance" )  AND  fraud )
```

FIGURE 4 - SEARCH QUERY USED FOR THIS LITERATURE REVIEW

Two scientific search engines are used to find articles: Scopus and Web of Science. Both have a good coverage of technical articles (related to data mining for fraud detection for example), as well as medical coverage. Web of Science also searches through PubMed, a primary database containing citations of biomedical literature. Only articles written in English will be included in the search. Furthermore, citations and patents are excluded. Only journal articles and scientific papers are included. In the case of enrollment only articles published after 1994 are taken into account, in order to not have to go through outdated articles.

Exclusion criteria are as follows for Q1. Articles have been selected if they were about the enrollment process and characteristics of that process. A lot of articles discussed the influence of Medicaid enrollment on access to healthcare, or on the treatment of specific diseases. These articles are not relevant, since our aim is to identify articles that might explain why might not enroll for Medicaid and what aspects of the enrollment process have been discussed in literature.

For Q2 the following exclusion criteria are used. Results were excluded from the initial listings if the articles discussed only specific fraud cases (e.g. an analysis of a fraud case that was in the news). Even though filters were in place, some results were not in English and had thus to be removed. Also, news articles have been removed. The included articles discuss types of healthcare fraud as a phenomenon, or are scientific publications about the detection of such fraud types using data mining methods.

Figure 5 shows that some papers might belong only to one of the two topics, while other papers might touch on both fraud descriptions and methods for detecting the fraud.

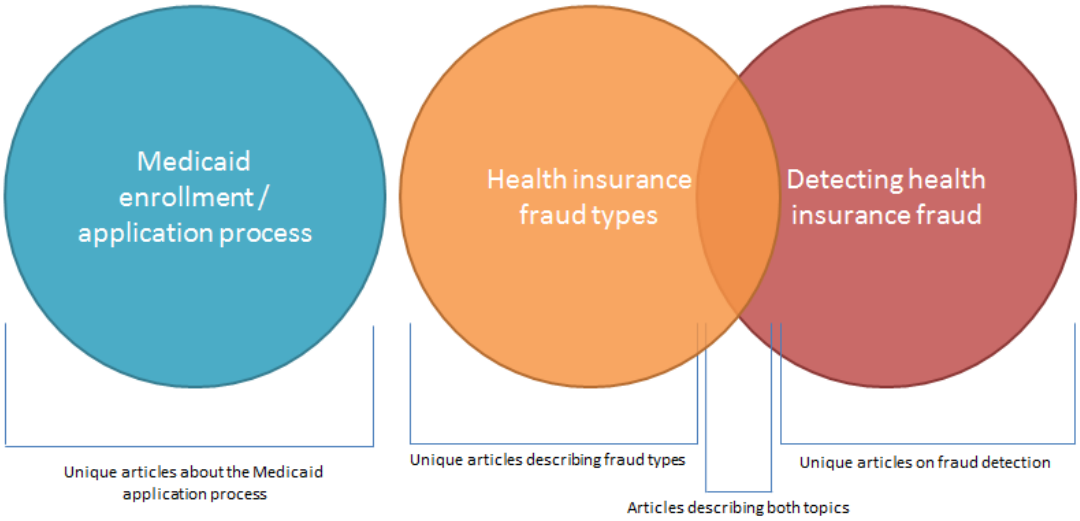


FIGURE 5 - LITERATURE STUDY TOPICS

Q1 resulted in over 600 search results, which after filtering out duplicated, and assessing them on solely the title resulted in a list of 114 articles. Of those 114 articles the abstract is assessed, resulting in a final set of 12 papers that were really about the topic we are researching.

Q2 resulted in a set of 252 unique articles matching the query. After filtering on abstract and title the resulting set yields 69 articles. The selected articles can be categorized into two different categories. They can describe fraud related to the healthcare system, and they can describe computerized fraud detection methods. The 69 articles have been categorized into these two categories. 50 Articles are related to healthcare fraud in general, while 19 articles are focused on topics like data mining in relation to health care fraud. After a more thorough review of the 50 selected articles on fraud types, 27 articles have been selected definitively.

3.2 The Medicaid application process

There is little literature available in the area of the Medicaid enrollment process from a more technical perspective. Most search results, which were deemed irrelevant for this research, describe the enrollment policies. A small number of authors however examined the enrollment process itself., with the majority of papers focusing on the enrollment of children into Medicaid/CHIP. (Hansen et al. 2011) examined the readability of Spanish Medicaid application forms. They conclude that most application forms use small font sizes and lack adequate white space. Furthermore, the language used is too complex, and they recommend following low-literacy guidelines to improve the accessibility for eligible Spanish speaking families.

When a citizen applies for Medicaid the citizen needs to provide information about his personal situation and his income. (Lessner 2006) describes the implications when more years of evidence are required to assess eligibility. For long-term care local Medicaid agencies need to assess eligibility information for several years. Lessner states that it is harder for families to gather all required information if the number of years they need to provide information on increases, but also this will lead to increased processing times at local Medicaid offices.

The implications of another documentation requirement, namely citizenship documentation, is discussed by (Hatch et al. 2014). Both the extended income requirements and the citizen documentation requirement are introduced in 2005, as part of the 'Deficit Reduction Act' (DRA). The researchers focused on the enrollment of children in to Medicaid in the state of Oregon. They found that children who were denied for Medicaid because they couldn't provide citizenship documents were likely to be U.S. citizens. Also they were medically and socially more vulnerable than their peers, had unmet health care needs and had gaps in their health insurance coverages. The authors mention that thanks to the Children's Health Insurance Program Reauthorization Act (CHIPRA, since 2009) and the 2010 Patient Protection and Affordable Care Act (PPACA), states now have the more options to confirm citizenship using electronic data-matching technologies instead of requiring physical documentation. This might reduce the number of people wrongly deemed ineligible. (Sommers 2010) mentions that enrolling in Medicaid and CHIP has become more difficult since the requirement for citizenship documentation is introduced. In another paper the same author states that the introduction of the DFA in 2005 led to a net loss of 600 million dollar because of increased administrative spending for states and also the increased difficulty for applicants to comply.

Another finding mentioned by (Hatch et al. 2014)is that there are no significant differences because of race, nationality or parental language when it comes to a right eligibility decision. (Kenney et al. 2012) also discussed the issue of language, and finds that parents not being a citizen and being unable to speak English is associated with lower Medicaid participation. It is suggested that this is possibly due to problems navigating enrollment systems or a lack of knowledge about the program.

Simplification of the enrollment process is an important topic. (Priest n.d.) describes the application for CHIP in California in 1999: at that time 443 items of information had to be provided, and a college reading level was needed to understand the application form. He also mentions that verification of income can be hard for people who get their income paid in cash, or who do not want to ask their employers to write letters of income verification. Simplification si also mentioned as an important factor in order to reduce the percentage of people misinformed about Medicaid enrollment, research by (Stuber and Bradley 2005) shows.

There is high interest in enrolling children in Medicaid, shows research by (Kenney et al. 2015). Many families however perceived difficulties with enrolling their children in the program. 5.4 Million children in the U.S. are uninsured, and 68.4% (3.693.600) of them are eligible for Medicaid, but not enrolled. The authors mention some children cannot enroll because of their immigration status, and for others their family income exceeds the eligibility thresholds. The majority of low-income parents with uninsured children that had experience with Medicaid/CHIP enrollment (60.2%) believes enrolling their child is going to be difficult. Of Parents without enrollment experience 43% believes it will be difficult to enroll their child.. The authors also looked into language, and found that children in immigrant families and non-English-speaking Hispanic families are less families with the possibilities, and are more confused about eligibility requirements. (Avruch et al. 1998) drew similar conclusions many years ago, the author also found that low income families may find enrolling their children in Medicaid difficult.

The importance of being able to enroll in person, next to for example applying by mail or online, is denoted by (Allison 2003): there may be a substantial number of CHIP eligible children at risk when they (or their parents) cannot apply at a local welfare office. The authors furthermore discuss the importance of good computerized eligibility systems. Families that are enrolled in multiple social programs, such as Medicaid, food stamps or child care assistance, might have to provide the same information several times a year for each of the individual programs. Case

workers might be able to align the dates, in order to align annual reviews for each of the programs. Computerized systems need to support such tasks.

Citizens do not specifically have to apply for Medicaid/CHIP in order to enroll. (DeLeire et al. 2012) describes the advantages 'auto-enrollment' can bring. The state of Wisconsin applied this concept and enrolled 44.000 individuals using an auto-enrollment system. This system involves electronic database matching, which applies program eligibility criteria to existing citizen data that is already yin state databases, and automatically converts eligible persons to coverage. The authors state that automatic enrollment takes away the application burden from potential eligible citizens. However, auto-enrollment seems an ineffective strategy for Medicaid eligible populations who are subject to premium payments.

3.3 Fraud types described in literature

In this subsection we will discuss the 24 fraud types identified in literature. Appendix 1 provides the complete table linking all identified literature to the identified methods of healthcare fraud. A summary of the article counts for each fraud type can be found in Figure 6Figure 7.

Kickback schemes

One of the most discussed types of fraud is fraud involving kickbacks. Kickbacks exist in different forms. For example, pharmacists can fill a prescription with a specific brand of medicines instead of another that yields a bonus from the pharmaceutical company (Rabecs 2006). Beyond financial implications, this might also be detrimental to the patient's health. Physicians themselves can fraudulently write prescriptions for money, essentially a kickback from the downstream illegal sale of these drugs (Morris 2009). (Bennett and Medearis 2003) point to the importance of complying with kickback legislation, and states that deals that seem too good to be true can be illegal.

Self-referral

(Rashidian et al. 2012) defines self-referrals as follows: *referring the patients to a clinic, diagnostic service, hospital etc. with which the referring physician has a financial relationship*. This might involve a kickback scheme if the referred-to party pays on a referral base to the physician, but also other financial relationships are conceivable.

Doctor shopping

If bribing a doctor does not work, a drug-seeking person might want to look for another doctor who wants to provide the desired prescriptions. One patient can register himself at multiple doctors, to get prescriptions multiple times, or can cherry-pick a doctor that is able (be it by means of a kickback) to fulfill the patient's regulatory needs. The latter is called 'doctor shopping'. (Carlson 2013) refers to a study by the Government Accountability Office that found out that in 2011 there were about 600 patients in the Medicare program that got prescriptions from more than 20 doctors each.

Identity fraud

Identity fraud may happen where an uninsured individual assumes the identity of a person with insurance coverages to obtain services or to hide a certain illness (Plomp and Grijpink 2011). They mention that in the end the personal health record of the person 'lending' their identity might be affected, since this will contain health information that is not related to the actual identities' owner. Identity theft might also happen without the owner of the identity knowing. (Dube 2011) mentions identity theft conducted by foreign gangs, that have scammed federal authorities for millions of dollars. The latter might be called identity theft, while if all actors are aware of the scam, it might be called abuse of an insurance card.

Fraud by pharmaceutical companies

(Sparrow 2008) describes pharmaceutical abuses beyond the kickbacks schemes are mentioned above. Specifically, off-label promotion of drugs involves the marketing of drugs for uses which are not approved by the Food and Drug Administration. Illegal price manipulation in collusion with downstream data providers or other pharmaceutical companies has been shown on multiple occasions.

Price manipulation of devices and services

Providers of medical equipment or health services can manipulate prices for certain groups of clients (Sparrow 2008). If they know Medicaid will pay varying rates for services, the may increases prices directly. Or, they may move across the street to the next zip code from which they can bill a higher rate.

Improper coding and upcoding

Improper coding (sometimes called upcoding) is another example of the most discussed fraud topics. This is sometimes called upcoding. Upcoding means *billing for a more expensive service*

or procedure than the one performed (Agrawal et al. 2013). However, improper coding might also happen because of an administrative error. (Agrawal et al. 2013) also mention 'incorrect coding' as an error, while upcoding might be a form of abuse.

Unbundling

Unbundling means creating separate claims for services or supplies that should be grouped together (Cady 2007). Unbundling can maybe be seen as a part of improper coding, but multiple authors mention unbundling as a separate form of fraud.

Submitting double bills

When it comes to submitting claims not only improper coding practices can be fraudulent, but also care providers can try to submit the same claim multiple times, in order to get paid two times for performing one action. (Byrd et al. 2013) name double-billing as *billing multiple times for the same service*. Automatic acceptance of claims is mostly done to improve processing speed, however Benzio (Benzio 2009) rightly mentions that for true efficiency not only speed matters. Tests for legitimacy are just as important.

Billing for services not provided

In the case of double billing at least care is provided to a patient. In the situation of billing for services not provided, claims are submitted for health care that has not been provided, or for medicines or medical devices that have not been delivered to the patient. This concept is also referred to as *phantom billing* (Rashidian et al. 2012). One of the examples mentioned of this is the one of health care providers that submit so many claims on one day that is not physically possible (or at least highly unlikable) to help so many patients (Stanton 2001) (Lubao 2008).

Related to this method of fraud is submitting false claims to the systems, and finding out how to pass a claim. Since claims are mostly automatically processed, finding the thresholds of the claim handling systems lead to submitting claims for services not provided, that do not trigger the monitoring systems (Morris 2009).

In order to submit these false claims information from patients is needed. There are several ways in which this information can be gathered. One is the use of information of patients that are no longer alive. There are more situations in which claims are submitted for care provided to people that actually shouldn't be seen as eligible anymore. In the Medicaid system people prisoners are

not eligible, as so are people who have been banished from the country and are prohibited from returning (Sparrow 2008).

Not only claims can be submitted stating care was provided to patients that passed away already, the other way around is also described in literature: claims that rely on the identifiers of deceased physicians (Morris 2009). Research on a population in Ontario (Canada) showed that for 1 per 3000 deaths providers submitted claims for medication more than one year after a patient deceased (Stelfox and Redelmeier 2003). Another example are ghost employees: employees on the health providers' payroll that do not actually exist (Brooks et al. 2012). (Evans and Porche 2005) show evidence of practices submitting bills for group sessions, while only one patient was treated.

Providing unnecessary care and maximizing care

It may also happen that more healthcare is provided than was actually needed to heal the patient; thus providing unnecessary care. Sometimes certificates are falsified (Rashidian et al. 2012) to show the medical necessity of certain actions in order to justify payments. (Morris 2009) also describes maximizing the number of claims, and claiming the most complex services, because of the so-called fee-for-service system that is in use. This model means that physicians get paid based on the services they provided; maximizing the number of services means maximizing their pay.

Other examples of unnecessary care include 'Rolling labs'; which are test provided by health care providers that temporary visit shopping centers or retirement houses. These are simple test, but billed as expensive tests to insurance programs. Furthermore sometimes care providers use unproven treatments, which might not work in the end and thus result in unnecessary care provided.

False negation cases

False negotiation cases are mentioned by (Doan 2011) as cases that arise from situation in which a health care provider makes false statements to induce the government ot enter into a contract for services or supplies. Sometimes this is also referred to as *frauds-in-the-inducement*.

Using the wrong diagnosis

Claims can be submitted claiming some kind of care has been provided based on a certain diagnosis. Fraud can also take place in the form of wrong diagnosis: a patient can get a certain

diagnosis while that is diagnosis is not actually true (Ogunbanjo and Knapp van Bogaert 2014). This type of fraud can be done to falsely prescribe certain medicines to a patient, for example. (Van Der Spoel et al. 2013) tried to predict cash flows in hospitals, using process mining methods, in processes that include human errors. Using the wrong diagnosis, but also 'providing unnecessary care', are seen as 'human noise'. This is one of the three factors that influences the power to predict cash flows most.

Billing for services rendered by unqualified personnel

Care can be provided by people who do not have the license to actually perform that kind of care (Byrd et al. 2013). An example of this is when an intern is providing a form of care, which he/she actually is unqualified for.

Lying about eligibility

Patients can lie about their situation when they visit a pharmacist or a physician. They can for example claim exemption from prescription charges, when they are not exempt (Rashidian et al. 2012) or they can misrepresent information about their dependents to get insurance coverage for them (Byrd et al. 2013).

Reverse false claim cases

False claims that are accepted by an insurance program result in receiving money from the insurer (in the case of Medicaid: receiving money from the government). The other way around is called *reverse false claims*. In these situations a care provider owes money to the government, and doesn't pay it back on time (Doan 2011).

Managed care fraud

Managed care is a part of the Medicaid program. This type of care is not provided on a fee-for-service basis. This provides new areas of fraud, as mentioned by (Sparrow 2008). This type of fraud includes denial of services to patients, providing substandard care and creating logistical and/or administrative obstacles for patients in order to receive the care they need.

Waiving co-payments

In some cases co-payments might be in place (Freeman and Loavenbruck 2001). In the Medicaid case this might happen if the income of a family exceeds a certain threshold. In such situations a deductible has to be paid. Some health care providers however sometimes waive co-pays or deductibles, resulting in higher cost for the government.

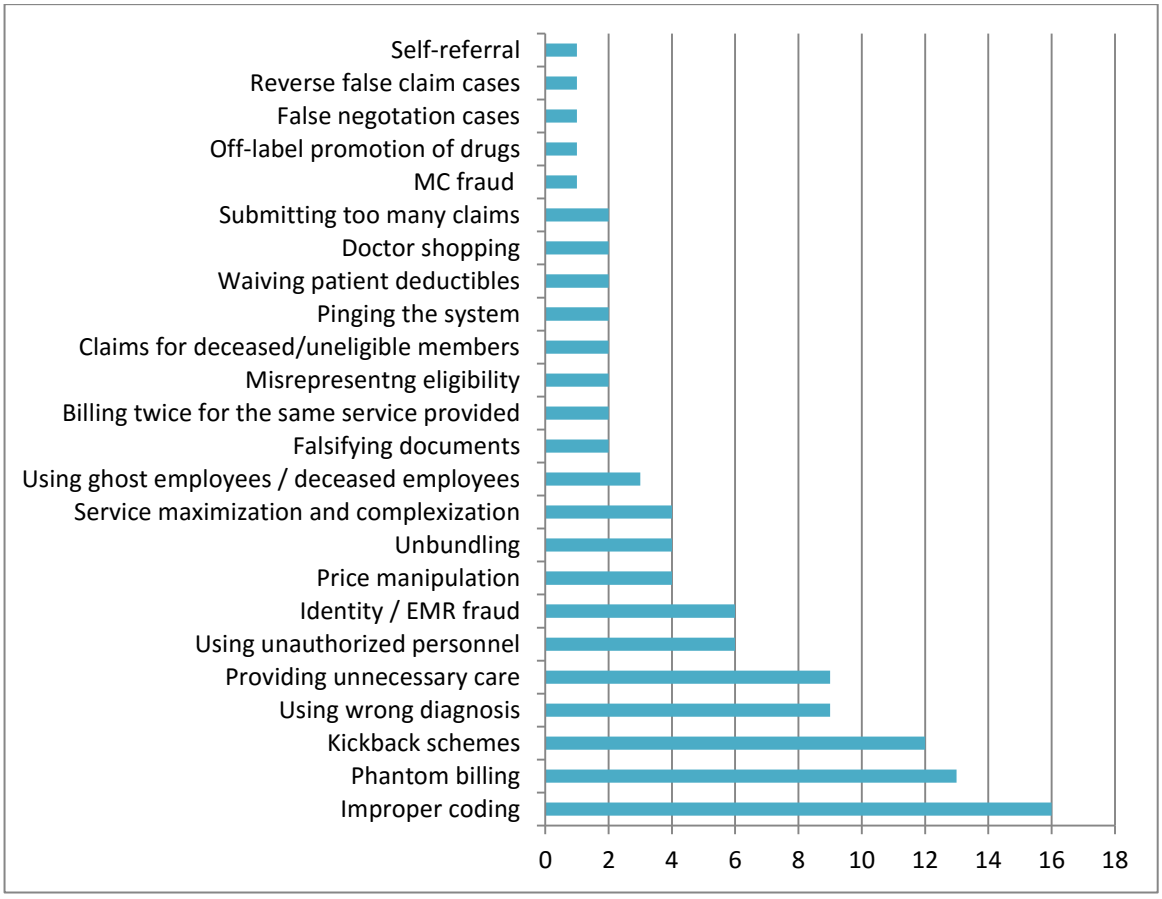


FIGURE 6 - INCIDENCE OF HEALTH INSURANCE FRAUD TYPES IN LITERATURE

3.4 Data mining methods

The search for papers about the detection of fraud using data mining resulted in a set of 22 papers discussing the topic. The papers have been classified in two categories: the data mining method used, and the purpose of the data mining method: which type of fraud did they aim to detect.

Some papers contain a literature review on data mining to detect fraud. However, since these overviews tend not to include all papers identified by our own literature search, a new overview has been created. This overview extends the existing literature review done by (Joudaki et al. 2014) with additional research in the field, creating one extensive overview of methods and their designated purposes consisting of 36 articles.

This part of the literature review shows that all data mining methods used are unsupervised data mining methods. Clustering, in different forms, is the most used. Other popular methods are outlier detection and classification. Classification is however closely linked to decision trees, therefore these two methods will be discussed together. In this sub section we start with describing the most popular fraud detection methods, and how they have been used in the literature. Figure 7 gives an overview of the number of articles in which each fraud detection method is discussed.

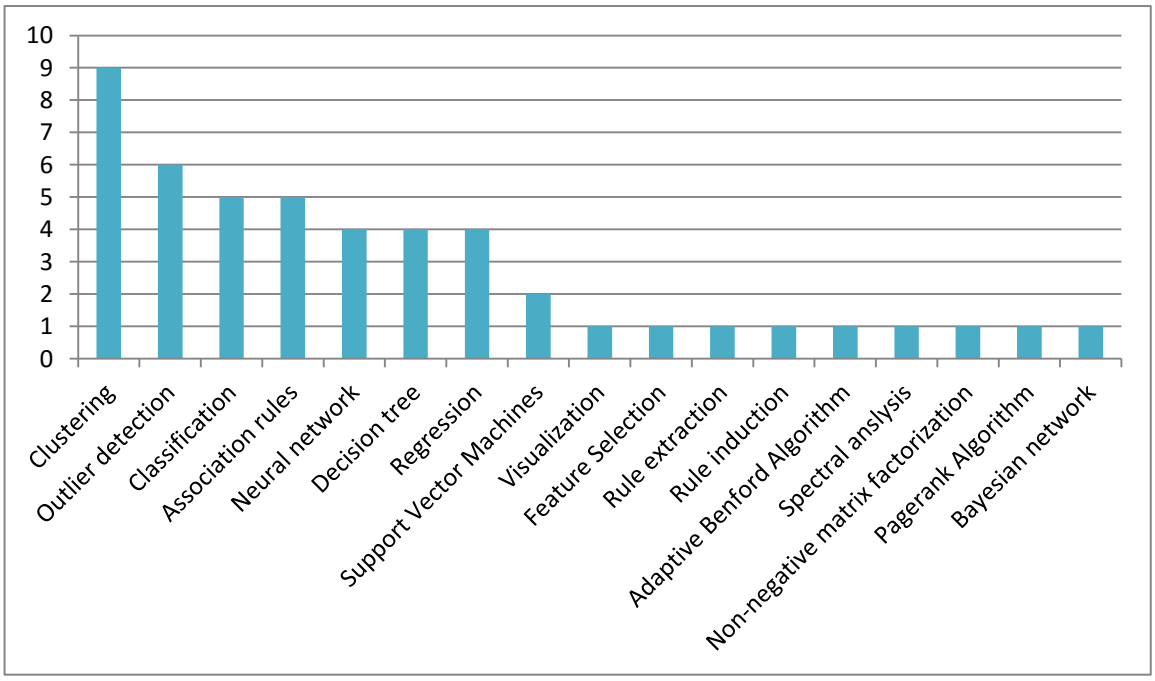


FIGURE 7 - FRAUD DETECTION METHODS IN LITERATURE

3.4.1 Clustering

The concept of clustering data is the method most used for the detection of health care data. As (Liu and Vasarhelyi 2013) explain, clustering in an unsupervised data mining method. The advantage of unsupervised methods over supervised methods is that there is no need for training data. Supervised algorithms require a set of data to be labeled (e.g. a training set of claims marked as fraudulent and non-fraudulent). The authors used geo-location information of patients and health care providers to find out people that get care from providers that are on a long distance. This might indicate fraudulent behavior. Their model is capable of both detecting long distance claims, claims with high payment amounts, as well as a detecting claims with relatively long distances and large amounts of payment. They don't provide information about the accuracy of the model, to provide insights to what extent this approach leads to finding false-

positives. (Ekina et al. 2013) developed a co-clustering method, looking at the memberships of both providers and beneficiaries (eligible people) to unusual groups. They state that using their approach it would be possible in the future to predict the possible behavior of new providers and beneficiaries, based on their characteristics.

Clustering can also be part of a more extensive fraud detection model. (Tang et al. 2011) used clustering as one of the preparation steps, to finally perform the detection of fraud using outlier detection. Their 'cluster builder' was used to examine data and label it based on certain criteria (frequency of drug prescriptions, or the similarity of temporal prescription sequences). Since these criteria are manually entered, this method is not an unsupervised clustering method. (Musal 2010) used clustering to flag health care providers that are potential frauds. They also combined clustering with several other methods, including regression analysis and various descriptive statistical models. Clustering was used to group zip code areas in combination with socioeconomic factors.

(Williams 1999) defined a clustering approach to highlight importance or interesting areas in large datasets, called hot spot data mining. However, a general clustering algorithm (k-means based clustering) is used. They don't have any claims about the usefulness of their method however. The approach described is more used to explore large data sets, instead of really identifying frauds in claims data for example.

A comparison of two clustering methods, SAS EM and CLUTO, was done by (Peng et al. 2006). SAS EM is an commercial enterprise mining solution developed by the company SAS, while CLUTO is a software package developed by the University of Minnesota. Their conclusion is that CLUTO is faster, while SAS EM provides more useful clusters. They suggest that clustering is mainly useful for labeling data, which in turn can serve as input for other algorithms. Clustering is more or less seen as a data preprocessing step. (Zhu et al. 2011) propose a solution using non-negative matrix factorization. They create a matrix with on one dimension patients, and on the other dimension their medicines used. They authors are unsure however if non-negative matrix factorization performs better than other clustering methods.

3.4.2 Outlier detection

The second most described method is outlier detection. (Capelleveen 2013) describes the use of outlier detection to find fraudulent dental claims. The precision rate of the methodology is estimated at 71%. However, the author concludes that outlier based predictors are not likely to

succeed as fraud classification technology. One of the conclusions is that fraud detection predictors should be seen within a program of multiple fraud detection methodologies. (Ngufor and Wojtusiak 2013) also looked into medical claims data. They place an interesting note on the use of outlier detection. According to the authors outlier detection methods are not suitable to handle *drifts* in data over time. Drifts are changes in data, and in the case of labeling: when the same example receives different labels at two different times. To clarify: what is considered an appropriate payment for a service at one point in time, may not be appropriate five years later (because prices rise over time). They propose however a modification to their outlier detection algorithms in order to handle this case, and are able to show an increase in detected outliers using the modified method.

Detecting prescription fraud in Turkey using outlier detection is described in a paper by (Aral et al. 2012). They propose a model which has a true positive rate of 77.4% and a false positive rate of only 6%. Based on the claims they created index matrices of combinations of claim characteristics. These include type of medicine and {age, sex, diagnosis, medicine} and diagnosis and cost and try to find outliers in these datasets. (Shan et al. 2009) tried to detect fraudulent optometrists claims in Australia by finding claims with unusual distances (between patient and clinic). They used outlier detection to do this. Their Local Outlier Factor was 'significantly better than randomly sampling, and at least as good as crude univariate method.' They propose to integrate this detection method in an application that includes multiple fraud detection methods.

Such an integrative system is proposed by (Major and Riedinger 2002), which combined behavioral heuristics with outlier detection and selection rules. Outlier detection was used to detect care providers that did not fit the pattern when compared with their peers.

Most detection methods are focused on finding claims fraud. (Konijn and Kowalczyk 2011) state that when searching for fraud it is more important to look on the level of single patients, pharmacies of GP's. They created a two-layer outlier detection approach. The first layer is an analysis on claim level, and the measurements of the first level are aggregated to detect outliers in the 'higher level' entities mentioned before.

3.4.3 Classification & Decision trees

Clustering and classification are easily mixed up. Both concepts apply to objects in a data set, for example claims. Using classification a claim will be assigned a label, based on a rule that assign labels to points. This is a form of supervised learning. The assignment of these classes can be

done using decision trees. On the other hand, when clustering claims, there are no labels. The clusters are created based on how close the claims are to each other. This approach is used to identify structure in data, and it is a type of unsupervised learning (MIT OpenCourseWare 2008).

Even though classification is introduced as a form of supervised learning, some authors have tried to perform unsupervised classification. (Ngufor and Wojtusiak 2013) describe an approach to label a dataset, and add 'normal' or 'abnormal' as a label to claims based on the amount paid. Another method of classification is executed by (Williams and Huang 1997), they used a C4.5 algorithm. This algorithm is able to generate decision trees and rule sets. These decision trees are in turn used for classifying a data set. They used the method on a case study containing Medicare claims, looking at claim amounts. Identified (potential fraudulent) segments need to be assessed by humans after classification in order to identify if this is a real fraud case.

Several classification methods can also be used together, as shown by (Phua et al. 2004). They used both a naïve Bayesian, C4.5 and Backpropagation algorithm. The authors got their inspiration from the movie Minority Report, in which crimes are predicted by *Precogs* (Precognitive Elements). In the movie these are mutant humans, while in the solution proposed by Phua et al. each precog is a classifier trained with one of the three classification algorithms. The authors describe an approach for selecting the best set of classifiers given a dataset. (Bonchi et al. 1999) also described the concept of combining classifiers. They describe conjunction (taking all cases marked as fraudulent by all classifiers used) and disjunction (taking all cases market as fraudulent by only one of the classifiers). Furthermore they explain the voting concept, as also used by Phua et al. A last method of combining classifiers described is taking all cases that have been selected as fraudulent by at least 3 of the 4 classifiers used. (Yang 2003) developed two algorithms for the classification of claims. These algorithms serve as a step in a more complex fraud detection system.

3.5 Conclusion

In this section literature regarding Medicaid application processes has been studies, as well as literature on health care fraud. Based on the identified articles about the application process some related factors to this process can be identified. These are:

- Language,
  Evidence requirements (such as income or citizenship)
- Method of application (paper forms, computer systems or local welfare offices)

Also fraud types in the field of Medicaid have been identified. The fraud types can be classified into several categories. Some types of fraud are related to claims submitted to the healthcare insurance system (in this case Medicaid). This is fraud committed by the health care provider, since they are the parties that submit claims to Medicaid. Other fraud involves kickbacks, and requires not only involvement of the health care provider (which is the party receiving money), but requires also involvement of a party that is providing the financial benefit. Patients can also fraud the system on their own, while their health care provider doesn't have to be aware of this. This can be the case when committing identity fraud, or when performing *doctor shopping*.

As the literature analysis shows there are many ways fraud can be committed. It happens at all stages of the process, and multiple actors tend to be involved. Fraud can both be committed by patients, by health care providers and by a combination of them. Furthermore, using identity fraud it is even possible for fraud to be committed without the care providers even knowing.

There is a lot of literature describing fraud related to claims. For example: phantom billing, providing fraudulent information on the claims, and sending in too many claims. A topic less discussed is committed at the enrollment stage. The only identified topics related to this practice are *identity fraud* and *misrepresenting eligibility*.

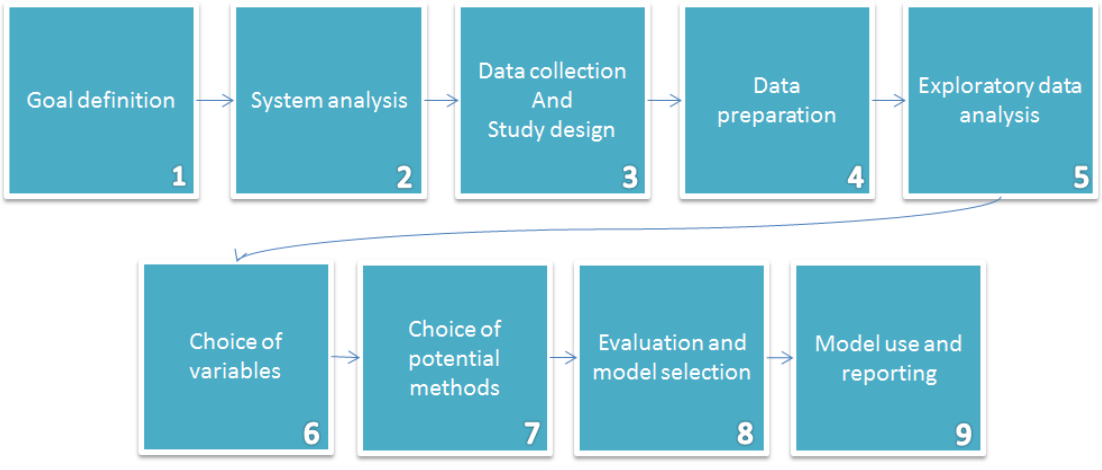
Researchers describe various ways of detecting fraud. The topic of fraud committed at the enrollment stage is however only not that frequent discussed.. Only (Suleiman et al. 2013) covers the topic. That specific paper covers only the validation of user input at the moment when someone applies for Medicaid eligibility.

4 RESEARCH METHOD

This chapter introduces the research method applied in this research. It starts with a justification for the chosen method in section 4.1, followed by a description of the steps the chosen method consists of in section 4.2.

4.1 Selection of the research method

In this research data analysis plays an important role. The aim is to optimize a process by looking at the historical data generated during that process. Based on the historic data we want to predict the processing time of future applications. This introduces the topic of *predictive analytics*. Predictive analytics are defined by (Shmueli and Koppius 2010) as *the empirical methods that generate data predictions as well as methods for assessing predictive power*. According to the authors predictive analytics can play an important role when it comes to explanatory modeling in theory building and testing. The authors define an eight step model for research projects in the field of predictive analytics. The model is extended by (Van der Spoel and Amrit 2015) with an additional step after the first step. The paper introduces system analysis as the additional step



4.2 Explanation of the research method

The selected research method is a nine-stage method. The individual steps will be explained below:

1. Goal definition In this step it is made clear what exactly needs predicting.
2. System analysis: Determine whether the system at hand is complex or simple.
3. Data collection and study design: In this step the actual data collection takes place. Four elements are important:
    a. Data collection instrument: ideally the same variables are in the training data set, as well as in the dataset that is used for prediction.
    b. Sample size: the size of the training data set
    c. Data dimension: explain which variables are included in the dataset
    d. Hierarchical designs: Increasing group sizes are needed when the number of groups increases
4. Data preparation: In this step the data is prepared for the analysis. Several factors have to be taken into account
    a. How to handle missing values
    b. Data partitioning: creating a training set, and a dataset to evaluate the trained model
5. Exploratory data analysis: During this stage the initial dataset is analyzed, in order to detect and remove outliers. This can be done using visualization, and it might lead to dimension reduction: narrowing down on the set of input variables.
6. Choice of variables: for the predictive models variables need to be selected. These variables need to be available at the time of prediction and their measurement qualities need to be sufficient.
7. Choice of potential methods: In this step the potential methods are selected.
8. Evaluation and model selection: the selected models are applied on a dataset ,and the best model is selected.
9. Model use and reporting: this is the final step of the analysis. Report the results of using the selected model.

Table 3 depicts how the individual elements of the methodology relate to the chapters in this thesis.

Step of the research method	Chapter
Goal definition	Chapter 1
System analysis	Chapter 5.1
Data collection and study design	Chapter 5.2
Data preparation	Chapter 5.2
Exploratory data analysis	Chapter 6

Choice of variables	Chapter 5.2
Choice of potential methods	Chapter 5.3
Evaluation and model selection	Chapter 6.3
Model use and reporting	Chapter 6.3

TABLE 3 - MAPPING OF CHAPTERS TO THE RESEARCH METHOD

5 ARTIFACT DESCRIPTION

In this section the designed artifact will be introduced. The artifact is a combination of both a data model and a data mining methodology to analyze the application process. The section starts with a description of the current systems and data structure in section 5.1. From this current structure we can derive what useful information is available, resulting in a new data model, described in section 5.2. This will provide an answer to SQ2: What data is needed to analyze the Medicaid application process? Based on the designed dataset suitable data mining algorithms can be selected which will be applied, which is done in 5.3.

5.1 Current systems

There are multiple systems in use for managing Medicaid in South Carolina. Separate systems for claim handling and eligibility determination exist. And within the eligibility determination there are multiple systems. Currently there are two applications used for the determination of eligibility of Medicaid applicants: MEDS and Cúram. MEDS stands for Medicaid Eligibility Determination System. This is a mainframe system, that is in use for a long time already. CTIT is working towards an 'end-of-life' moment for this system, and migrating all its functionality and data to Cúram. In the past MEDS was the only system in use for the eligibility decisions. Cúram is a software package from IBM aimed at supporting governments in their health and human services processes.

The decision about in which system the information of a person is stored is based on if a person is in a MAGI eligible group. MAGI stands for Modified Adjusted Gross Income. MEDS stores all non-MAGI enrollees, while Cúram is responsible for all MAGI eligible people. The difference between MAGI and non-MAGI eligible people is shown in Table 4.

MAGI Eligibility Groups	Non-MAGI Eligibility Groups
Pregnant women	SSI - SSI recipients - State Supplement only
Infants and children under 19 years old	SSI-related Medically needy - Aged - Blind - Disabled
New adults (childless adults, which include individuals: - Are not pregnant - Are age 19 – 64 (19&20 living alone) - Do not have Medicare - Could be certified disabled but do not have Medicare yet	ADC-related Medically Needy - Under 21 years old - For reasons other than income, would meet the eligibility requirements of the aid to Dependent Children program as it existed on the 16th day of July 1996.
Parents/Caretaker relatives	Foster Care (IV-E or Non-IV-E)
19&20 year olds living with parents	Medicaid Buy-In for working people with disabilities (Basic Group and Medical Improvement Group)
Family Planning Benefit Program	Medicaid Cancer Treatment Program
Childs in Foster Care (Chaffee)	Individual under 26 years who was in Foster Care and in receipt of Medicaid on 18th birthday
	Resident of Home for Adults run by LDSS, OMH Residential Care Centers/Community Residences
	Medicare Saving Program
	Individuals applying for Cobra continuation of premium payments
	Medicaid continuation of Pickle, Widow and Widowers and DAC eligible individuals

TABLE 4 - MAGI VERSUS NON MAGI GROUPS (WEILL CORNELL CENTER FOR HUMAN RIGHTS 2015)

5.1.1 Cúram

Cúram is the name of an off-the-shelve product by IBM, that offers full lifecycle support for managing health and social programs from need to outcome (IBM 2015). The system is used to handle more and more eligibility decisions, to in the end be the successor of the MEDS system. In contrast to MEDS, which is a mainframe system, Cúram is a web-based system built in Java.

5.1.2 MMIS

MMIS is the Medicaid Management Information System. Both MEDS and Cúram applications are connected to MMIS, and provide information with MMIS. Where the first two systems are focused solely on handling eligibility decisions, MMIS is solely processing Medicaid claims. Furthermore, the system is used to generate reports to submit to several authorities that require periodical reports.

5.1.3 Interfaces

MEDS and Cúram are connected to several external data sources. These sources are called *interfaces*. These interfaces provide information for the validation of enrollment information. This includes income information, information about someone's imprisonment and information about enrollment in Medicaid programs. We will provide some examples of interfaces the system is connected to, to give an impression of this functionality:

- State Data Exchange (SDX): Receive data from the Social Security Administration for individuals receiving SSI payments (Supplementary Security Income). Data is exchanged on a weekly based. Individuals receiving SSI payments are also eligible for the Medical program.
- Coordination of Benefits (COB): A data exchange with the Centers for Medicare and Medicaid Services (CMS) to exchange information about enrollment in the Medicare program.
- Enumeration Verification System (EVS): A data exchange with the Social Security Administration for their South Carolina Retirement System (SRS): Receive income data from the South Carolina State Retirement System. The verification of primarily SSN and identification data.

5.1.4 Reporting

MMIS not only processes claims data, but is also used to generate reports for several instances. The image below shows a number of these reporting systems. We will give an overview of three reporting systems (which are also displayed in Figure 8).

- Truven: MMIS sends MEDS data to load into a Decision Support System (of which Truven is a vendor). From this system, reports may be run. The data is also provided for research from this system.
- Management Administrative Reporting Subsystem (MARS): This system generated federally mandated reports
- Expenditures (EXP): This is a monthly financial reporting of money spent on Medicaid services.

The relation between the systems explained above is depicted in Figure 8.

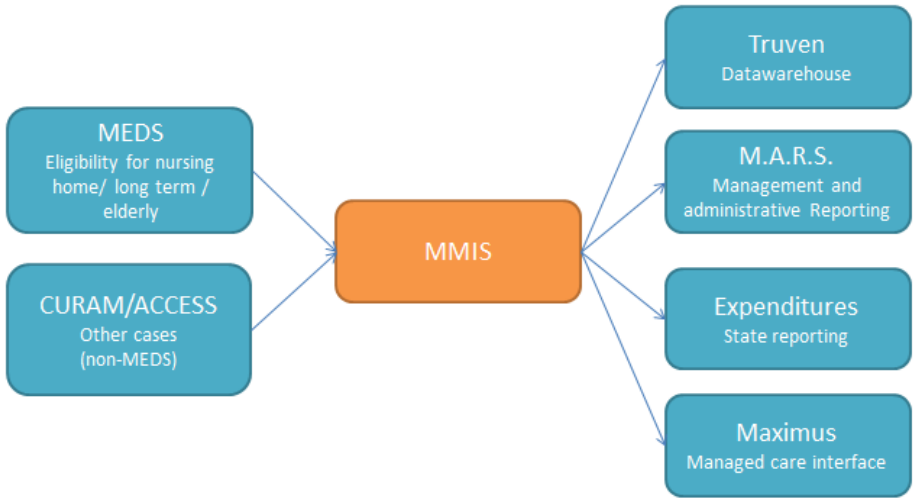


FIGURE 8 - MEDS, MMIS, CÚRAM RELATIONSHIP

5.1.5 From Cúram and MEDS to MMIS

Data is exchanged between Cúram, MEDS and MMIS. The image below shows the most important eligibility data that is transferred. This contains information about the person (called the Card data: the information that is on the physical Medicaid card of a recipient), the Budget Group (information about the circumstances the recipient lives in) and information about managed care plans (if applicable).

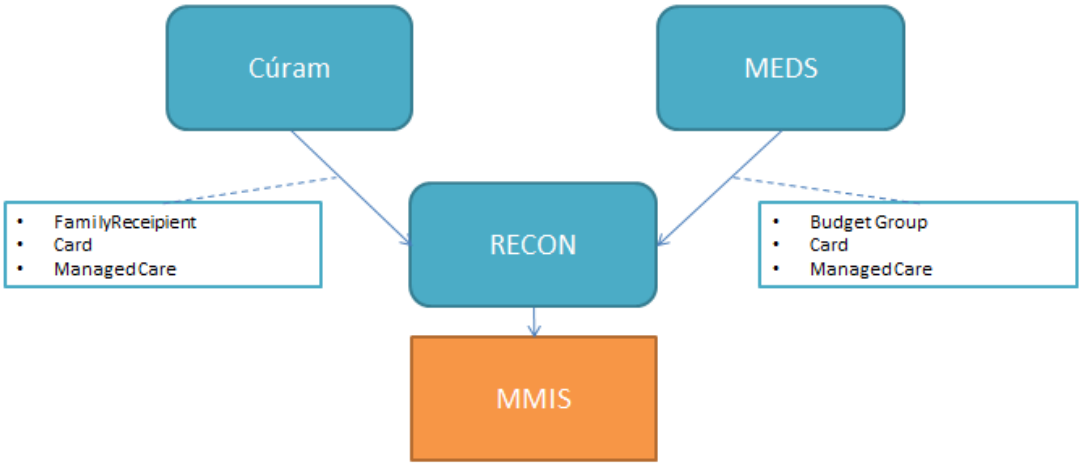


FIGURE 9 - FROM MEDS AND MMIS TO MMIS

Cúram and MEDS store eligibility data different. The data needs to be converted to a format that MMIS understands, so there is a conversion layer between MMIS and MEDS and Cúram. This conversion layer is a system called RECON. RECON is short for reconciliation. In this sub section we will discuss the way the eligibility information is stored in MEDS and Cúram, before it enters the RECON system. The relations between the systems are shown in Figure 9.

Both MEDS and Cúram consist of hundreds of tables. However, in both cases there is a 'core set' of tables that contains the most important information. For MEDS the structure is shown in Figure 10. For the determination of eligibility the following situation applies. An eligible person is an entity type Member. This Member is part of a Budget Group, which in turn is a part of a Household. The Household is the physical Household a person lives in. One person can be a member of several Budget Groups. The Period tables are keeping the changes in data over time. A Member can be pregnant for a certain time, and therefore be member of a specific Budget Group.

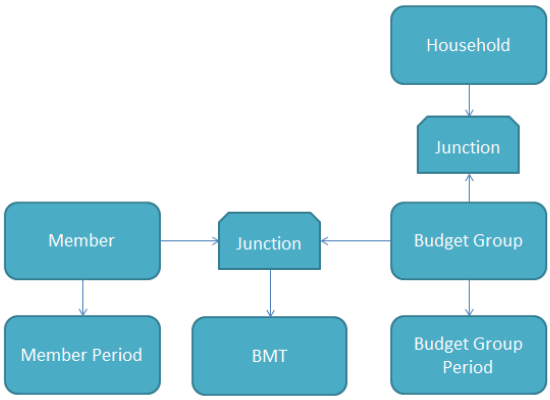


FIGURE 10 - BASIC MEDS DATASTRUCTURE

In Cúram a person receiving Medicaid/CHIP is called a 'concern role'. The concern role entity stores all information about the person. A person can be part of multiple 'cases'. At the same time multiple persons can be part of that same case. If there is a family for which a child goes into CHIP a case is created, containing the concern roles of the parents and the child. The information about the dates that the child is eligible for CHIP is stored as a 'product delivery'.

5.2 Selecting the data

After having clear what systems are in use, a choice has to be made what data to include in the analysis. Cúram is currently the only platform that allows both citizens and caseworkers to enter applications. Applications in MEDS can only be entered by caseworkers in a terminal screen. Furthermore, data in Cúram is more accessible. The system uses an Oracle database designed for reporting purposes that can be directly assessed. Exporting data from the mainframe systems involves more actions, because a specific reporting database as in Cúram does not exist.

The decision is made to focus on enrollment in Cúram. Not only because of the reasons mentioned before, but also because SCDHHS is slowly migrating all their application processes towards Cúram. SCDHHS is working towards an "end-of-life" situation for MEDS, and therefore results based on the new system are more valuable.

The process of selecting the right data out of the Cúram system will now be discussed. Figure 12 shows the process that is executed for eligibility determination. The process starts with a citizen applying for Medicaid or CHIP, or a citizen applying on behalf of another citizen applying for one of the two programs. There are several methods an application process can be initiated, which will be briefly discussed below.

Online

A citizen can apply for Medicaid or Chip using an online portal, which is called the 'Citizen portal'. This portal is a web application on which a citizen can do a 'quick scan' to see if he/she or his/her family is eligible for Medicaid (Figure 11). Three questions are being asked:

- What is your income?
- How many adults do live in your household?
- How many children under the age of 19 do live in your household?

After answering the questions the portal shows if the citizen seems to qualify for Medicaid. If so, he can fill out the online application forms. First, the user creates an account and using the account he goes through a multi-page application form. After filling out all forms the system will tell if the citizen needs to submit additional proof of statements he made (e.g. to verify a social security number or evidence of his income stated).

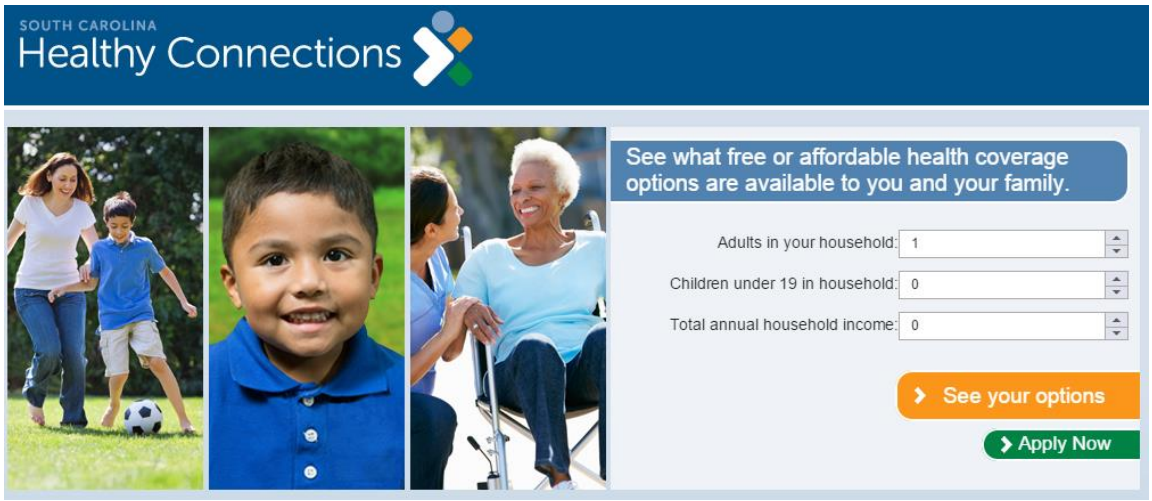


FIGURE 11 - CÚRAM ENROLLMENT HOMEPAGE

If no additional evidence has to be provided by the applicant(s) the system can process the application directly. This is called straight-through processing, or in terms of this specific application it goes by the name 'Streamline Medicaid'.

Finally, the system will show for which programs the people in the household may be eligible, based on the information provided. This is however a preliminary check: once the citizen decides to apply for a program, a more extensive ruleset will be executed. If the person is deemed eligible directly, it called a 'Streamlined Medicaid case'. In case some data cannot be validated by the rule engine, an employee of SCDHHS will look through the application and the employee will try to

get additional evidence and run the rule engine again. That will result in the final eligibility decision. The online eligibility determination process is shown in Figure 12.

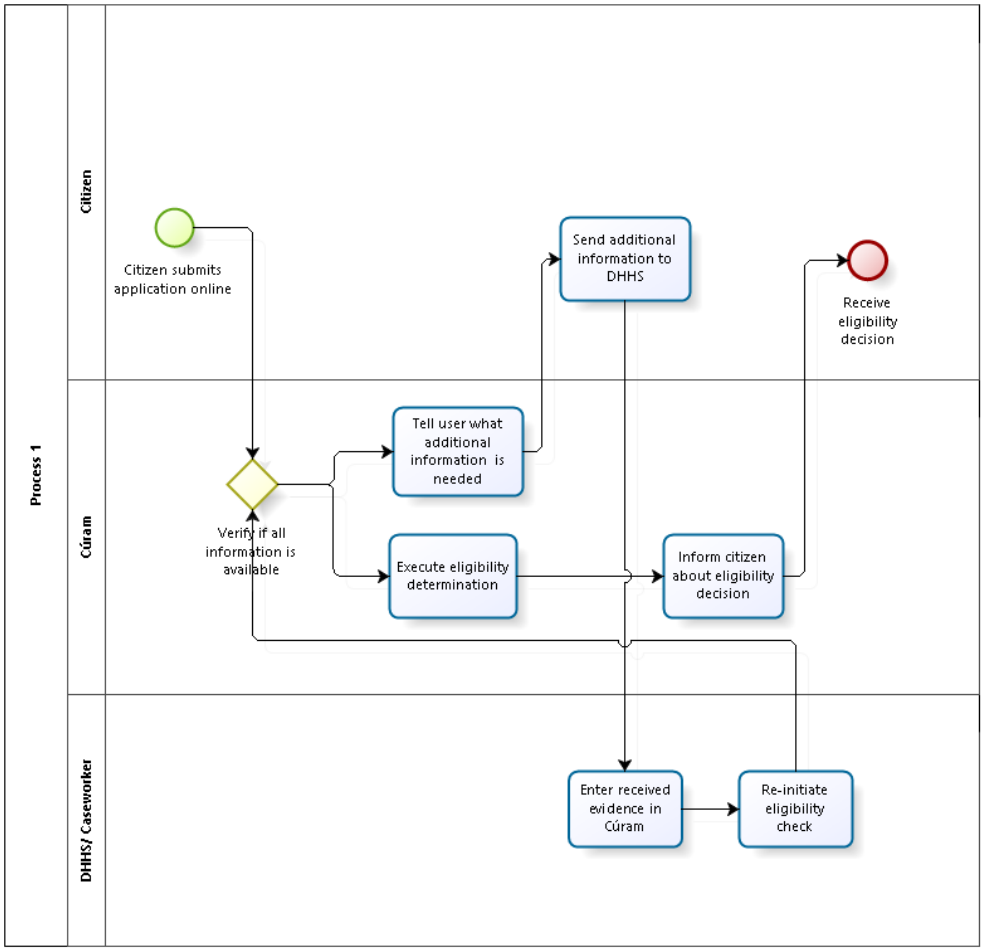


FIGURE 12 - ELIGIBILITY DECISION WORKFLOW

In person

A citizen can apply for Medicaid in person, for example at one of the offices of SCDHHS. An employee will create a new case, and fill out all information on behalf of the applicant. The caseworker will enter all this information in Cúram, so the same rulesets will be ran on the application as when a citizen would apply online.

By phone

This works basically the same as when one would apply in person. The main difference is that missing evidence cannot be handed to an employee directly, but has to be send by mail or delivered to a SCDHHS office by the applicant.

By mail

There are paper application forms available for Medicaid. A citizen can fill out such an application form and send it to SCDHHS. A case worker will then enter all information from the paper form in Cúram, using the same form as used in the 'In person' option.

The process looks like a straight-forward, and easy to understand process. The question rises however: why does one application takes just three days before the citizen knows he or she is eligible, and another application takes over 100 days before the citizen is informed of his eligibility decision. In order to find this out, a dataset has been created that contains characteristics of all applications processed. Cúram stores its information in hundreds of tables in an Oracle database. Since Cúram is not only used for initial eligibility determination, but also for yearly redeterminations not all data is useful for analyzing the initial application process. In order to find out what information is available, an overview of tables involved during each step of the process, and the information they contain about the application has been created. Table 5 shows the information gathered during this process.

Initiated by actor	Step in the process	Database tables used
Citizen	#1 Citizen submits the application online	ApplicationCase AppCaseEligiblityResult
Cúram	#2 Inform user about information needed	VerifiableDataItem
Cúram	#3 Execute eligibility determination	CaseHeader CaseParticipantRole ProgramAuthorisationData MMEligiblityDecisionDetails
SCDHHS/CS	#4 Enter received information in Cúram	EvidenceDescriptor
SCDHHS/CS	#5 Re-initiate eligibility check	Refers to #3

TABLE 5 - TABLES USED TO CREATE DATASET

Based on the information derived from the process and its related tables as explained above a dataset is created. This dataset contains a subset of all information as stored in the tables mentioned before. Information about each application is stored in one row in the new dataset. This dataset is created using a query in Oracle SQL Developer and exported to a .csv file. This .csv file is processed with PHP to do some minor tweaks to the data, which would otherwise make the SQL query way more complex. For example, the PHP script changes a string containing multiple

types of missing evidence to columns in the dataset containing the number missing items for each specific evidence type. Also all applications for which 'SYSTEM' had made a decision have been removed. The focus here is on 'real applications', and those SYSTEM applications are automatically processed applications (e.g. streamlined Medicaid applications, or applications imported from other systems.

The exact query can be found in Appendix B. For each application the following information is stored in the dataset. The specific selection of data elements has been made during a two stage process. First, the application process from a user and case worker perspective is analyzed. We went through the online application process, and looked at what information users have to enter that is not part of the automatic rule engine, and might thus be of influence on the processing speed. Furthermore, the potential items have been discussed with multiple IBM consultants and Business Analysts to see if they agreed on the selection, or had suggestions for other variables to include. According to all persons spoken missing evidence seems to be the main cause for delays in processing time. They did not mention other metrics that are currently in the system. Table 6 shows all elements that are included in the dataset.

Column name	Column description
NumberOfPeople	The number people that is in the application. One applicant can for example apply for his whole family. In case of a mom, dad and two children there would be four children in the application.
MethodOfApplication	An application can be entered in Cúram online (so by the citizen himself) or by a SCDHHS caseworker. The caseworker can in turn receive the application from the applicant in person, by phone or by mail.
NumberOfDays	The number of days between the moment the application was submitted (entered in Cúram) and the first eligibility decision has been made by the Cúram rule engine. This is calculated by subtracting the application date from the first decision date, since Cúram does not store this number by itself.
ProgramType	The program the main applicant got accepted or rejected for. If a family applies, it can happen that the parents are

	determined eligible for Medicaid, while the children are accepted in CHIP. Since the first decision is always made at the same time for all participants
NumberOfEvidence	The number of different types of evidence that are missing at the time the application was submitted, and had to be provided by the applicants.
EvSSNDetails	Missing evidence: the SSN could not be confirmed
EvCitizenStatus	Missing evidence: the citizenship of the applicant could not be confirmed
EvIncome	Missing evidence: the income could not be verified
EvExcludedIncome	Missing evidence:
EvGenderDetails	Missing evidence: the gender details provided could not be confirmed
EvIncarceration	Missing evidence: it could not be confirmed the applicant is alive
EvStateResidency	Missing evidence: it could not verified the applicant is a resident of the state of South Carolina
PaymentCategory	The payment category
FinalDecision	The eligibility decision that has been made for this person regarding the ProgramType as mentioned before. This can be either a 1 (=eligible) or a 0 (ineligible).
CaseworkerName	The caseworker that executed the final eligibility check.
CaseID	The ID of the case within Cúram. This helps to trace back specific cases throughout all Cúram tables.
PreferredLanguage	The preferred language as entered by the citizen
PreferredMethodOfCommunication	The preferred communication method as entered by the citizen
ApplicationMonth	The month the application is submitted
DecisionMonth	The month the eligibility decision is made

TABLE 6 - ELEMENTS OF THE CREATED DATASET

After the initial dataset was created the dataset was reviewed to see if there was a need for optimization of both the SQL query and the PHP code. It turned out not every row (meaning every application) had all data for each of the columns defined in the table above. Therefore it

was decided to create a version of the dataset which does not contain empty or NULL columns. This can happen in for example applications that not have been completed yet. These applications have no final eligibility decision yet, and might not have a case worker name assigned to them.

The created dataset contains 13038 rows. Each row represents a unique 'ApplicationCase' and has a unique 'CaseID'. For rows containing NULL at the PreferredLanuage of PreferredMethodOfCommunication fields the fields are manually filled. That means English is then selected as default language, for the communication method 'not selected' is entered. If these fields would've left empty, the rows would be ignored during data analysis, resulting in a significant smaller dataset.

Within the dataset there is some data that is known directly after an user submits an application, and there is information that becomes available after the decision is made. The items known for sure at moment of submission are the number of applicants, the number of missing evidence items, the method of applicant, the preferred language, the preferred communication method and the specific types of missing evidence.

In addition to the data gathered from the database each application has been labeled with a class, which identifies if the decision was made within 45 days or not. 45 Days is the general limit used by SCDHHS in which they try to come up with an eligibility decision (SC DHHS 2015). The dataset includes applications that took less than 101 days to process. Most applications that exceed this 100 day limit are entered in the system on incorrect dates, not in line with the majority of the applications and thus might be inaccuracies.

5.3 Selecting data mining methods

Having the dataset created, exploratory data analysis can be performed. The data analysis has been done in R Studio using the R statistical programming language. The goal of exploratory data analysis is to find if there are interesting relations between factors, that might be useful in the prediction of the duration of the application process for a specific application. The task that we try to accomplish is an example of classification. There are four classes: application processed within 25 days, within 25-50 days, within 50-75 days and within 75-100 days. Based on the input variables the algorithm should be able to predict the correct class. If it turns out the methods succeed in doing this, we can analyze the results to find out what factors influence the processing time.

The data mining applied is a form of supervised data mining. We have a training set that contains complete cases: applications and the outcomes. The selected algorithms can be trained on a selection of this dataset, and their performance can be validated using a training data set. The selected algorithms are Random Forests, Support Vector Machines and Neural Networks. These are selected because they support training data, in contrast to unsupervised algorithms. Also, these algorithms are able to handle datasets with a large set of input variables. Also, they are able to handle large datasets relatively quick. We will now introduce the three algorithms and briefly explain their workings. For a detailed explanation of the mathematical theories behind the algorithms we refer to the specific documentation published by the authors of the individual algorithms.

Random Forest

The implementation of Random Forest in R is developed by (Breiman and Cutler 2015). The package used is 'randomForest'. Random Forests are based on the construction of many classification trees. The random forest is trained on a dataset, which means a large number of classification trees is created. When the forest is then provided with new input, it will send the input to all classification trees and keep track of the classification provided by each tree. The classification that is most provided 'wins': the input is then classified according to that classification. It is basically an example of voting: the result of each classification tree can be seen as a vote for a certain class. The input it classified to the most voted class. Random forest are capable of handling large datasets in an efficient way, while at the same time supporting numbers sets of input variables. The default call is used in R, so no specific parameters are used. This way R will try to optimize the required settings automatically. This results in the following R-call to do the training:

```
randomForest(toBePredicted ~ ., data=dataset)
```

Support Vector Machines

(Fan et al. 2005) developed an implementation of Support Vector Machines, on which the implementation in the used R-package 'svmLinear' is based. The core functionality of a SVM is to find out the formula for a hyperplane. The hyperplane is the mathematical function that separates the different classes within the data. If for example all Medicaid applications are plotted, the SVM will try to find the function that separates the most applications handled on time from the applications handled too late. A linear SVM is used, because as (Joachims 2006)

states: *Linear SVMs provide state-of-the-art prediction accuracy*. Below the R-call for training the SVM is shown.

```
train(toBePredicted ~ ., data= dataset, method="svmLinear")
```

Neural network

(Ripley 1996) states that neural networks have arisen from analogies with models of the way humans approach pattern recognition tasks, although they have developed a long way from the biological roots. The used implementation, in the package 'nnet', is a feed-forward neural network with a single hidden layer. This means that the used neural network will consist of three layers: an input layer (these are the input variables), a hidden layer and the output layer (the predicted outcome). The hidden layer consists of a number of nodes. All nodes are connected, and the weights are initially randomly determined. During training based on feeding the neural network its training input, and comparing the predicted output with the actual output (from the training set) the weights of the nodes are adjusted. This process is called backpropagation. A feedforward neural network means that from the input layer to the output layer there is a one-directional information flow.

The default R settings for the 'nnet' package are used. The only modified parameter is the number of iterations. By default this is 100, but in this case it is manually set to 1000 iterations, because it showed better performance during preliminary tests. The R-call for training the neural network is shown below.

```
train(toBePredicted ~ ., data=dataset, method = "nnet", maxit=1000)
```

6 DATA ANALYSIS

In this chapter we will start with an exploratory data analysis and basic data mining on the created dataset, in section 6.1. After this analysis, the three selected machine learning algorithms will be applied to see if based on the dataset the expected processing time can be predicted (6.2). The results of both approaches are discussed in section 6.3.

6.1 Exploratory data analysis & data mining

Table 7 shows the number of applications in each processing class. This shows that over 65% of all applications is completed by a caseworker within 5 days after submission. It also shows 20,67% of all applications is not being completed within 45 days.

Class number	Number of days	Frequency	Percentage
1	0 – 5 days	8528	65,41%
2	6 – 10 days	272	2,09%
3	11 – 15 days	258	1,98%
4	16 – 20 days	282	2,16%
5	21 – 25 days	240	1,84%
6	26 – 30 days	233	1,79%
7	31 – 35 days	167	1,28%
8	36 – 40 days	162	1,24%
9	41 – 45 days	189	1,45%
10	46 – 50 days	172	1,32%
11	51 - 55 days	113	0,87%
12	56 – 60 days	149	1,14%
13	61 – 65 days	133	1,02%
14	66 – 70 days	117	0,90%
15	71 – 75 days	167	1,28%
16	76 – 80 days	298	2,29%
17	81 – 85 days	376	2,88%
18	86 – 90 days	516	3,96%
19	91 – 95 days	490	3,76%
20	95 – 100 days	176	1,35%

TABLE 7 - SPREAD OF APPLICATION PROCESSING TIMES

Table 8 shows the number of 'hard case' applications that are on time and too late. An application is considered a hard case in the following situations:

- An application submitted online, that needs a caseworker to be completed
- Applications submitted on paper, by phone and in person, that at least need one evidence item to be verified by a caseworker.

Processing times over 45 days are considered 'too late'.

	Total	'Hard cases'	On time	Too late
Online	28493	2137	482	1655
Paper	10408	9513	8634	879
Phone	40	37	30	7
In person	453	424	421	3

TABLE 8 - TIMELINESS OF APPLICATIONS BY METHOD

Table 8 makes it clear that the way an application enters the system, might influence the time it takes to process the application. Knowing this, it is interesting to find out why are there so many online applications handled too late, versus for example applications submitted in person or on paper.

In Table 7 an increase in the number of applications after day class 15 was noticed (this means all applications that took longer than 75 days to process). Figure 13 shows the relative number of applications and their processing time in days. Both paper and online applications are shown. Phone and 'In Person' applications are not reflected in the graph, because their number of applications by day was too low. The upper part of the image shows the division in full dataset. From this it becomes clear that a large amount of online applications that need a caseworker take over 75 days, whereas for paper applications that have at least one required evidence item most applications are completed on the same day. The second part shows the same data, but now the applications from November '14, December '14 and January '15 are removed from the dataset.

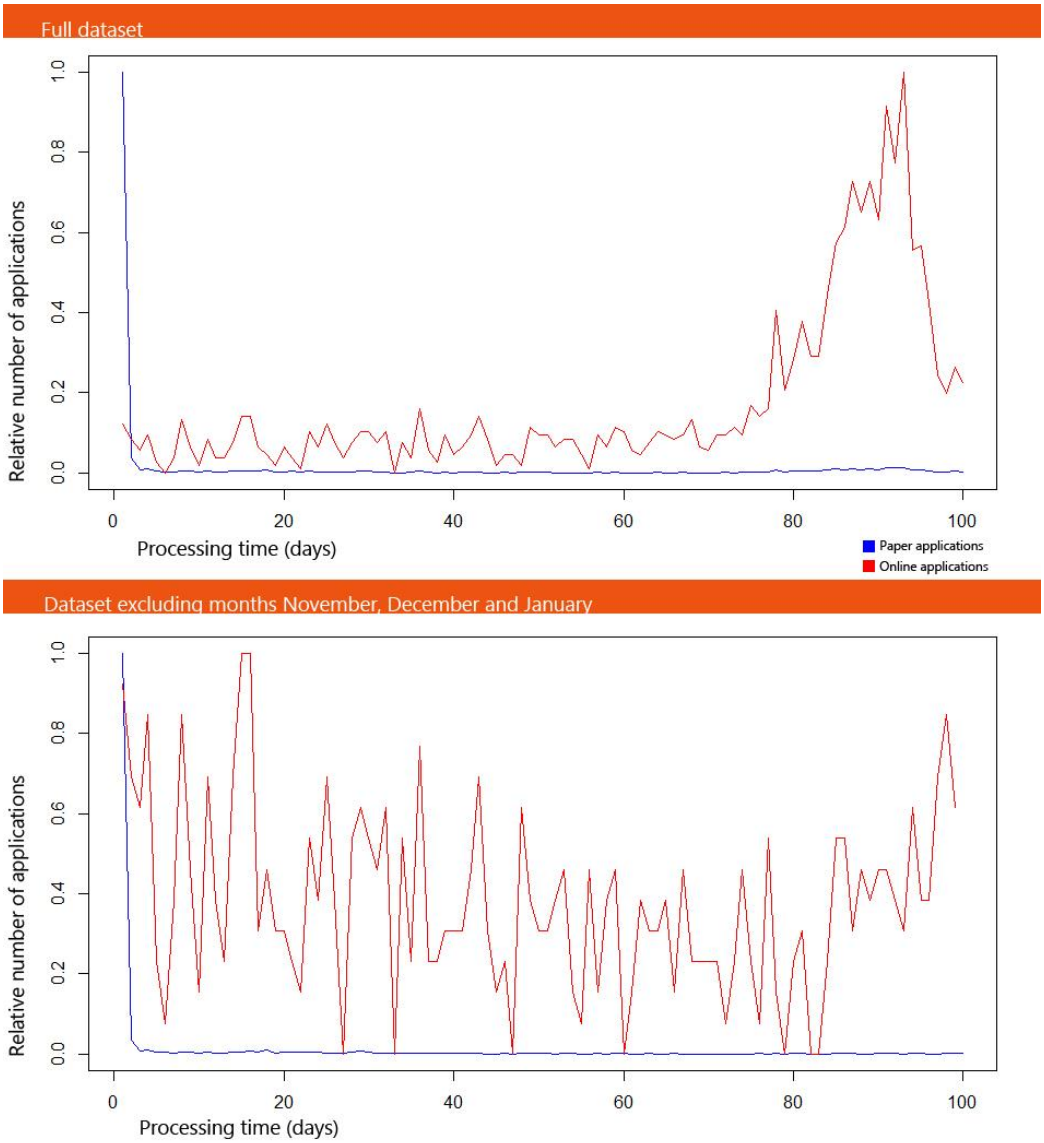


FIGURE 13 - DURATIONS OF APPLICATIONS

The second part of Figure 13 thus shows a different distribution of the processing time for online applications.

In order to find out if input variables correlate with the number of days for each method, we use a combination of Pearson's r score, Pearson's Chi Square test and Crámers V-value. Pearson's r score can be used to calculate the correlation between numerical values, while the combination of Chi Square and Crámers' V-value is to handle categorical variables.

	Online	Paper	Phone	In Person
Number of applicants	-0.14**	0.04**	-0.05	0.01
Number of missing evidence	-0.07**	0.01	-0.25	0.10**
SSN missing	-0.03	0.05**	-0.13	0.07
Citizenship missing	-0.04**	0.00	-0.22	0.06
Income missing	-0.08**	0.15**	0.00	0.06
Excluded income missing	0.03	-0.01	N/A	N/A
Gender missing	0.01	-0.01	-0.08	-0.01
Incarceration missing	N/A	-0.01	N/A	N/A
State residency missing	0.00	-0.01	N/a	-0.01
Benefits missing	0.00	-0.13**	-0.20	0.05
Former foster care missing	0.01	-0.02**	N/A	-0.01

TABLE 9 - CORRELATION BETWEEN VARIABLES AND NUMBER OF DAYS, BY APPLICATION METHOD
** = WITHIN 0.05 CONFIDENCE INTERVAL

Table 9 shows some correlations within the significance level. However, there are only weak correlations. The least weak correlation is the relation between the number of days and the missing income for paper applications. Because it is a positive number, this means that the more of this evidence is missing, it will lead to a higher number of processing days.

When looking at the categorical variables, as shown in Table 11, we used the categorical variable on time / too late as the dependent variable. The least weak correlation is between the payment category and the timeliness. There is too little spread in different preferred communication methods to calculate correlation using the Crámers V-value.

	Online	Paper	Phone	In Person
Preferred language	0.12**	0.10**	0.35	0.08
Eligibility decision*	0.06**	0.03**	N/A	0.02
Payment category*	0.22**	0.21**	0.36	0.21
Program determination*	0.11**	0.14**	N/A	0.64

TABLE 10 - CRÁMERS V-VALUES FOR CATEGORICAL CORRELATIONS BY APPLICATION METHOD

* = Fishers test used instead of chi square test for calculation of significance level

** = within 0.05 confidence interval

After calculating all correlations for the dataset, it becomes clear that there are no strong correlations between the various variables, and the length it takes to complete an application. However, the amount and types of evidence missing might differ for the different application methods. Table 11 shows for example that online applications have the highest percentage of missing social security number validations required, whereas for phone applications income information has to be validated in over 50% of the applications. Information about benefits is missing mostly at paper applications, compared to other application methods.

	SSN	Former foster care	State residency	Income	Gender	Citizenship	Benefits	Excluded income	Incarceration
Online	0.457	0.012	0.085	0.353	0.000	0.520	0.324	0.003	0.000
Paper	0.276	0.004	0.011	0.368	0.004	0.412	0.482	0.000	0.001
Phone	0.297	0.000	0.000	0.514	0.027	0.514	0.351	0.000	0.000
In Person	0.316	0.002	0.014	0.488	0.002	0.467	0.325	0.000	0.000

TABLE 11 – PERCENTAGE OF APPLICATIONS THAT REQUIRE A CERTAIN TYPE OF EVIDENCE TO BE VALIDATED

Concluding: for both online and paper applications most variables show a weak significant correlation to the time it takes to complete an application. Therefore, by combining them all, it might be possible to predict the application time. This is where machine learning comes in place: those algorithms are capable of handling a set of input variables, and based on that are able to come up with a prediction if an application is processed on time or not.

6.2 Predicting the number of days

As explained before a selection of data mining technologies has been made: Random Forest, Support Vector Machines and Neural Networks. Each of the technologies is applied on the dataset. A selection of test cases has been designed to see if variables are of influence, if so: which variables, and which machine learning method offers the best performance?

During the exploratory data analysis we saw that the amount of applications handled too late is the highest for online applications. The analysis will compare the set of online applications with the total set and the paper applications. The latter one seem to have a lower processing time on average. We can then see if there are differences, and if so: where those difference are.

Based on the results, the best performing algorithm will be selected. The best method is the one with the highest accuracy. The results of this algorithm will be validated using a ten-fold cross validation, to make sure the algorithm performs well in multiple runs.

The two tests that have been performed are:

1. Predict if an application will be assessed on time, based on the information known at the moment of submission. The variables are: number of people in the application, information known about missing evidence, the preferred communication method and the preferred language.
2. Predict if an application will be assessed on time , now with all the variables from step 1, and the program determination, eligibility decision, payment category and the caseworker experience.

The algorithms are trained on a dataset with a four-class dependent variable for processing time. This dependent variable has four possible values: less than 25 days, between 25 and 50 days, between 50 and 75 days and between 75 and 100 days. The exact prediction in Table 12 shows the percentage of all items from the validation set that were correctly classified by the algorithm. A correct classification means that the algorithm predicted the right class of time. 1 Off means that a class to the left and right of the actual class are also seen as correct. So if an application belongs to the class ‘between 50 and 75 days’ also the classes ‘between 25 – and 50 days’ and ‘between 75- 100 days’ are seen as a correct prediction. A detailed breakdown of all data mining methods and their accuracies can be found in Appendix C.

All variables known at moment of submission						
	Random forest		SVM		Neural network	
Accuracy	**Exact**	**1 Off**	**Exact**	**1 Off**	**Exact**	**1 Off**
All application methods	0.4492	0.8233	0.4222	0.7778	0.4572	0.7833
Only online applications	0.2800	0.6470	0.2736	0.6651	0.3066	0.6934
Only paper applications	0.3753	0.7193	0.3425	0.6438	0.3630	0.6678

All variables						
	Random forest		SVM		Neural network	
Accuracy	**Exact**	**1 Off**	**Exact**	**1 Off**	**Exact**	**1 Off**
All application methods	0.4791	0,8239	0.4012	0.8236	0.4322	0.8063
Only online applications	0.3620	0.7127	0.3125	0.7604	0.2760	0.5729
Only paper applications	0.3793	0.7372	0.3714	0.7607	0.3571	0.7179

TABLE 12 - PERFORMANCE OF MACHINE LEARNING ALGORITHMS ON DATASET. PERCENTAGES. SEE APPENDIX C FOR AN EXTENDED VERSION OF THIS TABLE

From the analysis in Table 12 the Neural Network seems to provide slightly better results than the SVM and the Random Forest when only looking at data available on submission. When looking at online applications only, the accuracy of the algorithm is 30,66%. This means that the majority of all applications are misclassified. For 1 off classification the accuracy of the Neural Network is just 69,34%. The performance is better when variables that are known after the decision is made are taken into account. SVM seem to give the best performance in that case.

These low accuracies show that it is hard for the selected machine learning methods to do predictions based on the input data set. Based on these results, combined with the insights from the correlations in the exploratory data analysis phase, the dataset does not seem to provide a clear indication of why some applications take longer to process than others.

6.3 Discussion of results

The results have been discussed a Production Support Manager. He explained that the system was put in use in November 2014, and that employees were trained until the 2nd of December. The high amount of applications can be explained because of the Open Enrollment Period, which took place in November 2014, December 2014 and January 2015. Employees might be overwhelmed by the amount of applications in this period.

He suggested that if only data starting from January/February 2015 is taken into account, the figures would look totally different. And they indeed do. Another explanation for possible longer processing times at the end of 2014 would be the fact that a lot of case workers take days of. Both Thanksgiving and Christmas take place in December, and a reason for a lot of employees to take days off.

Another suggestion by the Production Support Manager was that there might be relations between types of evidence missing, and the type of applicant. Pregnant woman, or deemed babies, are known as 'hard cases' when it comes to the validation of information. A newborn baby for example might not be in the systems of the Social Security Administration at the moment someone tries to enroll the baby in Medicaid. Therefore his/her social number cannot be automatically validated. In order to see if this is a correlation that can be derived from the dataset, the correlation between the payment category and the evidence types was calculated. The payment category can be used to identify the different types of applicants. This shows only a weak correlation between the payment category and the number of evidence ($p=0.000$, Cramers' $V=0.072^{**}$). In the case of deemed babies, it was suggested that state residency, citizenship or SSN might be troublesome during applications. Also for these evidence items and the payment category the correlation was calculated, to see if certain payment categories (e.g. the one for deemed babies) correlate with these specific evidence types. Only very weak correlations were found, and thus the data does not support this presumption.

7 CONCLUSION

This chapter is the concluding chapter of this thesis. In this chapter the answers to the research questions will be given. First the sub questions will be answered (sections 7.1 – 7.5), and finally an answer to the main research question will be given in section 7.6.

In this research the aim is to improve the Medicaid enrollment process for the state of South Carolina. Based on the categories of waste identified by (Hackbarth 2012) two areas can be linked to the Medicaid enrollment process: administrative complexity and fraud. The first two research questions are answered by performing a literature study.

7.1 Medicaid enrollment

The first research question helps to get an insight in the state-of-art, from a scientific perspective, of the Medicaid enrollment process. The question is formulated as follows:

SQ1. What is the state-of-art in literature regarding Medicaid enrollment processes?

Based on a literature study 12 articles were relevant to this question. Even though there is a large selection of literature on the Medicaid program, and on people who are enrolled in the program, literature on the enrollment process itself is barely available. From the result set some factors regarding Medicaid enrollment could be identified. These are:

- Language
- Types of proof that have to be provided
- Method of application

It turns out the complexity of the language influences how citizens perceive the Medicaid enrollment process. So, for non-English speakers, enrollment forms should be simple and use easy language. The types and amount of proof people have to provide influence both the perceived difficulty of enrolling in Medicaid, as well influence the time it takes for case workers to handle applications. Requiring more proof of income (e.g. statements covering multiple years) or proof of citizenship has an influence on how easy people can enroll in Medicaid. Furthermore, local welfare offices play an important role in the application process for children.

7.2 Healthcare fraud

Healthcare fraud is the second topic that has been studied using literature. The research was set up around the following sub question:

SQ2. How can Medicaid enrollment fraud types be classified?

Based on the literature two classifications were actually created. First, a classification of literature to fraud topics was done. This helps to identify which fraud topics in healthcare are frequently discussed, and which ones lack attention. An overview of 24 classes of health care fraud was created.

Improper coding, phantom billing and kickback schemes are the most discussed methods. The methods that have only been discussed in one paper are: self-referrals, reverse false claim cases, false negotiation cases, off-label promotion of drugs and managed care fraud. When looking at fraud related to the enrollment process, there are two related fraud types. The first one is identity fraud, which is identified by 6 authors. Second related fraud type is misrepresenting eligibility, which is mentioned by two authors. From this part of the literature study it can be concluded that fraud related to enrollment in Medicaid gets little attention in literature.

The second part of the literature analysis on health care fraud is concerned with the identification of fraud detection methods. 17 Fraud detection methods are described in literature, that can be useful for the detection of health care fraud. Clustering is the most discussed method, mentioned 9 times. Outlier detection, classification and association rules are also amongst the most popular methods. Only one paper is concerned with optimizing the enrollment process to reduce the chance of fraud (Suleiman et al. 2013).

7.3 Data involved in the Medicaid application process

The third sub question is concerned with the data involved in the Medicaid enrollment process. The analysis will be done to answer the following research question:

SQ3. What data is needed to analyze the Medicaid application process?

First, an analysis of all systems and the relations between the systems was made. SCDHHS uses two systems for eligibility determination: Cúram and MEDS. The systems exchange information with each other, but also interact with other systems. One of those other systems is MMIS, the

Medicaid Management Information System. This system is used for the handling of Medicaid claims. For this research the focus was put on Cúram, because this data was easily accessible, and CCIT/SCHHS plan on migrating to this system as the main Medicaid eligibility decision system. This means that MEDS is going towards an 'end-of-life' situation. This system handles most applications already, and has the most 'entrances': the systems handles online applications, as well as all applications entered by caseworkers. MEDS only handles applications handled by caseworkers.

A selection of variables are selected for the dataset, based on the three factors identified in SQ1: language, types of proof needed and the application method. The set is extended with some more characteristics of an application: the number of people in the application and the preferred communication method. Furthermore, some post-application characteristics are selected: the program determination (if a person was deemed eligible for Medicaid or CHIP) and a more specific subcategory of that decision: the payment category. Also the decision of the application is taken into account: was the applicant eligible or not, after processing his information? For each application also the caseworker that handled the application is included in the dataset.

7.4 Data mining method

Several data mining method are selected, to compare their performance. This will help to answer the research question:

SQ4. What is the best suitable data mining approach to analyze the Medicaid application process?

First, an exploratory data analysis was conducted. In this analysis correlations between input factors and the output factor were determined. To do this, Pearson's r score was used for the non-categorical input variables, and a combination of Chi-square test and Crámers V-value was used for the categorical variables. No strong correlations are identified.

Secondly three machine learning algorithms are trained on the dataset. The selected algorithms are Support Vector Machines, Random Forests and a Neural Network. These algorithms support a large set of input variables, and are able to execute tasks on the dataset quickly. All algorithms trained on several combinations of the dataset. A total of 11 cases was executed. Not only to see what algorithm would give the best performance, but also to find the input variables that had a

large influence on the outcome. All executed tests had unsatisfying accuracies, all below 50% accuracy. Of the excluded variables the Method of Application seems to have the biggest influence: the accuracy is lowest when this variable is excluded from the dataset.

The best performing data mining method are neural networks. Even though accuracies are low in general in this case, neural networks tend to give the 'highest' accuracy. However, the difference between the three algorithms was small in most cases. This algorithm is able to predict with an accuracy of 47,92 percent if an application would be handled on time. However, based on this result it turns out that even the best performing method used in this research does not provide sufficient performance to be usable.

7.5 Evaluation

The last sub question has the aim to find out what the data analysis tells about the Medicaid application process. The question is formulated as follows:

SQ5. What do the results of the data analysis tell about the Medicaid application process?

There are no strong correlations found between the different characteristics of Medicaid applications. The method an application is submitted seems to influence the processing time however. For online submitted applications which require a caseworker to be completed the data analysis shows that for a large amount of applications submitted in November '14, December '14 and January '15 the processing took longer than compared to other application methods and more recent online applications. Discussing these results with a Production Support Manager made clear that the Cúram system for handling online applications was introduced in November '14, and that employees were trained until December '14. Furthermore, in these months more employees take days off, because of for example Christmas and Thanksgiving. These factors might influence the processing time of the online applications during the before mentioned months.

From the data analysis we can derive that there are certain types of evidence that are missing mostly using specific application methods. SSN information has to be manually validated mostly in the case of online applications, whereas information about benefits an applicant already has is a evidence type that has to be mostly validated at paper applications. Based on the findings the

four evidence types that mostly need to be validated, amongst all application methods are: social security numbers, income information, citizenship information and information about benefits.

7.6 Conclusion of this research

Using the answers to all sub questions, it is now possible to answer the main research question. The question is:

How can the Medicaid eligibility determination process be successfully optimized?

The literature study shows there is little research into the technical aspects of managing Medicaid. Waste in the Medicaid eligibility determination process can occur because of fraud or administrative complexity. Citizens can apply using multiple methods for Medicaid: online, by mail, by phone or at a local office. In this research the focus was on the one application that handles applications by all those methods: Cúram. Based on the data it becomes clear there are big differences in the processing times of applications. A lot of applications are processed on the same day as they are submitted, but over 20 percent of all applications that involve a caseworker take more than 45 days to process.

Looking in more detail to this data, it turns out that online applications that need caseworker involvement have the most delays. Over 75 percent of them takes more than 45 days to process. These applications are analyzed, by looking at correlations that might influence the processing time and by applying machine learning methods. There are no strong correlations found and also the machine learning methods do not provide insights in what might cause those delays.

Based on this we can conclude that in order to optimize the Medicaid application process the aim should be to get as little online applications that need caseworker assistance. The main reason for requiring caseworker assistance is when evidence that is required for the application cannot be validated during the online application process. Suggestions on how to improve the evidence checks will be given in the next section, as part of the 'recommendations for practice'.

8 RECOMMENDATIONS

Based on the results we can come up with recommendations for science, as well provide recommendations for practice, that can be used CCIT and SCDHHS. For science suggestions for future research are provided in section 8.1, whereas for CCIT/SCDHHS suggestions for anticipating on the findings of this research are given in section 8.2.

8.1 Recommendations for science

The literature study shows there is little interest from a research perspective in the Medicaid application process. We found that authors mention the fact that increasing the need for people to provide evidence increases the burden to enroll in Medicaid. On the other hand: the healthcare system suffers from fraud. This is supported by our research to healthcare fraud methods. There is no research on the trade-off: how can states find the best balance between the amount of information they require from applying citizens and the chance of fraud?

Related to the topic of asking citizens to provide proof is the complexity of an application. Citizens need to provide a lot of information about their situation. An assessment to which extent the requested information is all needed by (SC)DHHS for eligibility determination seems interesting. When more information is asked from citizens, more information needs to be validated. This might lead to administrative complexity, which is one of the areas of waste in healthcare as identified by (Hackbarth 2012).

Medicaid is managed on a state-level. The Medicaid program is thus managed at 50 different places. Some states use the same computer systems to manage the program. States have subtle differences in their eligibility criteria. Especially on the topic of fraud in Medicaid there is almost no literature available. The costs to society of fraud are expected to be significant. Given the fact Medicaid contributes for a big part to the healthcare spending in the U.S., more research in the detection of fraud in Medicaid is needed in order to create an efficient and future-proof program. Especially on an inter-state level this could be interesting: are there approaches possible to improve Medicaid process while reducing fraud, that are applicable on a general level?

8.2 Recommendations for practice

8.2.1 Incorporation of more external data sources

A topic that is both mentioned during the literature study, as well during data analysis: the validation of information people provide. Citizenship, social security information and income information are some examples of information that applicants provide which requires validation. In the current situation some external data sources are used for validation. As the data analysis shows there are a lot of applications which needs additional evidence verification. By adding more external data sources, more information can be validated for the citizen. This will reduce the need for manual verification and might therefore reduce the number of applications that need to be handled by caseworkers. They can be automatically handled by Cúram.

(Suleiman et al. 2013) describe an approach for the state of North Carolina, on how to verify user-provided information. They start the process with verification of the applicants' Social Security Number (SSN). That number is sent to the BENDEX system (Beneficiary and Earnings Data Exchange). This way they are able to verify if the SSN is an existing number. If the SSN exists, the next database consulted is USCIS (United States Citizenship and Immigration Services), the system containing all immigration data. If the person is not a US citizen or an Alien Resident, the application is rejected. Next, characteristics of the applicant are being reviewed. This is a threefold process, in which the assets, DMV information (Department of Motor Vehicles) and income are assessed. Based on the results of these individual checks, a score is calculated. If the score is below a certain threshold the application is accepted. Cases in which the score can be higher than a threshold are for example: a person owns more motor vehicles than allowed, or has a too high income.

Another overview of possibilities for determining eligibility using external data sources is given by the FGA (Foundation for Government Accountability). This organization identifies three areas of which information on an enrolled person could be verified. These are 'Identity verification', 'Earnings & assets verification' and 'Additional benefits verification'. The categories can be linked to the types of missing evidence which were commonly amongst the missing evidence types. Table 13 displays all steps that are suggested.

Category	Proposed solution
Identity verification	Verify and confirm identity of all applicants before granting benefits
	Check a nationwide best-address and driver's license data source to verify individuals are residents of the state
	Check a comprehensive public records database that identifies potential identity fraud or identity theft that can closely associate name, Social Security Number, date of birth, phone and address information.
	Check immigration status information maintained by U.S. Citizenship and Immigration Services
	Check death register information maintained by SSA
	Check prisoner information maintained by the SSA
	Check national fleeing felon information maintained by the FBI
Earnings & Assets Verification	Check for unearned income with the IRS
	Check employer quarterly reports of income and unemployment insurance payments
	Check earned income information maintained by the SSA
	Check wage reporting and similar information maintained by bordering states
	Check earnings information maintained by the SSA in the BENDEX
	Check earnings and pension information maintained by the SSA in BEERS (Beneficiary Earnings Exchange Record System)
	Check employment information maintained by the state
	Check employment information maintained by the U.S. Department of Health and Human Services in its National Directory of New Hires database
	Check a database of all persons who currently hold a license, permit or certificate from any state agency the cost of which exceeds $500
	Check income and employment information maintained by the state's and the USDHHS office Child Support Enforcement
	Check earnings and pension information maintained by the state

	Check a nationwide public records data source of physical asset ownership: such as real property, automobiles, watercraft, aircraft and luxury vehicles or any other vehicle
Additional Benefits Verification	Check public housing payment information maintained by the Department of Housing and Urban development
	Check child care services information maintained by the state
	Check utility payments information maintained by the state under the Low Income Home Energy Assistance Program
	Check emergency utility payment information maintained by the state or local entities
	Check supplemental security income information maintained by the SSA in its SDX
	Check state veterans' benefits information against the federal Public Assistance Reporting Information System database maintained by DHHS
	Check any existing real-time database of persons currently receiving benefits in other states, such as the National Accuracy Clearinghouse

TABLE 13 - AUTOMATIC ELIGIBILTY DETERMINATION

8.2.2 Process monitoring

Currently there is little process monitoring in place for the eligibility determination at SCDHHS. There is no insight in how long it takes to handle applications, and no insight in who is handling what kind of cases.

When caseworkers are able to see what applications are open for a long time, they might be prioritized in order to reduce the average handling time. If it is clear which caseworkers handle applications that take a long time to complete, they can be asked to provide feedback on the system. Maybe there are improvements possible from a systems' perspective. Improvements that cannot be derived by just looking at the database, but which can be suggested by caseworkers who handle the actual 'hard' applications.

Insight into the process provides the possibility to measure the result of changes in the system. When more external data sources are integrated for example, a dashboard can provide insight into the changes in processing times. Having those insights helps in the assessment of what

changes to the system actually improve the system. From a 'customer perspective': it is in the citizens' interest if his or her application is handled as quickly as possible.

REFERENCES

Agrawal, S., Tarzy, B., Hunt, L., Taitsman, J., and Budetti, P. 2013. “Expanding Physician Education in Health Care Fraud and Program Integrity:,” *Academic Medicine* (88:8), pp. 1081–1087 (doi: 10.1097/ACM.0b013e318299f5cf).

Allison, R. A. 2003. “The impact of local welfare offices on children’s enrollment in Medicaid and SCHIP,” *INQUIRY: The Journal of Health Care Organization, Provision, and Financing* (40:4), pp. 390–400.

Aral, K. D., Güvenir, H. A., Sabuncuoğlu, İ., and Akar, A. R. 2012. “A prescription fraud detection model,” *Computer Methods and Programs in Biomedicine* (106:1), pp. 37–46 (doi: 10.1016/j.cmpb.2011.09.003).

Avruch, S., Machlin, S., Bonin, P., and Ullman, F. 1998. “The demographic characteristics of Medicaid-eligible uninsured children.,” *American journal of public health* (88:3), pp. 445–447.

Bennett, R. S., and Medearis, D. M. 2003. “Health Care Fraud; Recent Developments and Timeless Advice,” *Texas medicine* (99), pp. 50–56.

Benzio, B. 2009. “Fee-for-Disservice: Medicare Fraud in the Home Healthcare Industry,” *Annals Health L.* (19), p. 229.

Bonchi, F., Giannotti, F., Mainetto, G., and Pedreschi, D. 1999. “A classification-based methodology for planning audit strategies in fraud detection,” in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 175–184 (available at http://dl.acm.org/citation.cfm?id=312224).

Borca, G. 2001. “Technology Curtails Health Care Fraud,” *Managed Care Magazine* (10:4), pp. 50–53.

Breiman, L., and Cutler, A. 2015. “Random Forests,” *Berkely Statistics*, June 28 (available at https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm; retrieved June 28, 2015).

Brooks, G., Button, M., and Gee, J. 2012. “The scale of health-care fraud: A global evaluation,” *Security Journal* (25:1), pp. 76–87.

Buppert, C. 2001. “Avoiding Medicare Fraud,” *Nurse Practitioner* (26:2), pp. 36–38,41.

Byrd, J. D., Powell, P., and Smith, D. L. 2013. “Health Care Fraud: An Introduction to a Major Cost Issue,” *Journal of Accounting, Ethics and Public Policy* (14:3) (available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2285860).

Cady, R. F. 2007. “Healthcare Fraud: A Primer for the Nurse Executive,” *JONA’S healthcare law, ethics and regulation* (9:2), pp. 54–61.

Capelleveen, G. C. 2013. “Outlier based predictors for health insurance fraud detection within US Medicaid,” (available at http://essay.utwente.nl/64417/).

Carlson, J. 2013. “Panful side effects,” Modern Healthcare.

Chen, S., and Gangopadhyay, A. 2013. “A Novel Approach to Uncover Health Care Frauds through Spectral Analysis,” IEEE, September, pp. 499–504 (doi: 10.1109/ICHI.2013.77).

Clements, P., and Bass, L. 2010. “Using Business Goals to Inform a Software Architecture,” IEEE, September, pp. 69–78 (doi: 10.1109/RE.2010.18).

Copeland, L., Edberg, D., Panorska, A. K., and Wendel, J. 2012. “Applying business intelligence concepts to medicaid claim fraud detection,” *Journal of Information Systems Applied Research* (5:1), p. 51.

DeLeire, T., Leininger, L., Dague, L., Mok, S., and Friedsam, D. 2012. “Wisconsin’s Experience with Medicaid Auto-Enrollment: Lessons for Other States,” *Medicare & Medicaid Research Review* (2:2) (doi: 10.5600/mmrr.002.02.a02).

Department of Health and Human Services. 2015. “Medicaid and medicare eligibility requirements.pdf,” Department of Health and Human Services (available at https://www.medicare.gov/Pubs/pdf/11306.pdf).

Van Der Spoel, S., Van Keulen, M., and Amrit, C. 2013. “Process prediction in noisy data sets: a case study in a dutch hospital,” in *Data-Driven Process Discovery and Analysis*, Springer, pp. 60–83 (available at http://link.springer.com/chapter/10.1007/978-3-642-40919-6_4).

Doan, R. 2011. “False Claims Act and the Eroding Scienter in Healthcare Fraud Litigation, The,” *Annals Health L.* (20), p. 49.

Dube, J. F. 2011. “Fraud in Health Care and Organized Crime,” *Medicine & Health* (94:9), pp. 268–269.

Ekina, T., Leva, F., Ruggeri, F., and Soyer, R. 2013. “Application of Bayesian Methods in Detection of Healthcare Fraud,” *in Chemical Engineering Transaction* (33) (available at http://www.aidic.it/cet/13/33/026.pdf).

Evans, R. D., and Porche, D. A. 2005. “The nature and frequency of medicare/medicaid fraud and neutralization techniques among speech, occupational, and physical therapists,” *Deviant Behavior* (26:3), pp. 253–270 (doi: 10.1080/01639620590915167).

Fan, R.-E., Chen, P.-H., and Lin, C.-J. 2005. “Working set selection using second order information for training support vector machines,” *The Journal of Machine Learning Research* (6), pp. 1889–1918.

Freeman, B. A., and Loavenbruck, A. 2001. “Complying with healthcare fraud laws: an overview for the hearing professional,” *The Hearing Journal* (54:5).

Galewitz, P. 2013. “Even Without Expansion, S.C. Will See 16% Jump In Medicaid Enrollment,” *Kaiser Health News*, Blog, , November 26 (available at http://kaiserhealthnews.org/news/south-carolina-medicaid-enrollment/; retrieved March 3, 2015).

Gangopadhyay, A., Chen, S., and Yesha, Y. 2012. “Detecting healthcare fraud through patient sharing schemes,” in *Information Systems, Technology and Management*, Springer, pp. 421–426 (available at http://link.springer.com/chapter/10.1007/978-3-642-29166-1_39).

Hackbarth, A. D. 2012. “Eliminating Waste in US Health Care,” *JAMA* (307:14), p. 1513 (doi: 10.1001/jama.2012.362).

Hansen, J. S., Wallace, L. S., and DeVoe, J. E. 2011. “How Readable are Spanish-Language Medicaid Applications?,” *Journal of Immigrant and Minority Health* (13:2), pp. 293–298 (doi: 10.1007/s10903-010-9435-4).

Hatch, B. A., DeVoe, J. E., Lapidus, J. A., Carlson, M. J., and Wright, B. J. 2014. “Citizenship Documentation Requirement for Medical Eligibility: Effects on Oregon Children,” (available at http://pdxscholar.library.pdx.edu/soc_fac/4/).

He, H., Graco, W., and Yao, X. 1998. “Application of Genetic Algorithm and k-Nearest Neighbour Method in Medical Fraud Detection,” Presented at the Second Asia-Pacific Conference on Simulated Evolution and Learning on Simulated Evolution and Learning, Springer-Verlag, pp. 74–81 (available at http://dl.acm.org/citation.cfm?id=669969).

He, H., Wang, J., Warwick, G., and Hawkins, S. (n.d.). “Application of neural networks to detection of medical fraud,” *Expert Systems With Applications*1997 (13:4), pp. 329–336.

Hill, C., Hunter, A., Johnson, L., and Coustasse, A. 2014. “Medicare Fraud in the United States: Can it Ever be Stopped?,” *The Health Care Manager* (33:3), pp. 254–260 (doi: 10.1097/HCM.0000000000000019).

IBM. 2015. “Health and social program delivery that improves outcomes and reduces costs,” *IBM.com*, June 23 (available at http://www-03.ibm.com/software/products/en/social-programs; retrieved June 23, 2015).

Joachims, T. 2006. “Training linear SVMs in linear time,” in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 217–226 (available at http://dl.acm.org/citation.cfm?id=1150429).

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., and Arab, M. 2014. “Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature,” *Global Journal of Health Science* (7:1) (doi: 10.5539/gjhs.v7n1p194).

Kenney, G. M., Haley, J. M., Anderson, N., and Lynch, V. 2015. “Children Eligible for Medicaid or CHIP: Who Remains Uninsured, and Why?,” *Academic pediatrics* (15:3), pp. S36–S43.

Kenney, G. M., Lynch, V., Haley, J., and Huntress, M. 2012. “Variation in Medicaid Eligibility and Participation among Adults: Implications for the Affordable Care Act,” *Inquiry* (49:3), pp. 231–253 (doi: 10.5034/inquiryjrnl_49.03.08).

Kirlidog, M., and Asuk, C. 2012. “A Fraud Detection Approach with Data Mining in Health Insurance,” *Procedia - Social and Behavioral Sciences* (62), pp. 989–994 (doi: 10.1016/j.sbspro.2012.09.168).

Konijn, R. M., and Kowalczyk, W. 2011. “Finding fraud in health insurance data with two-layer outlier detection approach,” in *Data Warehousing and Knowledge Discovery*, Springer, pp. 394–405 (available at http://link.springer.com/chapter/10.1007/978-3-642-23544-3_30).

Kumar, M., Ghani, R., and Mei, Z.-S. 2010. “Data mining to predict and prevent errors in health insurance claims processing,” in *Proceedings of the 16th ACM SIGKDD international*

conference on Knowledge discovery and data mining, ACM, pp. 65–74 (available at http://dl.acm.org/citation.cfm?id=1835816).

Lessner, J. F. 2006. "'You Need How Many Months of Bank Statements□' Medicaid Long-Term Care Eligibility Changes After the DRA," *Geriatric Nursing* (27:6), pp. 334–335.

Lin, K.-C., and Yeh, C.-L. 2012. "Use of Data Mining Techniques to Detect Medical Fraud in Health Insurance," *International Journal of Engineering and Technology* (2:2), pp. 126–137.

Liou, F.-M., Tang, Y.-C., and Chen, J.-Y. 2008. "Detecting hospital fraud and claim abuse through diabetic outpatient services," *Health Care Management Science* (11:4), pp. 353–358 (doi: 10.1007/s10729-008-9054-y).

Liu, Q., and Vasarhelyi, M. 2013. "Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information," in *29th world continuous auditing and reporting symposium* (available at http://raw.rutgers.edu/docs/wcars/29wcars/Health%20care%20fraud%20detection%20A%20survey%20and%20a%20clustering%20model%20incorporating%20Geo-location%20information.pdf).

Lubao, M. 2008. "Claiming responsibility," *Health Management Technology* (June 2008).

Lu, F., and Boritz, J. E. 2005. "Detecting fraud in health insurance data: Learning to model incomplete Benford's law distributions," in *Machine Learning: ECML 2005*, Springer, pp. 633–640 (available at http://link.springer.com/chapter/10.1007/11564096_63).

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

Medicaid.gov. 2015. "Children's Health Insurance Program (CHIP) Program Information," *Medicaid.gov*, Government website, , July 7 (available at http://www.medicaid.gov/chip/chip-program-information.html; retrieved July 7, 2015).

Michael, J. E. 2003. "What Home Healthcare Nurses Should Know about Fraud and Abuse," *Home healthcare nurse* (21:8), pp. 522–530.

MIT OpenCourseWare. 2008. "Computational Biology: Genomes, Networks, Evolution," MIT (available at http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-047-computational-biology-genomes-networks-evolution-fall-2008/lecture-notes/MIT6_047f08_lec04_slide04.pdf).

Morris, L. 2009. "Combating Fraud In Health Care: An Essential Component Of Any Cost Containment Strategy," *Health Affairs* (28:5), pp. 1351–1356 (doi: 10.1377/hlthaff.28.5.1351).

Musal, R. M. 2010. "Two models to investigate Medicare fraud within unsupervised databases," *Expert Systems with Applications* (37:12), pp. 8628–8633 (doi: 10.1016/j.eswa.2010.06.095).

National Medicaid Fraud and Abuse Initiative. 2000. "Guidelines for addressing fraud and abuse in Medicaid Managed Care," National Medicaid Fraud and Abuse Initiative (available at http://www.cms.gov/Medicare-Medicaid-Coordination/Fraud-

Prevention/FraudAbuseforProfs/Downloads/GuidelinesAddressingfraudabuseMedMngdCare.pdf).

Ngufor, C., and Wojtusiak, J. 2013. “Unsupervised labeling of data for supervised learning and its application to medical claims prediction,” *Computer Science* (14:3), p. 191 (doi: 10.7494/csci.2013.14.2.191).

Ogunbanjo, G. A., and Knapp van Bogaert, D. 2014. “Ethics in health care: healthcare fraud: ethics CPD supplement,” *South African Family Practice* (56:1), pp. S10–S13.

OIG US DHHS. 2014a. “Compendium of Priority Recommendations,” US Department of Health and Human Services (available at http://oig.hhs.gov/reports-and-publications/compendium/files/compendium2014.pdf).

OIG US DHHS. 2014b. “Challenge 3: Protecting an Expanding Medicaid Program From Fraud, Waste, and Abuse,” *Reports and publications of the office of the inspector general for U.S> Department of Health and Human Services*, December 31 (available at https://oig.hhs.gov/reports-and-publications/top-challenges/2014/challenge03.asp; retrieved January 7, 2015).

Ormerod, T., Morley, N., Ball, L., Langley, C., and Spenser, C. 2003. “Using ethnography to design a Mass Detection Tool (MDT) for the early discovery of insurance fraud,” in *CHI’03 Extended Abstracts on Human Factors in Computing Systems*, ACM, pp. 650–651 (available at http://dl.acm.org/citation.cfm?id=765910).

Ortega, P. A., Figueroa, C. J., and Ruz, G. A. 2006. “A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile.,” *DMIN* (6), pp. 26–29.

Peng, Y., Kou, G., Sabatka, A., Chen, Z., Khazanchi, D., and Shi, Y. 2006. “Application of clustering methods to health insurance fraud detection,” in *Service Systems and Service Management, 2006 International Conference on* (Vol. 1), IEEE, pp. 116–120 (available at http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4114418).

Phua, C., Alahakoon, D., and Lee, V. 2004. “Minority report in fraud detection: classification of skewed data,” *ACM SIGKDD Explorations Newsletter* (6:1), pp. 50–59.

Plomp, M. G. A., and Grijpink, J. 2011. “Combating identity fraud in the public domain: information strategies for healthcare and criminal justice,” in *Proceedings of the 11th European Conference on e-Government*, pp. 451–458 (available at http://books.google.com/books?hl=en&lr=&id=cIkxJXGyhv8C&oi=fnd&pg=PA451&dq=%22interdependent.+ICT+can+support+the+development+of+interorganisational+relations+through%22+%22shares+which+information+and+why,+are+complex+and+important+as+well.+It+can+therefore%22+%22deployed,+there+are+many+potential+problems+in+their+use+that+need+to+be+tak&ots=RfYa1PlnvL&sig=P7z719JXdFh79V9tIP_XYKSwGNc).

Priest, D. (n.d.). “CHIP and Medicaid Outreach and Enrollment: A Hands-On Look at Marketing and Applications,” (available at http://www.w.nhpf.org/library/issue-briefs/IB748_SCHIPOutreach_10-19-99.pdf).

Rabecs, R. N. 2006. “Health care fraud under the new Medicare Part D prescription drug program,” *The Journal of Criminal Law and Criminology*, pp. 727–756.

Rashidian, A., Joudaki, H., and Vian, T. 2012. “No Evidence of the Effect of the Interventions to Combat Health Care Fraud and Abuse: A Systematic Review of Literature,” *PLoS ONE* (H. R. Baradaran, ed.) (7:8), p. e41988 (doi: 10.1371/journal.pone.0041988).

Ripley, B. D. 1996. *Pattern Recognition and Neural Networks* (1st ed.), Cambridge University Press (available at https://books.google.com/books?hl=nl&lr=&id=2SzT2p8vP1oC&).

Rudman, W. J., Eberhardt, J. S., Pierce, W., and Hart-Hester, S. 2009. “Healthcare fraud and abuse,” *Perspectives in Health Information Management/AHIMA, American Health Information Management Association* (6:Fall) (available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2804462/).

SC DHHS. 2015. “Frequently Asked Questions,” *South Carolina Healthy Connections*, January 7 (available at https://www.scdhhs.gov/FAQs; retrieved January 7, 2015).

Shan, Y., Jeacokcke, D., Murray, D. W., and Sutinen, A. 2008. “Mining medical specilist billing patterns for health service management,” Presented at the Seventh Australasian Data Mining Conferenc, Glenelg, Australia: Australian Computer Society, Inc.

Shan, Y., Murray, D. W., and Sutinen, A. 2009. “Discovering inappropriate billings with local density based outlier detection method,” in *Proceedings of the Eighth Australasian Data Mining Conference-Volume 101*, Australian Computer Society, Inc., pp. 93–98 (available at http://dl.acm.org/citation.cfm?id=2449380).

Shin, H., Park, H., Lee, J., and Jhee, W. C. 2012. “A scoring model to detect abusive billing patterns in health insurance claims,” *Expert Systems with Applications* (39:8), pp. 7441–7450 (doi: 10.1016/j.eswa.2012.01.105).

Shmueli, G., and Koppius, O. 2010. “Predictive analytics in information systems research,” *Robert H. Smith School Research Paper No. RHS*, pp. 06–138.

Sommers, B. D. 2010. “Enrolling Eligible Children In Medicaid And CHIP: A Research Update,” *Health Affairs* (29:7), pp. 1350–1355 (doi: 10.1377/hlthaff.2009.0142).

Sozio, S. G., and Noss, W. W. 2002. “Medicare Fraud-Busters Target Patient Transfer Reporting,” *Health Law* (55:8), pp. 46–47.

Sparrow, M. K. 2008. “Fraud in the US Health-Care System: Exposing the Vulnerabilities of Automated Payments Systems,” *social research*, pp. 1151–1180.

Van der Spoel, S., and Amrit, C. 2015. “The Relation between method and system complexity for predictive analytics in the supply chain (Working paper),” University of Twente.

Stanton, T. H. 2001. “Fraud-And-Abuse Enforcement In Medicare: Finding Middle Ground,” *Health Affairs* (20:4), pp. 28–42 (doi: 10.1377/hlthaff.20.4.28).

Stelfox, H. T., and Redelmeier, D. A. 2003. “An analysis of one potential form of health care fraud in Canada,” *Canadian Medical Association Journal* (169:2), pp. 118–119.

Stuber, J., and Bradley, E. 2005. “Barriers to Medicaid Enrollment: Who is at risk?,” *American journal of public health* (95:2), pp. 292–298.

Suleiman, M., Agrawal, R., Grosky, W., and Andres, F. 2013. “A generic data driven approach for Medicaid fraud detection,” in *Proceedings of the Fifth International Conference on*

Management of Emergent Digital EcoSystems, ACM, pp. 233–234 (available at http://dl.acm.org/citation.cfm?id=2536182).

Tagaris, A., Konnis, G., Benetou, X., Dimakopoulos, T., Kassis, K., Athanasiadis, N., Rüping, S., Grosskreutz, H., and Koutsouris, D. 2009. "Integrated Web Services Platform for the facilitation of fraud detection in health care e-government services," in *Information Technology and Applications in Biomedicine, 2009. ITAB 2009. 9th International Conference on*, IEEE, pp. 1–4 (available at http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5394355).

Tang, M., Mendis, B. S. U., Murray, D. W., Hu, Y., and Sutinen, A. 2011. "Unsupervised fraud detection in Medicare Australia," in *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*, Australian Computer Society, Inc., pp. 103–110 (available at http://dl.acm.org/citation.cfm?id=2483641).

The Henry J. Kaiser Family Foundation. 2015. "Total Medicaid Spending," *State Health Facts*, March 3 (available at http://kff.org/medicaid/state-indicator/total-medicaid-spending/; retrieved March 3, 2015).

Tsai, Y.-H., Ko, C.-H., and Lin, K.-C. 2014. "Using CommonKADS Method to Build Prototype System in Medical Insurance Fraud Detection," *Journal of Networks* (9:7) (doi: 10.4304/jnw.9.7.1798-1802).

Viveros, M. S., Nearhos, J. P., and Rothman, M. J. 1996. "Applying data mining techniques to a health insurance information system," in *VLDB*, pp. 286–294 (available at http://gbook.yolasite.com/resources/P286.PDF).

Webster, J., and Watson, R. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Quarterly* (26:2), pp. xiii – xxiii.

Weill Cornell Center for Human Rights. 2015. "MAGI and NON-MAGI Eligibility Groups," Will Cornell Center for Human Rights (available at http://wcchr.com/sites/default/files/MAGI%20criteria.pdf).

Wieringa, R. J. 2014. "Research Goals and Research Questions," in *Design Science Methodology for Information Systems and Software Engineering*, Springer, pp. 13–23 (available at http://link.springer.com/chapter/10.1007/978-3-662-43839-8_2).

Williams, G. J. 1999. "Evolutionary hot spots data mining," in *Methodologies for Knowledge Discovery and Data Mining*, Springer, pp. 184–193 (available at http://link.springer.com/chapter/10.1007/3-540-48912-6_26).

Williams, G. J., and Huang, Z. 1997. "Mining the knowledge mine," in *Advanced Topics in Artificial Intelligence*, Springer, pp. 340–348 (available at http://link.springer.com/chapter/10.1007/3-540-63797-4_87).

Yang, W.-S. 2003. "A Process Pattern Mining Framework for the Detection of Health Care Fraud and Abuse," Kaohsiung: National Sun Yat-Sen University.

Zhu, S., Wang, Y., and Wu, Y. 2011. "Health care fraud detection using nonnegative matrix factorization," in *Computer Science & Education (ICCSE), 2011 6th International Conference on*, IEEE, pp. 499–503 (available at http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6028688).

APPENDIX A – LITERATURE

Below the identified fraud categories and the related papers from the literature study are shown.

Improper coding	(Agrawal et al. 2013), (Borca 2001), (Brooks et al. 2012), (Buppert 2001), (Byrd et al. 2013), (Doan 2011), (Evans and Porche 2005), (Freeman and Loavenbruck 2001), (Hill et al. 2014), (Morris 2009), (Ogunbanjo and Knapp van Bogaert 2014), (Rashidian et al. 2012), (Rudman et al. 2009), (Sozio and Noss 2002), (Sparrow 2008), (Stanton 2001)
Phantom billing	(Agrawal et al. 2013), (Benzio 2009), (Borca 2001), (Brooks et al. 2012), (Cady 2007), (Doan 2011), (Gangopadhyay et al. 2012), (Hill et al. 2014), (Ogunbanjo and Knapp van Bogaert 2014), (Rashidian et al. 2012), (Rudman et al. 2009), (Sparrow 2008)
Kickback schemes	(Bennett and Medearis 2003), (Benzio 2009), (Borca 2001), (Brooks et al. 2012), (Byrd et al. 2013), (Cady 2007), (Carlson 2013), (Michael 2003), (Morris 2009), (Rabecs 2006), (Rashidian et al. 2012), (Sparrow 2008)
Using wrong diagnosis	(Borca 2001), (Buppert 2001), (Byrd et al. 2013), (Freeman and Loavenbruck 2001), (Hill et al. 2014), (Morris 2009), (Ogunbanjo and Knapp van Bogaert 2014), (Rashidian et al. 2012), (Sparrow 2008)
Providing unnecessary care	(Agrawal et al. 2013), (Borca 2001), (Brooks et al. 2012), (Byrd et al. 2013), (Cady 2007), (Michael 2003), (Ogunbanjo and Knapp van Bogaert 2014), (Rashidian et al. 2012), (Rudman et al. 2009)
Using unauthorized personnel	(Brooks et al. 2012), (Buppert 2001), (Byrd et al. 2013), (Doan 2011), (Morris 2009), (Rashidian et al. 2012)
Identity / EMR fraud	(Borca 2001), (Byrd et al. 2013), (Dube 2011), (Ogunbanjo and Knapp van Bogaert 2014), (Plomp and Grijpink 2011), (Rashidian et al. 2012)
Price manipulation	(Bennett and Medearis 2003), (Cady 2007), (Morris 2009),

	(Rashidian et al. 2012), (Sparrow 2008)
Unbundling	(Borca 2001), (Byrd et al. 2013), (Cady 2007), (Rashidian et al. 2012)
Service maximization and complexation	(Benzio 2009), (Byrd et al. 2013), (Morris 2009), (Ogunbanjo and Knapp van Bogaert 2014)
Using ghost employees / deceased employees	(Brooks et al. 2012), (Morris 2009)
Falsifying documents of medical necessity or prescriptions	(Byrd et al. 2013), (Rashidian et al. 2012), (Sparrow 2008)
Billing twice for the same service provided	(Byrd et al. 2013), (Morris 2009)
Misrepresenting eligibility	(Byrd et al. 2013), (Rashidian et al. 2012)
Submitting bills for deceased / ineligible people	(Sparrow 2008), (Stelfox and Redelmeier 2003)
Pinging the system, and submitting claims that are just below a threshold	(Morris 2009), (Sparrow 2008)
Waiving patient deductibles	(Borca 2001), (Freeman and Loavenbruck 2001)
Doctor shopping	(Carlson 2013), (Rashidian et al. 2012)
Submitting so many claims that it is physical impossible to do so much in one day	(Lubao 2008), (Stanton 2001)
Managed care fraud	(Sparrow 2008)
Off-label promotion of drugs	(Sparrow 2008)
False negotiation cases	(Doan 2011)
Reverse false claim cases	(Doan 2011)
Self-referral	(Rashidian et al. 2012)

Below the data mining methods for each of the identified papers in the literature study on fraud detection methods are shown.

Author	**Supervised / unsupervised**	**Method**
(Liu and Vasarhelyi 2013)	Unsupervised	Clustering
(Ekina et al. 2013)	Unsupervised	Bayesian co-clustering
(Ngufor and Wojtusiak 2013)	Hybrid supervised and unsupervised	Unsupervised data labeling and outlier detection, classification and regression
(Capelleveen 2013)	Unsupervised	Outlier detection
(Shin et al. 2012)	Supervised	Six statistical techniques - correlation analysis, logistic regression and classification tree
(Kirlidog and Asuk 2012)	Supervised	Support Vector Machine (SVM)
(Copeland et al. 2012)	Unsupervised	Visualization by histogram
(Aral et al. 2012)	Hybrid supervised and unsupervised	Distance based correlation and outlier detection
(Tang et al. 2011)	Unsupervised	Clustering, feature selection and outlier detection
(Musal 2010)	Unsupervised	Clustering algorithms, regression analysis, and various descriptive statistics
(Kumar et al. 2010)	Supervised	Support Vector Machine (SVM)
(Shan et al. 2009)	Unsupervised	Local density based outlier detection
(Shan et al. 2008)	Unsupervised	Association rules
(Liou et al. 2008)	Supervised	Logistic regression, neural network, and classification trees
(Yang 2003)	Supervised	Classification based on associations algorithm, feature selection by Markov

		blanket filter
(Ortega et al. 2006)	Supervised	Neural network
(Major and Riedinger 2002)	Hybrid supervised and unsupervised	Outlier detection and rule extraction
(He et al. 1998)	Unsupervised	Genetic algorithm and k-Nearest Neighbor clustering
(Williams 1999)	Hybrid supervised and unsupervised	Clustering and rule induction
(Williams and Huang 1997)	Hybrid supervised and unsupervised	Clustering and C5.0 classification algorithm
(He et al. n.d.)	Supervised	Neural network
(Tsai et al. 2014)	Unsupervised	Decision trees
(Suleiman et al. 2013)	Unsupervised (not data mining: upfront checking)	Decision trees
(Lin and Yeh 2012)	Unsupervised	Multiple models; rule-based
(Konijn and Kowalczyk 2011)	Unsupervised	Outlier detection, using multiple statistic methods (rank-based; weighted rank-based, binomial distribution, mean of the outlier score)
(Lu and Boritz 2005)	Unsupervised	Adaptive Benford Algorithm
(Peng et al. 2006)	Unsupervised	Two clustering methods: SAS EM and CLUTO
(Chen and Gangopadhyay 2013)	Unsupervised	A community detection algorithm through spectral analysis
(Tagaris et al. 2009)	Hybrid supervised and unsupervised	Multiple models: subgroup discovery, similarity learning and RapidMiner
(Zhu et al. 2011)	Unsupervised	Clustering using Non-negativeMatrix Factorization

(Gangopadhyay et al. 2012)	Unsupervised	Pagerank algorithm
(Ormerod et al. 2003)	Supervised	Bayesian Network
(Viveros et al. 1996)	Supervised	Association rules and a neural segmentation algorithm
(Phua et al. 2004)	Supervised	backpropagation (BP), together with naive Bayesian (NB) and C4.5 algorithms
(Bonchi et al. 1999)	Supervised	Decision trees
(Yang 2003)	Supervised	C4.5 decision tree algorithm

APPENDIX B – QUERY USED FOR DATASET CREATION

In this appendix the SQL query used for dataset generation is given. After executing the query the dataset was fine-tuned by PHP. For example a column was added stating the application was completed within 25 days, within 50 days and so on. However, the SQL query below is used to get the data out of the source system (Cúram).

```
SELECT
  DISTINCT AC.applicationCaseID as ApplicationCaseId,
  CPR.NUMBEROFAPPPLICANTS AS NumberOfApplicants,
  AC.METHODOFAPPLICATION AS MethodOfApplication,
  (PAD.AUTHORISATIONDATE - AC.APPLICATIONDATE) AS NumberOfDays,
  (PAD.AUTHORISATIONDATE-TRUNC(AC.SUBMITTEDDATETIME)) AS SystemNumberOfDays,
  MMEDD.PRODUCTNAME AS ProgramType,
  VER.NumberOfEvidence AS NumberOfEvidence,
  VER."EvidenceTypes" AS EvidenceTypes,
  MMEDD.ELIGIBILITYSTATUS as CaseworkerDecision,
  MMEDD.PaymentCategory as PaymentCategory,
  PAD.authorisedBy as caseworker,
  CH.caseid as CaseId,
  CR.prefCommMethod as prefCommMethod,
  CR.preferredLanguage as preferredLanguage,
  PAD.AUTHORISATIONDATE as DecisionDate,
  TRUNC(AC.SUBMITTEDDATETIME) as ApplicationDate
FROM
  APPLICATIONCASE AC
LEFT JOIN
  PROGRAMAUTHORISATIONDATA PAD ON PAD.APPLICATIONCASEID =
AC.APPLICATIONCASEID
LEFT JOIN
  CASEHEADER CH ON CH.INTEGRATEDCASEID = PAD.INTEGRATEDCASEID
LEFT JOIN
```

```
  (
    SELECT CASEID,COUNT(distinct participantRoleID) as NUMBEROFAPPPLICANTS
    FROM CASEPARTICIPANTROLE
    WHERE TYPECODE='MEM' OR TYPECODE='NOM'  OR TYPECODE = 'PRI'
    GROUP BY CASEID
  ) CPR
  ON CPR.CASEID=CH.CASEID
LEFT JOIN
  (
    SELECT
      CASEID,
      COUNT(1) as NumberOfEvidence,
      listagg (EVIDENCETYPE, '|')
    WITHIN GROUP
      (ORDER BY EVIDENCETYPE) "EvidenceTypes"
    FROM
    (
      SELECT  ED.caseid, VDI.EVIDENCETYPE
      FROM
      VDIEDLINK VLINK
      LEFT JOIN
        VERIFIABLEDATAITEM VDI
        ON
          VLINK.VERIFIABLEDATAITEMID=VDI.VERIFIABLEDATAITEMID
      LEFT JOIN
        EVIDENCEDESCRIPTOR ED
        ON
          VLINK.evidenceDescriptorID = ED.EVIDENCEDESCRIPTORID
    )
    GROUP BY CASEID
  ) VER
  ON
    VER.caseid = CH.INTEGRATEDCASEID
LEFT JOIN
  MMELIGIBILITYDECISIONDETAILS MMEDD
```

```
  ON
    MMEDD.concernroleid = CH.concernRoleID
LEFT JOIN
  CONCERNROLE CR
  ON
    CR.CONCERNROLEID = CH.CONCERNROLEID
WHERE     PAD.authorisedBy != 'SYSTEM' AND PAD.AUTHORISATIONDATE IS NOT NULL
GROUP BY AC.applicationCaseID, CPR.NUMBEROFAPPPLICANTS, AC.METHODOFAPPLICATION,
(PAD.AUTHORISATIONDATE - AC.APPLICATIONDATE),
MMEDD.PRODUCTNAME, VER.NumberOfEvidence, VER."EvidenceTypes",
MMEDD.ELIGIBILITYSTATUS,
MMEDD.PaymentCategory,  PAD.authorisedBy,
CH.caseid,(PAD.AUTHORISATIONDATE-
TRUNC(AC.SUBMITTEDDATETIME)),CR.prefCommMethod,CR.preferredLanguage,PAD.AUTHORISA
TIONDATE,TRUNC(AC.SUBMITTEDDATETIME);
```

APPENDIX C – DATA MINING RESULTS

Three data mining algorithms are applied on several versions of the generated dataset. The table below shows the confusion matrix for each algorithm and each version of the dataset used, combined with the accuracy for each prediction class. Both the exact prediction as the 1-away prediction are given.

All variables known at moment of submission, all application methods							
Neural network		Predicted					
Accuracy	0,4572	0	1	2	3	Exact	1-away
Actual	0	111	21	2	1	0.8222	0.9778
	1	50	40	5	40	0.2963	0.7037
	2	38	27	4	66	0.0296	0.7185
	3	19	17	6	91	0.6842	0.7293
SVM		Predicted					
Accuracy	0,4222	0	1	2	3	Exact	1-away
Actual	0	99	35	0	1	0.7333	0.9926
	1	61	35	0	39	0.2493	0.7111
	2	38	29	1	67	0.7407	0.7185
	3	19	23	0	93	0.6889	0.6889
Random forest		Predicted					
Accuracy	0,4492	0	1	2	3	Exact	1-away
Actual	0	415	221	25	18	0,6112	0,9367
	1	125	333	63	158	0,4904	0,7673
	2	80	277	59	263	0,0869	0,8822
	3	49	150	67	413	0,6082	0,7069
All variables known at moment of submission, only online applications							
Neural network		Predicted					
Accuracy	0,3066	0	1	2	3	Exact	1-away
Actual	0	14	16	9	14	0,2642	0,5660
	1	8	22	12	11	0,4151	0,7925
	2	9	16	13	15	0,2453	0,8302
	3	10	12	15	16	0,3019	0,5849

SVM		Predicted					
Accuracy	0,2736	0	1	2	3	Exact	1-away
Actual	0	11	20	14	8	0,2075	0,5849
	1	12	18	7	16	0,3396	0,6981
	2	4	23	7	19	0,1321	0,9245
	3	4	25	2	22	0,4151	0,4528
Random forest		Predicted					
Accuracy	0,2800	0	1	2	3	Exact	1-away
Actual	0	99	68	34	66	0,3708	0,6255
	1	71	72	40	84	0,2697	0,6854
	2	76	51	30	110	0,1124	0,7154
	3	54	63	52	98	0,3670	0,5618
All variables known at moment of submission, only paper applications							
Neural network		Predicted					
Accuracy	0,3630	0	1	2	3	Exact	1-away
Actual	0	47	7	2	17	0,6438	0,7397
	1	22	16	6	29	0,2192	0,6027
	2	16	14	8	35	0,1096	0,7808
	3	19	14	5	35	0,4795	0,5479
SVM		Predicted					
Accuracy	0,3425	0	1	2	3	Exact	1-away
Actual	0	44	9	2	18	0,6027	0,7260
	1	24	6	5	38	0,0822	0,4795
	2	19	1	10	43	0,1370	0,7397
	3	19	8	6	40	0,5479	0,6301
Random forest		Predicted					
Accuracy	0,3753	0	1	2	3	Exact	1-away
Actual	0	218	37	51	61	0,5940	0,6948
	1	69	78	87	133	0,2125	0,6376
	2	53	62	99	153	0,2698	0,8556
	3	55	59	97	156	0,4251	0,6894
All data from dataset, all application methods							
Neural network		Predicted					
Accuracy	0,4322	0	1	2	3	Exact	1-away

Actual	0	85	34	2	8	0,6589	0,9225
	1	41	49	3	36	0,3798	0,7209
	2	16	42	4	67	0,0310	0,8760
	3	11	25	8	85	0,6589	0,7209
SVM		Predicted					
Accuracy:	0,4012	0	1	2	3	Exact	1-away
Actual	0	97	27	1	4	0,7519	0,9612
	1	61	37	5	26	0,2868	0,7984
	2	42	43	3	41	0,0233	0,6744
	3	27	26	6	70	0,5426	0,5891
Random forest		Predicted					
Accuracy	0,4791	0	1	2	3	Exact	1-away
Actual	0	440	123	64	19	0,6811	0,8715
	1	151	248	145	102	0,3839	0,8421
	2	104	206	166	170	0,2570	0,8390
	3	49	117	96	384	0,5944	0,7430
All data from dataset, only online applications							
Neural network		Predicted					
Accuracy	0,2760	0	1	2	3	Exact	1-away
Actual	0	3	3	2	40	0,0625	0,1250
	1	7	4	2	35	0,0833	0,2708
	2	3	2	0	43	0,0000	0,9375
	3	1	1	0	46	0,9583	0,9583
SVM		Predicted					
Accuracy	0,3125	0	1	2	3	Exact	1-away
Actual	0	13	12	6	17	0,2708	0,5208
	1	14	11	5	18	0,2292	0,6250
	2	10	10	8	20	0,1667	0,7917
	3	9	5	6	28	0,5833	0,7083
Random forest		Predicted					
Accuracy	0,3620	0	1	2	3	Exact	1-away
Actual	0	85	65	40	51	0,3527	0,6224
	1	71	77	45	48	0,3195	0,8008
	2	66	49	54	72	0,2241	0,7261

	3	42	30	36	133	0,5519	0,7012
All data from dataset, only paper applications							
Neural network		Predicted					
Accuracy	0,3571	0	1	2	3	Exact	1-away
Actual	0	34	18	10	8	0,4857	0,7429
	1	12	24	20	14	0,3429	0,8000
	2	19	22	7	22	0,1000	0,7286
	3	8	20	7	35	0,5000	0,6000
SVM		Predicted					
Accuracy	0,3714	0	1	2	3	Exact	1-away
Actual	0	36	17	10	7	0,5143	0,7571
	1	12	13	14	31	0,1857	0,5571
	2	17	12	18	23	0,2571	0,7571
	3	13	5	15	37	0,5286	0,7429
Random forest		Predicted					
Accuracy	0,3793	0	1	2	3	Exact	1-away
Actual	0	189	73	55	35	0,5369	0,7443
	1	70	84	94	104	0,2386	0,7045
	2	63	73	96	120	0,2727	0,8210
	3	46	67	74	165	0,4688	0,6790

Sub measurements

All data from dataset, excluded method of application							
Neural network		Predicted					
Accuracy	0,3852	0	1	2	3	Exact	1-away
Actual	0	77	38	11	9	0,5704	0,8519
	1	35	48	23	29	0,3556	0,7852
	2	18	56	32	29	0,2370	0,8667
	3	23	36	25	51	0,3778	0,5630
SVM		Predicted					
Accuracy	0,3685	0	1	2	3	Exact	1-away
Actual	0	97	24	1	13	0,7185	0,8963
	1	56	43	0	36	0,3185	0,7333
	2	47	37	0	51	0,0000	0,6519

	3	46	30	0	59	0,4370	0,4370
Random forest		Predicted					
Accuracy	0,3833	0	1	2	3	Exact	1-away
Actual	0	331	211	76	61	0,4875	0,7982
	1	120	312	116	131	0,4595	0,8071
	2	88	281	125	185	0,1841	0,8704
	3	78	207	121	273	0,4021	0,5803
All data from dataset, excluded missing evidence							
Neural network		Predicted					
Accuracy	0,4426	0	1	2	3	Exact	1-away
Actual	0	113	19	0	3	0,8370	0,9778
	1	64	33	1	37	0,2444	0,7259
	2	40	37	0	58	0,0000	0,7037
	3	29	13	0	93	0,6889	0,6889
SVM		Predicted					
Accuracy	0,4333	0	1	2	3	Exact	1-away
Actual	0	112	18	2	3	0,8296	0,9630
	1	66	25	7	37	0,1852	0,7259
	2	39	34	4	58	0,0296	0,7111
	3	30	11	1	93	0,6889	0,6963
Random forest		Predicted					
Accuracy	0,4190	0	1	2	3	Exact	1-away
Actual	0	463	183	14	19	0,6819	0,9514
	1	284	197	63	135	0,2901	0,8012
	2	171	186	61	261	0,0898	0,7482
	3	110	83	69	417	0,6141	0,7158
All data from dataset, excluded number of people							
Neural network		Predicted					
Accuracy	0,4870	0	1	2	3	Exact	1-away
Actual	0	104	25	0	6	0,7704	0,9556
	1	33	60	0	42	0,4444	0,6889
	2	32	44	3	56	0,0222	0,7630
	3	14	24	1	96	0,7111	0,7185
SVM		Predicted					

Accuracy	0,4389	0	1	2	3	Exact	1-away
Actual	0	102	28	0	5	0,7556	0,9630
	1	56	38	0	41	0,2815	0,6963
	2	52	28	1	54	0,0074	0,6148
	3	24	15	0	96	0,7111	0,7111
Random forest		Predicted					
Accuracy	0,4529	0	1	2	3	Exact	1-away
Actual	0	409	229	22	19	0,6024	0,9396
	1	107	337	73	162	0,4963	0,7614
	2	79	286	79	235	0,1163	0,8837
	3	51	158	65	405	0,5965	0,6922
All data from dataset, excluded preferred language							
Neural network		Predicted					
Accuracy	0,4778	0	1	2	3	Exact	1-away
Actual	0	94	27	9	5	0,6963	0,8963
	1	32	53	4	46	0,3926	0,6593
	2	22	44	11	58	0,0815	0,8370
	3	10	19	6	100	0,7407	0,7852
SVM		Predicted					
Accuracy	0,4130	0	1	2	3	Exact	1-away
Actual	0	101	28	0	6	0,7481	0,9556
	1	71	18	0	46	0,1333	0,6593
	2	52	21	0	62	0,0000	0,6148
	3	23	8	0	104	0,7704	0,7704
Random forest		Predicted					
Accuracy	0,5109	0	1	2	3	Exact	1-away
Actual	0	407	238	14	20	0,5994	0,9499
	1	101	636	43	172	0,6681	0,8193
	2	78	294	70	237	0,1031	0,8851
	3	44	171	50	414	0,6097	0,6834
All data from dataset, excluded preferred communication method							
Neural network		Predicted					
Accuracy	0,4481	0	1	2	3	Exact	1-away
Actual	0	112	22	0	1	0,8296	0,9926

	1	59	41	2	33	0,3037	0,7556
	2	41	39	4	51	0,0296	0,6963
	3	35	15	0	85	0,6296	0,6296
SVM		Predicted					
Accuracy	0,4296	0	1	2	3	Exact	1-away
Actual	0	109	25	1	0	0,8074	0,9926
	1	63	38	1	33	0,2815	0,7556
	2	43	39	2	51	0,0148	0,6815
	3	35	14	3	83	0,6148	0,6370
Random forest		Predicted					
Accuracy	0,4470	0	1	2	3	Exact	1-away
Actual	0	403	246	20	10	0,5935	0,9558
	1	142	336	40	161	0,4948	0,7629
	2	103	259	73	244	0,1075	0,8483
	3	50	157	70	402	0,5920	0,6951
All data from dataset, and the final program determination							
Neural network		Predicted					
Accuracy	0,4683	0	1	2	3	Exact	1-away
Actual	0	104	23	0	7	0,7761	0,9478
	1	38	49	2	45	0,3657	0,6642
	2	34	37	2	61	0,0149	0,7463
	3	15	22	1	96	0,7164	0,7239
SVM		Predicted					
Accuracy	0,4646	0	1	2	3	Exact	1-away
Actual	0	99	28	1	6	0,7388	0,9478
	1	35	53	0	46	0,3955	0,6567
	2	32	42	1	59	0,0075	0,7612
	3	17	21	0	96	0,7164	0,7164
Random forest		Predicted					
Accuracy	0,4549	0	1	2	3	Exact	1-away
Actual	0	459	170	20	21	0,6851	0,9388
	1	167	283	65	155	0,4224	0,7687
	2	112	241	73	244	0,1090	0,8328
	3	60	136	70	404	0,6030	0,7075

All data from dataset, and the eligibility decision							
Neural network		Predicted					
Accuracy	0,4500	0	1	2	3	Exact	1-away
Actual	0	104	26	2	3	0,7704	0,9630
	1	46	45	7	37	0,3333	0,7259
	2	28	37	4	66	0,0296	0,7926
	3	26	16	3	90	0,6667	0,6889
SVM		Predicted					
Accuracy	0,4407	0	1	2	3	Exact	1-away
Actual	0	98	32	4	1	0,7259	0,9630
	1	47	48	3	37	0,3556	0,7259
	2	31	38	3	63	0,0222	0,7704
	3	26	17	3	89	0,6593	0,6815
Random forest		Predicted					
Accuracy	0,4503	0	1	2	3	Exact	1-away
Actual	0	431	196	38	14	0,6348	0,9234
	1	133	308	83	155	0,4536	0,7717
	2	97	257	73	252	0,1075	0,8571
	3	48	145	75	411	0,6053	0,7158
All data from dataset, and the payment category							
Neural network		Predicted					
Accuracy	0,4468	0	1	2	3	Exact	1-away
Actual	0	97	28	2	2	0,7519	0,9690
	1	48	40	3	38	0,3101	0,7054
	2	32	49	5	53	0,0360	0,7698
	3	11	21	4	93	0,7209	0,7519
SVM		Predicted					
Accuracy	0,4123	0	1	2	3	Exact	1-away
Actual	0	82	37	7	3	0,6357	0,9225
	1	41	42	5	41	0,3256	0,6822
	2	48	42	3	56	0,0201	0,6779
	3	12	21	2	94	0,7287	0,7442
Random forest		Predicted					
Accuracy	0,4675	0	1	2	3	Exact	1-away

Actual	0	451	151	35	9	0,6981	0,9319
	1	146	265	116	119	0,4102	0,8158
	2	80	226	127	213	0,1966	0,8762
	3	56	126	99	365	0,5650	0,7183