Data Mining for Description and Prediction of Antibiotic Treated Healthcare-Associated Infections

EMMY DAMBERG

Master of Science Thesis in Medical Engineering
Stockholm 2014
This master thesis project was performed in collaboration with Mawell Scandinavia AB
Supervisors at Mawell Scandinavia AB: Eva Biberg & Torbjörn Dahlin

Data Mining for Description and Prediction of Antibiotic Treated Healthcare-Associated Infections

EMMY DAMBERG

Master of Science Thesis in Medical Engineering
Advanced level (second cycle), 30 credits
Supervisor at KTH: Heikki Teriö
Reviewer: Altug Akay
Examiner: Mats Nilsson
School of Technology and Health
TRITA-STH. EX 20XX:XX

Royal Institute of Technology
KTH STH
SE-141 86 Flemingsberg, Sweden
http://www.kth.se/sth
Abstract

Healthcare-associated infections is the most common healthcare related injury and affect almost every tenth patient. With the purpose of reducing these infections Infektionsverktyget was developed for registration and feedback of infection data. The tool is now used in all Swedish county councils resulting in a wealth of data. The purpose of this thesis was thus to investigate how data mining can be applied to describe patterns in this data and predict patient outcomes regarding healthcare-associated infections that need to be treated with antibiotics.

Data mining was performed with Microsoft SQL Server 2008 in which models based on six different data mining algorithms with different parameter settings were developed. They used the attributes gender, age and previous diagnoses and medical actions as inputs and antibiotic treated healthcare-associated infection outcome as output. The predictive performance of the models was evaluated using 5-fold cross validation and macro averaged measures of recall, precision and F-measure. Patterns generated by selected models were extracted.

Models based on the Naive Bayes algorithm showed the highest predictive capabilities with respect to recall and models based on the Decision Trees algorithm with low pruning had the highest precision. Although, none were considered to perform sufficiently well and several areas of improvement were identified. The most important factor in the inadequate performance is believed to be the relatively rare occurrences of infections in the dataset. Extracted patterns based on the Association Rules algorithm were considered the easiest to interpret. Patterns included clinically valid and invalid as well as trivial relationships.

Future studies should be focused on further model improvements and gathering of more patient data. The idea is that data mining in Infektionsverktyget in the future could be used both to provide ideas for further medical research and to identify risk patients and prevent healthcare-associated infections in daily clinical work.
Sammanfattning

Vårdrelaterade infektioner är den vanligaste vårdskadan och drabbar nästan var tioende patient. Med syfte att minska antalet vårdrelaterade infektioner utvecklades Infektionsverktyget för registrering och återkoppling av infektionsdata. Verktyget används nu i alla Sveriges landsting vilket resulterar i stora mängder data. Syftet med detta examensarbete var därför att undersöka hur data mining kan användas för att beskriva mönster i denna data och för att förutsäga om patienter kommer att drabbas av en vårdrelaterad infektion som behöver antibiotikabehandlas.


Framtida studier bör fokuseras på att förbättra modellerna ytterligare och att samla in mer patientdata. Idén är att data mining i Infektionsverktyget i framtiden skulle kunna användas för att ge uppslag till medicinsk forskning och för att identifiera riskpatienter och därmed förebygga vårdrelaterade infektioner i den dagliga kliniska verksamheten.
Acknowledgements

This master thesis has been performed at KTH Royal Institute of Technology, School of Technology and Health, in collaboration with Mawell Scandinavia AB.

For accepting me as a thesis student and for guiding and helping me during the course of the project I would like to thank Eva Biberg and Torbjörn Dahlin, my supervisors at Mawell. I also want to thank Rikard Lövström and Ann-Sofie Geschwindt for taking their time to contribute to the project.

Thanks to my internal supervisor Heikki Teriö for guidance and providing new perspectives on struggles I encountered.

To my friends and family I would like to say thank you for bearing with me and supporting me during both the ups and downs of this project.

Emmy Damberg
Stockholm 2014-08-22
List of Abbreviations

CDI  Clostridium difficile infection
CeHis Centrum för eHälsa i Samverkan
DPI Deep postoperative wound infection
FN False Negative
FP False Positive
HAI Healthcare-associated infection
ICD-10-SE International Statistical Classification of Diseases and Related Health Problems, tenth revision, Swedish version
KVÅ Klassifikation av Vårdåtgärder
SITHS Säker IT i Hälso- och Sjukvården
SKL Sveriges Kommuner och Landsting
SPI Superficial postoperative wound infection
TN True Negative
TP True Positive
UTI Urinary tract infection
# Contents

Abstract \hspace{1cm} v

Sammanfattning \hspace{1cm} vii

Acknowledgements \hspace{1cm} ix

List of Abbreviations \hspace{1cm} xi

Contents \hspace{1cm} xiii

List of Figures \hspace{1cm} xv

List of Tables \hspace{1cm} xvii

1 Introduction \hspace{1cm} 1

1.1 Project objective and definition \hspace{1cm} 2

1.2 Demarcations \hspace{1cm} 2

2 Theoretical background \hspace{1cm} 3

2.1 Healthcare-associated infections \hspace{1cm} 3

2.2 Infektionsverktyget \hspace{1cm} 4

\hspace{1cm} 2.2.1 ASP.NET \hspace{1cm} 6

2.3 Legal aspects \hspace{1cm} 6

2.4 Data mining \hspace{1cm} 7

\hspace{1cm} 2.4.1 Data mining cycle \hspace{1cm} 8

\hspace{1cm} 2.4.2 Structure and model \hspace{1cm} 8

\hspace{1cm} 2.4.3 Data mining tasks and algorithms \hspace{1cm} 9

\hspace{1cm} 2.4.4 Cross validation and holdout method \hspace{1cm} 12

\hspace{1cm} 2.4.5 Classification matrix and evaluation measures \hspace{1cm} 13

\hspace{1cm} 2.4.6 Medical data mining \hspace{1cm} 14

2.5 Previous studies \hspace{1cm} 15

3 Method \hspace{1cm} 17

3.1 Data mining environment \hspace{1cm} 17

3.2 Data mining cycle \hspace{1cm} 18
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.1</td>
<td>Problem formation</td>
<td>18</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Data preprocessing</td>
<td>19</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Model building</td>
<td>26</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Model evaluation</td>
<td>31</td>
</tr>
<tr>
<td>3.3</td>
<td>Pattern extraction</td>
<td>33</td>
</tr>
<tr>
<td>3.4</td>
<td>Web application development</td>
<td>33</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Environment</td>
<td>34</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Login</td>
<td>34</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Analysis Services connection</td>
<td>35</td>
</tr>
<tr>
<td>3.4.4</td>
<td>CrossValidation page</td>
<td>35</td>
</tr>
<tr>
<td>3.4.5</td>
<td>PatternExtraction page</td>
<td>36</td>
</tr>
<tr>
<td>4</td>
<td>Result</td>
<td>39</td>
</tr>
<tr>
<td>4.1</td>
<td>Metadata</td>
<td>39</td>
</tr>
<tr>
<td>4.2</td>
<td>Quantitative evaluation</td>
<td>39</td>
</tr>
<tr>
<td>4.3</td>
<td>Extracted patterns</td>
<td>40</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Association Rules algorithm</td>
<td>41</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Naive Bayes algorithm</td>
<td>41</td>
</tr>
<tr>
<td>5</td>
<td>Discussion</td>
<td>47</td>
</tr>
<tr>
<td>5.1</td>
<td>Metadata</td>
<td>47</td>
</tr>
<tr>
<td>5.2</td>
<td>Quantitative evaluation</td>
<td>48</td>
</tr>
<tr>
<td>5.3</td>
<td>Extracted patterns</td>
<td>49</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Association Rules</td>
<td>49</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Naive Bayes</td>
<td>50</td>
</tr>
<tr>
<td>5.4</td>
<td>Improvements and future work</td>
<td>51</td>
</tr>
<tr>
<td>6</td>
<td>Conclusion</td>
<td>53</td>
</tr>
<tr>
<td>Bibliography</td>
<td></td>
<td>55</td>
</tr>
<tr>
<td>A</td>
<td>ICD-10-SE hierarchy</td>
<td>59</td>
</tr>
<tr>
<td>B</td>
<td>KVA hierarchy</td>
<td>61</td>
</tr>
</tbody>
</table>
List of Figures

2.1  Schematic of Neural Network and Logistic Regression. . . . . . 11
3.1  Nested table setup with example data. . . . . . . . . . . . . . . 21
3.2  Database diagram over tables and columns used. . . . . . . . . 23
3.3  Screenshot of the CrossValidation page. . . . . . . . . . . . . . 36
3.4  UML diagram of the CrossValidation page. . . . . . . . . . . . . 37
3.5  Screenshot of the PatternExtraction page. . . . . . . . . . . . . 37
3.6  UML diagram of the PatternExtraction page. . . . . . . . . . . . 38
A.1  Extract from the ICD-10-SE hierarchy tree. . . . . . . . . . . . 59
B.1  Extract from the KVÅ hierarchy tree. . . . . . . . . . . . . . . 61
# List of Tables

2.1 Classification matrix. ............................................. 13  
2.2 Formulas and purposes of the evaluation measures. ............... 14  

3.1 The number of action and diagnosis terms. ....................... 28  
3.2 Models used in round one of quantitative evaluation. ............ 29  
3.3 Models used in round two of quantitative evaluation. ............ 30  
3.4 Formulas for micro and macro averaging in multi-state classification. ......................................................... 33  
3.5 The models used for pattern extraction. .......................... 33  

4.1 The number of care events included. ............................ 39  
4.2 Distribution of infection outcomes. .............................. 40  
4.3 Results of cross validation for all clinics. ....................... 40  
4.4 Results of cross validation for women’s health clinic. .......... 41  
4.5 Examples of rules extracted for all clinics. ..................... 42  
4.6 Examples of rules extracted for women’s health clinic. .......... 43  
4.7 Examples of attribute characteristics extracted for all clinics. 44  
4.8 Examples of attribute characteristics extracted for women’s health clinic. ......................................................... 45
Chapter 1

Introduction

Healthcare-associated infections (HAIs) affect approximately every tenth hospitalized patient in Sweden [1], with the consequence of increased human suffer, increased costs as a result of prolonged hospital stays and even deaths [2].

Infektionsverktyget, The Anti-Infection Tool, is a Swedish eHealth tool for consistent registration, storage and feedback of data concerning HAIs and prescriptions of antibiotics. It is developed by Mawell on behalf of Sveriges Kommuner och Landsting (SKL, Swedish Association of Local Authorities and Regions) and Centrum för eHälsa i Samverkan (CeHis, Center for eHealth in Cooperation). The purpose of the tool is to reduce the number of HAIs and to decrease the number of incorrect antibiotic prescriptions in order to prevent bacteria resistance to antibiotics. The aim is increased patient safety and a more efficient use of the hospital resources. Infektionsverktyget is used in hospitals in all of the county councils and this results in enormous amounts of data that are today only used for simple follow-ups of the past.

An idea was that this data could be analyzed with data mining, which is the process of discovering patterns in large sets of data. If successful, it could lead to the discovery of interesting relationships in healthcare that are harder to find with regular statistical data analyses. Data mining could also be used to identify patients with an increased risk of incurring a HAI that needs to be treated with antibiotics.

Data mining in Infektionsverktyget has been somewhat studied before, but only as proof of concept with made up test data and a narrow focus on the already known relationship between urinary tract infections and the risk factor urine catheter [3]. This study aims at being a broader exploratory study which for the first time uses real patient data and takes more aspects into account.
1.1 Project objective and definition

The objective of the project was to investigate how data mining can be applied to describe patterns in data registered in Infektionsverktyget and predict patient outcomes regarding antibiotic treated HAIs.

The project consisted of the following parts:

- Development of data mining models for analysis of relationships between patient gender, age, diagnoses, medical actions and healthcare-associated infection outcome using data registered in Infektionsverktyget.

- Development of a web application extension to Infektionsverktyget for evaluation of models with real patient data.

- Quantitative evaluation of the predictive performance of the developed models.

- Extraction of patterns generated by selected models.

1.2 Demarcations

The study used data registered in Infektionsverktyget regarding cases of HAIs that were treated with antibiotics. Only these are registered because the registration occurs in connection with the prescription of antibiotics. Data regarding antibiotics and bacterial resistance to antibiotics was not considered, nor was data concerning community acquired infections. Only patient data registered in Uppsala County Council was used.

The HAIs that were taken into account were the ones that are registered in Infektionsverktyget; urinary tract infection (UTI), pneumonia, sepsis, superficial postoperative wound infection (SPI), deep postoperative wound infection (DPI) and Clostridium difficile infection (CDI).

HAIs that do not belong to any of these categories and are registered as “another type of HAI” were not considered.

The legal restrictions presented are valid for Sweden. No other countries were considered because Infektionsverktyget is currently only used in Sweden.
Chapter 2

Theoretical background

2.1 Healthcare-associated infections

A healthcare-associated infection is a type of injury that a patient incurs when in contact with healthcare. According to Socialstyrelsen, The National Board of Health and Welfare, the definition of a HAI is (translated from Swedish):

“An infection that appears in a patient during inpatient care or as a result of a medical action in the form of diagnostics, treatment or care in other healthcare or social care instances, or that staff working in healthcare and social care incur due to their exercise of profession.” [4]

An infection is classified as healthcare-associated if the symptoms debut 48 hours or more after admission to hospital, or if it is related to any medical action affecting the patient independently of debut time [4]. This definition does not take into account whether the pathogenic substance originated from the healthcare or from the patient [4].

According to a study conducted by SKL the average point prevalence of HAIs in Swedish hospitals was 8.9% during spring 2014 [1]. The number varied between 4.8 and 12.9 depending on county council. For Uppsala County Council the point prevalence was 11.9% [1]. A healthcare records review conducted by Socialstyrelsen in 2013 showed that 40% of healthcare-associated injuries are HAIs which makes it the most common healthcare-associated injury [5]. The prevalence of HAIs varies between hospitals of different sizes and between different specializations within the same hospital. The reason is that some specializations deal more frequently with certain risk factors for HAIs such as central vein catheters, urinary catheters and endotracheal tubes [2]. Studies have shown that the prevalence of HAIs is relatively equal among industrialized countries with similar economical and medical standards [2].
Patients with reduced immune system functionality are more likely to be affected by HAIs. Factors that influence the immune system are [2]:

- **Age.** Premature infants and elderly are generally immunodeficient.

- **Health and trauma.** Major surgeries or traumas, some malignant diseases and for instance malnutrition and uncontrolled diabetes affect the immune system negatively.

- **Medical treatments.** There is a wide variety of treatments that can increase the risk of HAIs, for instance implantations, treatments with antibiotics and immunosuppressive treatments such as chemotherapy.

HAIs not only cause human suffering and is one of the most serious threats to patient safety but also often prolongs the hospital stay of the patient which decreases the capacity and throughput of patients in the hospital. This leads to increased costs and fewer beds available for patients who need care. It is estimated that HAIs in Swedish hospitals cause 500 000 extra hospital days yearly (in average 4 extra days per patient suffering from a HAI) resulting in an additional cost of approximately 3.7 billion SEK per year [2].

The most common transmission route for HAIs is via the hands of healthcare professionals [2]. Therefore the most important hygiene care is for the healthcare staff to comply with the basal hygienic routines. These together with other routines to increase the patient safety when it comes to HAIs should be included in the patient safety and quality management system established by each healthcare provider. The Code of Statutes 2011:9 of Socialstyrelsen states how this systematic improvement work should be carried out and mentions self-monitoring as one important factor [6].

Studies have shown that 20-30% of all HAIs are possible to prevent with optimal routines for infection control and systematic improvement work. The greatest improvement potential lies in the areas of intensive care, lung infections and infections connected to catheters in the urinary tract and central vessels [2].

### 2.2 Infektionsverktyget

Apart from following the basal hygiene and clothing routines, studies have shown that registration of infections can contribute to decreasing the prevalence of HAIs [2]. Registration is a part of the important self-monitoring activity because the registered information can be used as feedback for the systematic quality improvement work.

SKL, together with all the local authorities and regions, focuses on patient safety, including reducing the prevalence of HAIs. The goal is to halve the
prevalence of HAIs in inpatient care, from 10% in 2006 to 5%, through systematic quality improvement work [7].

With this goal in mind the need of a national tool for documentation and feedback of information regarding HAIs and prescriptions of antibiotics was identified. It resulted in the eHealth tool Infektionsverktyget which was developed by Mawell on behalf of SKL and CeHis during 2010 and 2011. It was first pilot tested in two regions and was thereafter introduced in all Swedish county councils [7].

The main purpose of Infektionsverktyget is to be used as a tool to improve the prevention of HAIs through increased awareness of risk factors, problems and potential areas for improvement of each healthcare provider. In the longer term it is thought to be relevant for achieving a more rational use of antibiotics to reduce bacterial resistance to antibiotics [7].

Six types of HAIs are registered in Infektionsverktyget:

- Pneumonia
- Urinary tract infection
- Clostridium difficile infection
- Superficial postoperative wound infection
- Deep postoperative wound infection
- Sepsis

The registration takes place during the electronic prescription of antibiotics. The responsible physician indicates whether the reason for the prescription is a healthcare-associated infection, a community-acquired infection or a prophylactic treatment. If the reason is a HAI, the type of infection (of the six above, or “another type of HAI”) is also stated before the prescription is finished. Additional information about the patient such as gender, year of birth, previous diagnoses and medical actions, laboratory test results and risk factors is acquired from the electronic healthcare record.

The information contained in Infektionsverktyget includes the following:

- Cause of prescriptions of antibiotics, i.e. the type of infection/treatment that caused the prescription, including the type of antibiotics prescribed, the date of prescription and the responsible care unit.

- Laboratory test results indicating incidence of the Clostridium difficile bacterium, including the date of the test and the care unit ordering the test.

- Location of patients with respect to care unit, including dates of admission to and discharge from the care unit.
2.3. LEGAL ASPECTS

• Medical actions performed on patients, including the type of activity, the date it was performed and the responsible care unit.

• Diagnoses of patients, including the type of diagnosis, the date that the diagnosis was registered and the responsible care unit.

Together with this information an ID unique to each patient is always stored and through this the personal information such as gender and birth month and year can be retrieved. However, the real social security number for identification is not present in any of this information because it is exchanged with the made up patient ID, i.e. the information is pseudonymized. The real social security number can only be retrieved from a separate table via a pseudo ID.

All this information serves as a base for reports that each healthcare provider can use for their own follow-up. The reports are rendered via the Infektionsverktyget web application to which each healthcare provider can login with a SITHS card to produce reports based on their own data. SITHS (Säker IT i Hälso- och Sjukvården - Secure IT in Healthcare) cards are used in healthcare for secure digital and physical identification of healthcare personnel. The reports can be arbitrarily focused, filtered and ordered to adapt to the needs of each particular healthcare provider [7]. Due to present regulations, each healthcare provider can only use their own data for follow-up and data exchange between different healthcare providers is not allowed [8].

The Infektionsverktyget database is managed in Microsoft SQL Server 2008 and the Infektionsverktyget reporting tool is developed in Visual Studio with ASP.NET.

2.2.1 ASP.NET

ASP.NET is a framework for creation of dynamic web pages, i.e. web pages that can be programmed to response to different kinds of events. During development there is one file with the extension .aspx that contains static HTML code as well as web controls such as buttons, text boxes, labels and more advanced controls. There is also a file with the extension .aspx.cs which contains the code behind the HTML which makes the web page dynamic. Events triggered by for instance a button defined in the static HTML file are handled in a method in the code behind.

2.3 Legal aspects

In the legal sense the patient data that a healthcare provider registers in Infektionsverktyget is a responsibility of the healthcare provider although it is stored centrally on servers administered by Mawell.
For Mawell to be allowed to handle the patient data on behalf of the healthcare provider, a written agreement between each healthcare provider and Mawell has been established in accordance with the Personal Data Act (2008:204). This agreement is called a personal data processor agreement (personuppgiftsbiträdesavtal in Swedish) and specifies the rights and obligations of both parties [9].

Even though the patient data registered in Infektionsverktyget is not considered part of the healthcare record it is still comprised by the same healthcare privacy and confidentiality.

2.4 Data mining

Data mining is the process of analyzing data already existing in a database to find hidden patterns in it. The extracted patterns must be meaningful in some sense, for instance economically or from a patient safety point of view [10, p. 5].

Apart from the term data mining there are other more or less equivalent terms used in literature, such as machine learning, predictive analytics and knowledge discovery in databases (KDD) [11, p. 1]. KDD follows a nine step process where data mining is considered one of the steps even though the terms are often used interchangeably in practice [12].

The digitalized society of today generates enormous amounts of data that are simply stored on hard disks of growing size. Extracting meaningful information from the data has proved to be challenging, leading to an increasing gap between the generation of data and the understanding of data. In the great amounts of data there is potentially valuable information hidden, that can possibly be found with data mining techniques [10, p. 4]. At the same time as hard disk drive capacities increase and more and more data is stored the processing power of regular computers increases according to Moore’s law, leading to increasing opportunities for data mining [11, p. 2].

Traditional analyses of data involve querying the data to answer specific questions or verify/falsify specific hypotheses. As the number of columns and possible column values grows, the number of required queries to cover all possible combinations increases rapidly. On the contrary, when analyzing data with data mining, there is no predefined hypothesis on which the analysis is based. Instead columns of data are marked as inputs and/or outputs and the data mining system will automatically show how the input attributes are related to the outputs [11, pp. 2-3]. Exactly how these calculations are performed and how the patterns are presented depend on the chosen data mining algorithm. An advantage of not having any pre-defined hypotheses is that new and unforeseen relationships can be found. Those relationships could otherwise have been missed because they are too odd to base a hypothesis on.
2.4. DATA MINING

2.4.1 Data mining cycle

The data mining project cycle contains roughly the following steps, even though it varies depending on the purpose of the data mining [11, pp. 10-13].

1. Business problem formation
   What are the problems to be solved? How can these problems be solved with data mining? Which data mining task is suitable?

2. Data preprocessing
   The data preprocessing step contains for instance collecting all the relevant data into one place, data cleaning to remove noise and redundant data, discretization of continuous attributes, grouping and aggregation of discrete attributes. The purpose is to transform the raw data so that is is useful for mining. This is the most resource-consuming step.

3. Model building
   One or several data mining algorithms are chosen depending on the data mining task and models are built. Usually several models are built with different algorithms or with different algorithm parameter settings to be able to compare their performance in describing data or predicting.

4. Model assessment
   The predictive performance of the models is evaluated and the revealed patterns are inspected and assessed in terms of usefulness and value to the area of study.

5. Prediction
   In many cases prediction is the goal of the data mining. Models that have been trained with data in the model building step can now be used for prediction on new cases of data.

6. Application integration
   Embedding the data mining into the business application.

2.4.2 Structure and model

Two important concepts in data mining are data mining structure and data mining model. A data mining structure defines the shape of the data mining problem. It holds information about the included data columns, such as gender and age, including their data types and whether they are discrete (i.e. have a set number of values, such as gender) or continuous (i.e. are numerical, such as age). The mining structure contains the source data that
is used for training and testing of the mining models and information about how much of the data that should be used for training and testing [11, pp. 93-94].

The data mining model transforms the rows of source data into cases and performs data mining with these. It uses some or all of the source data depending on filters applied to the model. The mining model uses one data mining algorithm and some or all mining structure columns as data mining attributes and specifies if these attributes are to be used as input, output or both. The model then uses the inputs to learn about the outputs. The idea behind data mining is to show a data mining model examples of data, containing both inputs and outputs, from which it can extract patterns. This is called the training phase. The patterns can then either be studied by themselves or be applied to new examples of data. To test the predictive performance the trained model is only given the input attributes of the new cases and from these it tries to predict the output attribute state. The prediction is compared to the known state of the output attribute of that case. In this way the performance of the model in predicting the outputs can be evaluated. This is called the testing phase [11, pp. 94-95].

2.4.3 Data mining tasks and algorithms

There are various data mining tasks that can be applied to solve a problem. Depending on the nature of the problem, one or more different tasks can be combined to solve it because all work in a different way. General data mining tasks include classification, clustering, association, regression, forecasting, sequence analysis and deviation analysis. The choice of data mining algorithm then decides exactly how the data is analysed for patterns and in what form the identified patterns are (e.g. trees and rules).

Below the tasks that will be applied in this study, i.e. classification and association analysis, are presented together with brief descriptions of the algorithms that can be used for performing each task. Four out of the nine algorithms available in Microsoft SQL Server Data Mining 2008 are used for classification, and one for association analysis.

Classification

Classification is the task of assigning a state to the output attribute of each case. The classification task consists of describing the patterns of the output attribute in terms of the input attributes. A model is trained with training data where the output attribute is known, and can then be used to classify the output attribute in test data. Classification is called a supervised task because it requires a pre-supposed target, the output attribute, to lean against [11, p. 6]. The classification algorithms available are Naive Bayes, Decision Trees, Neural Network and Logistic Regression, described below.
Naive Bayes algorithm  Naive Bayes is the simplest of the classification algorithms. It builds patterns by counting the correlations between all different states of the input attributes and all different states of the output attributes. Attributes can only have discrete values. Naive Bayes is based on Bayes’ theorems and is naive in the sense that it does not take possible dependencies among the input attributes into account. Strong dependencies among the input attribute can therefore bias the identified patterns. Naive Bayes is often used in the beginning of the data mining process to quickly explore the data but can also be a powerful predictor in some situations. More advanced algorithms such as Decision Trees and Neural Network are typically used for prediction when available [11, pp. 216-217].

Patterns generated by Naive Bayes include so called attribute characteristics which can be interpreted as the main influencers. Attribute characteristics are expressed as an attribute-state combination with an associated frequency which indicates the proportion of cases with the target output state that also had this specific input attribute-state combination. The frequency can be interpreted as the strength of the influence on the infection outcome [11, p. 224].

Decision Trees algorithm  Decision Trees can handle both discrete and continuous attributes but bins the continuous values if appropriate. The algorithm works recursively to build a tree that afterwards can be used for prediction. It searches for the input attribute that most cleanly divides the data across the states of the output attribute. That input attribute is used to split the data into subsets and then the same procedure is repeated for each of the subsets and so forth. When a new case of data is to be classified it is compared to the splits of the built tree thus creating a path from the root to a leaf node. That leaf node contains the predicted state of the output attribute. During training the tree is pruned using two algorithm parameters so that the resulting tree is not too deep which may cause over-training. Very deep trees tend to overrepresent the training data instead of generalizing rules from it, which may result in a bad performance when classifying new cases of data [11, p. 252]. Decision Trees is one of the most popular algorithms because it is fast, easily understood and accurate if used properly [11, p. 236].

Neural Network algorithm  The Neural Network algorithm is an artificial neural network that mimics the way the human mind works when presented with a problem. It analyzes all possible combinations of inputs and outputs and assigns weights to their relationships. It also looks for combinations of inputs that correlate to an output even though the inputs alone do not. There is also a hidden layer of nodes between the inputs and outputs, so that the inputs do not have to be directly correlated to an output.
Instead the inputs can be related to a node in the hidden layer which in turn is related to an output, see Figure 2.1a. The resulting network can be used for prediction of new cases of data. Neural Network is suitable when trying to detect very complex relationships between inputs and outputs. However, the patterns extracted from the model are not well-suited for exploration because they are hard to interpret [11, pp. 371-373].

**Logistic Regression algorithm**  Logistic Regression is a special case of the Neural Network algorithm in the way that it contains no hidden layer, see Figure 2.1b, but besides that they are identical and therefore behave similarly. The removed hidden layer does not necessarily make it a weaker algorithm when it comes to predicting new cases of data. In some situations it can even perform better than Neural Network because the reduced complexity implies less risk of overtraining [11, pp. 372-373]. Both Neural Network and Logistic Regression are able to handle both discrete and continuous attributes.

![Figure 2.1: Schematic representation of Neural Network and Logistic Regression. Adapted from [11, p. 386]. Figure (a) has a hidden layer while (b) do not.](image)

**Association analysis**

Association involves analyzing sets of items to find frequently occurring item-sets and from these, formulate association rules. Association data mining is often called “market basket analysis” because increasing cross-selling by analyzing customers’ shopping behaviours in the form of transaction tables is a common application [11, p. 7]. Association analysis is mainly a descriptive task, i.e. its purpose is to describe patterns in the data, but it is also possible to use it for prediction of the output attribute.

**Association Rules algorithm**  Association Rules simply counts the input and output attribute states and how often combinations of those appear together. This number is called the support of an item or itemset. The strongest correlations found in the counting are generating association rules, each with a measure of support, probability and importance.
2.4. DATA MINING

The probability, also known as confidence, of a rule $A \Rightarrow B$ indicates the fraction of cases containing $A$ that also contains $B$, i.e. the conditional probability of $B$ given $A$:

$$\text{Probability}(A \Rightarrow B) = P(B \mid A) \quad (2.1)$$

The importance indicates the correlational strength of a rule and is defined as:

$$\text{Importance}(A \Rightarrow B) = \log \frac{P(B \mid A)}{P(B \mid \text{not}A)} = \log \frac{\text{Probability}(A \Rightarrow B)}{P(B \mid \text{not}A)} \quad (2.2)$$

Zero importance means that there is no correlation between $A$ and $B$, because the probability of $B$ is the same given either $A$ or not $A$, see Equation 2.2. A positive importance indicates that when $A$ is true the probability of $B$ increases, i.e. there is a positive correlation. Similarly, a negative importance indicates that when $A$ is true the probability of $B$ decreases, i.e. there is a negative correlation.

The difference between probability and importance is that while probability is concerned with the accuracy of the rule independently of the rest of the dataset, importance also takes the influence of the overall dataset into account when calculating the rule accuracy.

2.4.4 Cross validation and holdout method

Multifold or k-fold cross validation is a technique that tests the predictable performance of a model in a systematic manner. The first step is randomized splitting of the data into $k$ partitions containing the same, or as close to the same as possible, number of cases. The next step is looping through the partitions and for every partition, training the model with the data in the other partitions ($k - 1$ number of partitions) and testing it with the data in the current partition [11, p. 174-175]. Each training and testing is performed in the same way as described in Section 2.4.2. Evaluation of the predictable performance of the model must be targeted against one state of the output attribute.

The most basic evaluation method is called the holdout method, in which one part of the data is used for training (normally around one third) and the other part is set aside for testing (the holdout data, normally about two thirds) [10, p. 149]. This method can be seen as one iteration of cross validation. The advantages of this method are that it is easy and fast. However, cross validation has two main advantages over the holdout method.

1. All data can be used for training and no data will need to be set aside for testing only. This is a great advantage especially if there is limited data in the dataset.
2. All partitions will be used for testing once and irregularities in the
distribution of data among the partitions will in average even out
more than when using the holdout method.

Cross validation is frequently used in data mining studies and the most
common number of partitions, the k value, is 10 [13]. However, other values
of k such as 5 and 20 are also being used and are likely to produce almost
as good results [10, p. 150].

2.4.5 Classification matrix and evaluation measures
The classification matrix, also known as the confusion matrix, is a tool for
quantitative evaluation of the predictive performance of a model. For a given
state of the output attribute, the classification matrix shows how many cases
the model correctly and incorrectly classified during testing. A classification
matrix is shown as Table 2.1.

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>Actual Positive</td>
<td>True Positive</td>
</tr>
<tr>
<td>Actual Negative</td>
<td>False Positive</td>
</tr>
</tbody>
</table>

Table 2.1: Classification matrix.

Pneumonia is a state of the output attribute infection. This example will
be used in the list below to illustrate the meaning of True Positive, False
Positive, False Negative and True Negative.

1. True Positive (TP) is the number of correct predictions that an in-
stance is positive, i.e. the number of times the model correctly classi-
fied an actual case of pneumonia as pneumonia.

2. False Positive (FP) is the number of incorrect predictions that an
instance is positive, i.e. the number of times the model classified a
case that was not pneumonia as pneumonia.

3. False Negative (FN) is the number of incorrect predictions that an
instance is negative, i.e. the number of times the model classified an
actual case of pneumonia as something else.

4. True Negative (TN) is the number of correct predictions that an in-
stance is negative, i.e. the number of times the model correctly classi-
fied a case that was not pneumonia as something else.

Based on these numbers different evaluation measures can be calculated.
The choice of appropriate measures to achieve a reliable evaluation depends
### 2.4. DATA MINING

<table>
<thead>
<tr>
<th>Measure</th>
<th>Formula</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>$\frac{TP+TN}{TP+FP+TN+FN}$</td>
<td>Measures the proportion of true predictions (TP+TN) in the whole population (TP+FP+TN+FN)</td>
</tr>
<tr>
<td>Recall, R</td>
<td>$\frac{TP}{TP+FN}$</td>
<td>Measures the proportion of actual positives (TP+FN) that are correctly classified as positives (TP)</td>
</tr>
<tr>
<td>Precision, P</td>
<td>$\frac{TP}{TP+FP}$</td>
<td>Measures the proportion of positive classifications (TP+FP) that are correctly classified (TP)</td>
</tr>
<tr>
<td>F-measure</td>
<td>$2 \cdot \frac{R \cdot P}{R+P} = \frac{2TP}{2TP+FP+FN}$</td>
<td>Represents a harmonic mean between recall and precision</td>
</tr>
</tbody>
</table>

Table 2.2: Formulas and purposes of the evaluation measures.

on both the nature of the data and the purpose of the data mining. Table 2.2 lists the evaluation measures used or discussed in this study.

Recall is also known as sensitivity or true positive rate (TPR) while precision is sometimes referred to as positive predictive value (PPV).

#### 2.4.6 Medical data mining

Because of its successful application in areas such as retail, e-commerce, banking and insurances, data mining has spread to other sectors such as healthcare [14]. Healthcare is an area with great potential for successful data mining because of the wealth of data available. However there is generally still a lack of effective analysis tools to maximize the utility of the data - the healthcare environment is often said to be information rich but knowledge poor [15]. Bellazzi and Zupan [14] states that data mining, and especially predictive data mining, has been used relatively limited in medicine but has started to become an important instrument for researchers but also for clinical practitioners.

Data mining with healthcare data differs from data mining within other areas in a number of ways. These differences might cause problems and require more data preprocessing and preparation to be performed, than when mining data from other areas [16].

The first difference concerns the data quality. The main purpose of healthcare data collection is to keep a record of a patient’s medical history for the benefit of the individual patient [16]. As a result the quality of the data is often lower with large heterogeneous datasets potentially containing missing values. As an example, not all patients, not even patients with the same diagnosis, undergo the same examinations and treatments which leads to a wide variety of data. Health records that are not migrated to electronic versions or that are only scanned can contain valuable data for the data mining task but cannot be used [13]. Another concern is that medical data is...
often of high dimension and thinly spread across these dimensions, meaning that extensive data is required to cover all possible combinations so that something meaningful can be obtained from data mining [17].

The second important factor in data mining with healthcare data is the patient privacy and legal issues. The privacy regulations regarding data for data mining are the same as for any medical data. A way of making data that is collected together with patient identification information, for instance social security number, less trackable to individual patients is to de-identify it. De-identification can consist of encoding of the patient identification information to a pseudo ID that is used to keep track of the patient data during the data mining process [16]. An even more secure option is to use anonymized data, which means that the patient identification information is removed forever and can never be restored [16].

2.5 Previous studies

Ramon et al. [18] performed a study to investigate the possibility of predicting outcomes of intensive care patients. Data that is known to be relevant for the outcome in intensive care units, such as inflammation and organ failure, was used to predict outcomes related to specific target attributes, such as “Severe-inflammation N days from today?”. Four different data mining algorithms were used and the results were evaluated with 10-fold cross validation and two evaluation criteria. They found that the decision tree algorithm produced visual results that could easily be interpreted by domain experts, but the nonvisual algorithms produced better prediction accuracy.

Wright, Chen and Maloney [19] investigated whether association techniques in data mining could be useful for automatic finding of relationships between patients’ medications, laboratory results and problem lists. This could be used to find and fill gaps in the problem lists. Found association pairs were evaluated quantitatively and strong candidates were also evaluated qualitatively by comparing them to gold standard references. Results showed that a large number of clinically accurate associations could be found and that the technique seems to be a useful approach.

Mullins et al. [17] studied the potential value of searching a large database of electronic patient records to find novel correlations. They mined data on conditions, laboratory test results and procedures using three unsupervised rule discovery techniques to find associations. The conclusion was that data mining without any particular hypotheses to falsify or verify can be a good alternative to traditional hypothesis-based research to find novel correlations for further investigation.

There is a couple of studies on data mining techniques for surveillance of healthcare-associated infections, of which Brossette et al. [20] was the
first that did not assume a predefined outcome to monitor. The purpose of their study was to mine surveillance data for new, unexpected and potentially interesting patterns to be used for detection, prevention and control of disease outbreaks. For this they proposed a new method using an association rules algorithm. To find non-trivial correlations the authors looked for high-support, low-probability association rules rather than high-support, high-probability rules. To analyze changes in time they compared the probability of a rule at two points in time. If the probability had increased significantly, this association rule was selected for further inspection. The method was demonstrated with infection control data and proved to be effective in identifying previously unknown and potentially interesting patterns. Ma et al. [21] also used the method proposed by Brossette et al. [20] but instead looked for low-support, low-probability rules and found that this approach is also likely to be valuable for infection control surveillance.
Chapter 3

Method

3.1 Data mining environment

The chosen data mining platform was Microsoft SQL Server 2008 Data Mining which is a part of Microsoft SQL Server Analysis Services. Analysis Services is an analysis tool for online analytical processing (OLAP), data mining and reporting in Microsoft SQL Server which is a relational database management system.

SQL Server Data Mining is widely used and Microsoft SQL Server is considered highly secure [22], but the main reasons for the choice of platform is simplicity and ease of use. Because the databases of Infektionsverktyget are managed in SQL Server it is easy to access the database tables and views from Analysis Services. There is no data importing issue and all database tables and views as well as data mining structures and models can be accessed through the same interface; Microsoft SQL Server Management Studio.

There is a wide variety of other data mining software available such as WEKA, Orange, IBM SPSS Modeler and STATISTICA Data Miner. A disadvantage with Microsoft SQL Server Data Mining is that it is not a freeware as some of the other data mining softwares. However this was not considered a problem because the SQL Server Data Mining rights are included in the SQL Server license. Overall, no other softwares than SQL Server Data Mining were considered due to the easy integration with the rest of SQL Server.

In SQL Server database queries were written in T-SQL (Transact-SQL) which is an extension to the standardized database query language SQL (Structured Query Language). T-SQL is very similar to SQL in many ways but offers some extra functionality. The language used for creation and handling of data mining structures and models in Analysis Services was DMX (Data Mining Extensions to SQL). DMX is a language with SQL-like syntax created specifically for performing data mining operations.
Data mining in SQL Server can be performed in three different interfaces. Data mining add-ins to Microsoft Excel is the simplest option but does not offer all the functionality needed for this study. For instance it does not support data mining on nested tables, only on relational tables. Why enabling nested tables was important is described later in this chapter. SQL Server Business Intelligence Development Studio (BIDS) is specifically designed for business intelligence questions within SQL Server and offers a wide variety of functionalities but is lacking the coding interface preferred for this study. The choice of interface was therefore SQL Server Management Studio (SSMS) where it is possible to manage all database tables and views as well as the data mining structures and models at the same time. It is possible to directly write T-SQL and DMX queries to create views, create and populate mining structures and models and to make predictions.

3.2 Data mining cycle

The methods used for data mining are presented according to the data mining project cycle steps introduced in Section 2.4. The last two steps, prediction and application integration, are not presented because they were not in the scope of this thesis.

3.2.1 Problem formation

The goal of the data mining was to use data registered in Infektionsverktyget to search for patterns that describe how patients’ gender, age, previous diagnoses and underwent medical procedures relate to different outcomes of healthcare-associated infections treated with antibiotics. It was desirable that these patterns would be either suitable to study in themselves or that they could be used for prediction of infection outcomes based on new cases of patient data. The prediction should have a high performance and what that means is discussed later in this chapter.

To achieve these goals classification was identified as a suitable data mining task because it involves describing the output attribute, in this case the infection outcome, in terms of the input attributes. Since classification is mainly a predictive task it is highly suitable for prediction based on the patterns between the input and output attributes.

Association analysis is another data mining task that was identified as potentially useful for this problem because it involves the identification of common itemsets and thereof applicable association rules. These rules can be studied to find interesting relationships among the input and output attributes. Prediction is possible when using the Association Rules algorithm available in SQL Server Data Mining, even though the task of association is mainly descriptive.
CHAPTER 3. METHOD

3.2.2 Data preprocessing

The data was preprocessed in order to be easily accessible and meaningful for the data mining later on. The following data was decided to be used as input attributes for the data mining, based on the problem formation presented above and the type of data available in the database, see Section 2.2:

- Gender
- Age
- Diagnoses
- Actions

The output attribute was the antibiotic treated healthcare-associated infection outcome, because this is the attribute that we want to study in terms of influence from the input attributes. The possible states of the output attribute was the six types of healthcare-associated infections registered in Infektionsverktyget, see Section 2.2, as well as a state indicating the absence of a HAI: “No infection”. The state “another type of HAI” that can be registered in Infektionsverktyget as a prescription cause, was excluded because it was unlikely to provide interesting and valuable results.

The data preprocessing phase needed to handle a number of issues specific to this data mining problem and the nature of the data in the Infektionsverktyget database. The following issues were identified:

1. The patient ID is the only identifier that ties the information such as diagnoses, actions and prescriptions together in the database, according to Section 2.2. Because the chosen data mining tasks are concerned with finding correlations between input and output attributes, there was a need to find a way to tie the diagnoses and actions to the infection outcomes that they may have caused. As an example of why this is important, it is certain that an appendix surgery 10 years ago would not be the cause of an urinary tract infection today, and therefore the relationship between them are uninteresting for the purpose of this study.

2. As the study is focused on real medical data there is an arbitrary number of diagnoses, actions and infection outcomes associated with each patient. Consequently, a solution that could hold all the data used for mining and simultaneously handle this variation was needed. An ordinary relational table would be large, inflexible for possible changes and contain missing values in the majority of the table cells.
3. There is a large number of unique diagnosis and action terms registered in Infektionsverktyget. ICD-10-SE (International Statistical Classification of Diseases and Related Health Problems, tenth revision, Swedish version), the coding system for diagnoses, contains tens of thousands of diagnosis codes whereas KVÅ (Klassifikation av Vårdgärder), the coding system for medical actions, includes around ten thousand codes for different actions. Data with such a large number of states, especially when it occurs in several attributes (both diagnoses and actions), is not suitable for data mining because of the phenomenon known as the curse of dimensionality. The complexity and computational cost of the problem quickly increases and thus an appropriate solution to this issue was required.

The following three sections describe how each of these issues were handled.

**Connection of data**

According to the first post of the list presented above, the input and output attributes had to be tied together to enable data mining. This was solved by using the information about medical care events registered in the database and demarcations in time. Each registered care event made up one row of data and diagnoses, actions and HAI outcomes were connected to the care event using different time frames. Diagnoses and actions were considered if they were registered within 30 days before hospital admission (the start of the care event) to the date of discharge (the end of the care event). This time frame was chosen because it was believed to include the factors that mainly affected the infection outcome.

Regarding the infection outcomes, the time frames varied depending on the characteristics of the infections. For superficial postoperative wound infections the infection arise within 30 days of the procedure and for deep postoperative infections within 1 year, according to Socialstyrelsen [2] and the American Centers for Disease Control and Prevention (CDC) [24]. Postoperative wound infections are normally related to a specific medical action. However, in this study no such connection was made because that would have introduced predefined patterns when the purpose was in fact to look for patterns without any specific hypotheses. Instead the postoperative wound infections were connected to a care event if they arose within a time frame from hospital admission to 30 days (for superficial infections) or 1 year (for deep infections) after discharge.

For urinary tract infection, pneumonia, sepsis and Clostridium difficile infection the infection is usually classified as healthcare-associated if it debuts after more than 48 hours from hospital admission, or less if it can be derived to a specific medical action [2, 25]. There is no general rule for
how long after discharge the infection can debut and still be classified as healthcare-associated. In this study the time frame used to tie these types of HAIs to a care event was from 48 hours after admission to 30 days after discharge.

**Nested tables**

The second issue is related to the organization of the data used for data mining. As mentioned above, a single relational table has many drawbacks when used for data with a lot of variation. Instead a nested data setup was used. A nested table is a regular relational table contained within a cell of another table called the case table. The nested table has its own primary key and when included in a mining structure, it is typical for a nested table to only have one single column that is the primary key. The identified goal was to organize the data to be used for mining as in Figure 3.1, where the Diagnoses and Actions attributes contain nested tables. Using a nested table data setup complex cases with an arbitrary number of diagnoses and actions connected to a patient were possible to use in data mining.

<table>
<thead>
<tr>
<th>Id</th>
<th>Gender</th>
<th>Age</th>
<th>Diagnoses</th>
<th>Actions</th>
<th>Infection</th>
</tr>
</thead>
<tbody>
<tr>
<td>925A9663-C2E2-45DC-A7A2-56CB3634FAAA</td>
<td>Female</td>
<td>78</td>
<td>Diagnosis</td>
<td>Action</td>
<td>Sepsis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Acute myocardial infarction</td>
<td>Operations on the pericardium</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Paroxysmal tachycardia</td>
<td>Operations on the coronary arteries</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Psoriasis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9FFC9C2B-99CB-4047-92F7-ADD0F8FC9E5BA</td>
<td>Male</td>
<td>63</td>
<td>Diagnosis</td>
<td>Action</td>
<td>No infection</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Chronic obstructive pulmonary disease</td>
<td>Operations on lungs</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Operations on the chest wall, pleurae and diaphragm</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Operations on bronchi</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>General anaesthesia, intravenous induction with relaxing agents</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3.1:** The nested table data setup with example data. Id is the primary key of the case table and Diagnosis and Action are primary keys of the respective nested table.

**Grouping**

The third and last issue on the list presented above is concerned with the numerous amounts of diagnosis and action codes which can pose a problem
for data mining. The chosen solution to this was to group the ICD-10-SE and KVÅ codes describing similar diagnoses and actions to reduce the complexity of the data mining problem [11, p. 11].

The ICD-10-SE codes consist of one letter and three trailing digits. Seen as a tree, there are four classification levels except the root node containing every single code. The first level contains the chapters which is the first division of diagnoses into groups such as neoplasms, diseases of the nervous system, diseases of the circulatory system etc. The second and third level contain more and more specific sub- and subsub-groups of the chapters and finally the fourth level contains the leaves of the hierarchy tree, i.e. the actual codes.

There are also four classification levels of KVÅ codes except the root node containing every single KVÅ code. The KVÅ codes are on the first level divided into KKÅ (Klassifikation av Kirurgiska Åtgärder) codes for surgical actions and KMÅ (Klassifikation av Medicinska Åtgärder) codes for nonsurgical actions [23]. The two types are distinguishable in that the KKÅ codes consist of three letters from A to Z and two trailing digits whereas the KMÅ codes consist of two letters from A to Z with three trailing digits. The second level represents the chapters which are distinguished by the first letter in each code. For KMÅ codes the chapters are divided into investigating actions, treating actions, anaesthetic actions etc. while for KKÅ codes the chapters represent different locations of the body where the surgical action is performed. The third level contains sub-chapters represented by the first two letters in the codes and the third level contains the actual codes, the leaves of the hierarchy tree.

The code hierarchies of ICD-10-SE and KVÅ illustrated with a selection of codes can be studied in Appendix A and B, respectively. The appropriate levels of grouping for the data mining were chosen in the model building step of the data mining project cycle.

Views

A number of different virtual views were created to be able to populate the case table and nested tables in Figure 3.1 with data. A virtual view is a query over database tables. Its result is not stored as a table in the database, which is the reason it is called “virtual”, but the view can be queried as if it existed. Before the created views are described an explanation of the organization of the database is given below.

Only selected tables in the Infektionsverktyget database were needed for this study. A schematic diagram of the tables that were used and the relationships between them can be studied in Figure 3.2. Every table has a primary key (a unique identifier of a row in the table) and one or more foreign keys (a unique identifier of a row in another table). The foreign key constraints are used to ensure that the value of the foreign key in the
CHAPTER 3. METHOD

The referencing table must also appear in the referenced table, and in this way the referential integrity is maintained on insertions or updates to the database.

Figure 3.2: Database diagram over tables and columns used. Primary keys are indicated with a key symbol and foreign key constraints are indicated with arrows pointing from the referencing table to the referenced table.

The Patient table has the columns PatientId, YearOfBirth and GenderId. GenderId is a foreign key referencing ConceptId of the Concept table. The Concept table holds all concept IDs in the database. The term that corresponds to each of the concept IDs, such as “Female” or “Male”, can be found in the lookup table Term. The tables Diagnoses, Actions and PrescriptionCause hold information about the patient ID as well as the concept ID and date of the diagnosis, action and cause of prescription (the type of infection that lead to the prescription of antibiotics). The PrescriptionCause table also has the column TransmissionId which indicates the route of transmission. The concept IDs DiagnosisConceptId, ActionConceptId and CauseConceptId are referencing the Concept table and the corresponding term can be looked up in the Term table in the same way as for gender. The corresponding ICD-10-SE, KVÅ or Snomed CT code can be found up in the Concept table. Snomed CT codes are used in Infektionsverktyget for coding of infections. The CareEvent table holds information about every
medical care event, including the patient ID of the treated patient and date of hospital admission and discharge.

The views vPatients, vInfections, vInfectionsAll and vPatientsInfections were created to populate the case table with data (columns Id, Gender, Age and Infection in Figure 3.1) and the views vDiagnosesHierarchy, vDiagnoses, ActionsHierarchy and vActions were created to populate the nested tables (tables Diagnosis and Action in Figure 3.1). Each of the views are described below.

**vPatients** The view vPatients selects each medical care event and includes the associated patient’s gender and age. The columns of this view are:

- *CareEventId* is the primary key of the view and is retrieved from the CareEvent table where it is simply called Id.
- *Gender* is generated by looking up the gender term (“Female” or “Male”) in the Term table that corresponds to the GenderId column in the Patient table.
- *Age* in years is calculated by subtracting the YearOfBirth in the Patient table from the year of the current database system timestamp.

**vInfections** The view vInfections contains all the infections that are associated to a medical care event, i.e. are within the given time frames. Each of the six HAI outcomes are selected individually from the Prescription-Cause table using their unique Snomed CT code and according to their specific time frame stated above. The infections that are not classified as healthcare-associated but rather as community acquired are filtered out using the TransmissionId column in the PrescriptionCause table. Thereafter all selected rows are put together into one single view with the UNION ALL SQL operation. UNION ALL was chosen over simply UNION because there are no duplicate rows and it executes faster. The resulting view has the following columns:

- *Id* is a randomized string that is the primary key of the view. It was needed because the CareEventId could not be assumed to be a primary key because theoretically one care event can have more than one HAI associated with it.
- *CareEventId* is the Id of the CareEvent table.
- *Infection* contains the term of one of the six HAIs that is associated with the care event.
CHAPTER 3. METHOD

vInfectionsAll  The view vInfectionsAll contains the outcomes of all medical care events. It is built up by first selecting all the care events that are present in the CareEvent table but not in the vInfections view. That is, all the care events that did not result in an infection. An ID is randomized and the CareEventId is selected in the same way as in vInfections though the value of the Infection column is constant; “No infection”. After that all the rows of the vInfections view are added to this view with the UNION ALL operator, so that the view contains all outcomes associated to all care events. The columns of this view are naturally the same as for the vInfections view.

vPatientsInfections  The view vPatientsInfections joins the views vPatients and vInfectionsAll together on the CareEventId, so that for every care event in vPatients the HAI outcome is added. The IDs from the vInfectionsAll view are kept to make sure that there is a unique identifier for each row. Hence, the vPatientsInfections view contains the following attributes: Id, CareEventId, Gender, Age, Infection.

vDiagnosesHierarchy  For the diagnosis grouping described earlier, there was a need for a view which contains all the diagnosis codes and terms of the ICD-10-SE code system on the four different levels. This would make it easy to select grouping terms on the desired depth of classification when performing data mining. The view vDiagnosesHierarchy was created by several subsequent self-joins of both the Concept and Term tables. To filter out all codes except the ICD-10-SE codes the column SubsetId of the help table SubsetConcept was used. SubsetId has a specific value for all concept IDs that belongs to an ICD-10-SE diagnosis. The eight columns of this view are Level1Code, Level1Term, Level2Code, Level2Term etc. up to Level4Term. The codes and terms on the different levels correspond to the hierarchy tree in Appendix A.

vDiagnoses  The view vDiagnoses selects all the diagnoses associated with each of the care events, i.e. within one month before admission to discharge. It has the following columns:

- **Id** is a randomized primary key of the view. CareEventId could not be the primary key because a patient can have more than one diagnosis associated with each care event.

- **CareEventId** from the CareEvent table.

- **Level1Term, Level2Term, Level3Term** and **Level4Term** are retrieved from the vDiagnosesHierarchy view and contain the diagnosis associated to the specific care event expressed on all four levels.
3.2. DATA MINING CYCLE

**vActionsHierarchy**  vActionsHierarchy works in exactly the same way as vDiagnosesHierarchy, but for the KVÅ codes and terms instead of ICD-10-SE. It was created in a similar way as vDiagnosesHierarchy but was filtered with a different SubsetId. It has the same eight columns with action codes and terms on all four levels which correspond to the hierarchy tree in Appendix B.

**vActions**  The view vActions selects all the actions within the specified time frame of each care event, i.e. from one month before the admission to the discharge. When selecting the actions, a filter is applied so that only those actions that are considered risk actions for HAIs are selected. This is achieved by using the SubsetConcept help table available in the database. A specific SubsetId is used to filter out all but those actions, which are decided upon by a group of specialists. The majority of the risk actions are surgical actions although some of the non-surgical actions are also included. The reason for the filtering was that the non-risk actions were considered irrelevant for the infection outcome and therefore could be ignored to reduce the complexity of the problem. The vActions view contains the following columns:

- *Id* is a randomized primary key of the view. CareEventId could not be the primary key because a patient can have more than one action associated with each care event.
- *CareEventId* from the CareEvent table.
- *Level1Term, Level2Term, Level3Term* and *Level4Term* are retrieved from the vActionsHierarchy view and contain the action performed on the patient under care in the specific care event, expressed on all four levels.

### 3.2.3 Model building

**Structures**

Two data mining structures were created, AllClinics and WomensHealth-Clinic. The first one was created to contain data from all clinics in Uppsala County Council and the second one to be used with data from only one of the clinics in Uppsala County Council, the women’s health clinic. The idea behind this was to study if there were any differences in prediction performance between models based on the two different datasets, i.e. if it is easier to predict HAIs in all clinics simultaneously or in one specific clinic at a time. Both of the structures had the same columns:

- *Id* (key)
CHAPTER 3. METHOD

- CareEventId (continuous)
- Gender (discrete)
- Age (continuous), for the algorithms that can handle continuous attributes (Decision Trees, Neural Network, Logistic Regression)
- AgeDisc (discretized), the continuous age discretized into five bins of equally many values, for the algorithms that can only handle discrete or discretized attributes (Naïve Bayes, Association Rules)
- Diagnoses (discrete), containing the nested table Diagnosis
- Actions (discrete), containing the nested table Action
- Infection (discrete)

No holdout of data for testing was specified because cross validation was used for the quantitative evaluation and therefore all data needed to be available for both training and testing.

After creation the structures were populated with data from the three resulting views described in Section 3.2.2; vPatientsInfections, vDiagnoses and vActions. Before insertion the views were filtered so that only data from the chosen clinics was used. This was achieved by first joining the views with the CareEvent table of the database and using the CareProviderId to filter out all care events belonging to clinics outside Uppsala County Council. For the structure WomensHealthClinic there was also a join with the table OrganizationalUnit to filter out care events from all clinics except the women’s health clinic, using the CareUnitId column.

For insertion into all columns except Diagnoses and Actions, the corresponding columns of the vPatientsInfections view were used. The data in the Age column of the view was inserted into both the columns Age and AgeDisc of the structure and the latter was discretized automatically. For insertion into the Diagnosis table and the Actions table the views vDiagnoses and vActions were used, respectively. During insertion into these nested tables, the appropriate classification levels of the diagnosis and action terms for each structure were specified. The CareEventId column coming from the vPatientsInfections view was related to the CareEventIds of the vDiagnoses and vActions views to tie the data from the different views together into correct rows in the data mining structures.

Regarding the classification level of the diagnosis and action terms, different depths were used for the two structures. By studying the different levels of terms and the number of terms at each level, see Table 3.1, appropriate depths were decided upon. The structure AllClinics used diagnosis terms on level 2 and action terms on level 3 because those levels contain
approximately the same number of terms, see Table 3.1, and because to-
gether they would result in a reasonable amount of terms for data mining.
A balance is required to reduce the complexity of the problem by grouping
the codes while allowing certain level of detail to provide more meaning.
The second structure, WomensHealthClinic, consists of data from only one
clinic and therefore the diagnosis and action terms used there are concen-
trated to fewer areas of medicine. Because of this the level of grouping can
be lower. This can also be seen as a requirement to make the data mining
meaningful for a single clinic, because using a higher level of grouping on
a specialized field of medicine would result in only a few wide terms. For
these reasons, the WomensHealthClinic structure used diagnosis terms on
level 3 and action terms on level 4.

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis terms</td>
<td>$10^1$</td>
<td>$10^2$</td>
<td>$10^3$</td>
</tr>
<tr>
<td>Action terms</td>
<td>$10^3$</td>
<td>$10^4$</td>
<td>$10^5$</td>
</tr>
</tbody>
</table>

Table 3.1: The order of magnitude of the number of diagnosis and action terms
on the different classification levels.

Models

Thirty-two models were created, sixteen for each structure. Ten out of those
sixteen were created and tested in round one of the quantitative evaluation
and six in round two. Out of all models, four were chosen to extract patterns
from.

All models had five attributes each:

- Gender
- Age or AgeDisc depending on algorithm
- Diagnoses
- Actions
- Infection

However, the total number of attributes was larger because in data min-
ing all values contained in nested table are considered attributes with the
binary states “Existing” or “Missing”. Using the nested table in Figure 3.1
as an example, the first row of data would have the attribute Gender with
the state Female, attribute Diagnoses(Acute myocardial infarction) with the
state Existing, attribute Actions(Operations on the pericardium) with the
state Existing and so forth.

Below the choice of models is described.
Quantitative evaluation, round one  The models in round one were based on three different data mining algorithms with slightly different parameter settings. However, only the parameters that were believed to affect the predictive performance were changed. The first round was limited to three of the simplest algorithms to evaluate their performance before using more advanced algorithms. For each structure, the models listed in Table 3.2 were used.

<table>
<thead>
<tr>
<th>Model number</th>
<th>Algorithm</th>
<th>Parameter settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Naive Bayes</td>
<td>Maximum input attributes = 100</td>
</tr>
<tr>
<td>2</td>
<td>Naive Bayes</td>
<td>Maximum input attributes = 300</td>
</tr>
<tr>
<td>3</td>
<td>Association Rules</td>
<td>Minimum probability = 20%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum importance = 0</td>
</tr>
<tr>
<td>4</td>
<td>Association Rules</td>
<td>Minimum probability = 80%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum importance = 0</td>
</tr>
<tr>
<td>5</td>
<td>Decision Trees</td>
<td>Maximum input attributes = 100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum support = 10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complexity penalty = 25%</td>
</tr>
<tr>
<td>6</td>
<td>Decision Trees</td>
<td>Maximum input attributes = 100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum support = 50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complexity penalty = 50%</td>
</tr>
<tr>
<td>7</td>
<td>Decision Trees</td>
<td>Maximum input attributes = 100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum support = 100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complexity penalty = 75%</td>
</tr>
<tr>
<td>8</td>
<td>Decision Trees</td>
<td>Maximum input attributes = 300</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum support = 10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complexity penalty = 25%</td>
</tr>
<tr>
<td>9</td>
<td>Decision Trees</td>
<td>Maximum input attributes = 300</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum support = 50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complexity penalty = 50%</td>
</tr>
<tr>
<td>10</td>
<td>Decision Trees</td>
<td>Maximum input attributes = 300</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum support = 100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complexity penalty = 75%</td>
</tr>
</tbody>
</table>

Table 3.2: The models used for each structure in round one of the quantitative evaluation. All algorithm parameters not specified used default values.

Two different values of the parameter maximum input attributes were tested for model number 1 and 2. This parameter is used to limit the number of input attributes considered in model training by ignoring the least important attributes.

Model number 3 and 4 both had a minimum importance of rules of 0 because a positive importance implies positive correlation. Two different values of minimum probability of rules generated by the model were used to investigate the possible effect of this parameter on the predictive performance. Even though association algorithms are mainly used for descriptive purposes, the Microsoft version of the algorithm can also be used for pre-
3.2. DATA MINING CYCLE

diction. This enabled the predictive performance of the Association Rules algorithm to be evaluated in this study.

For models number 5 to 10, two behaviours of the Decision Trees algorithm were controlled which required as many as six models. Two different values of the parameter maximum input attributes, which has the same effect as for Naive Bayes, were tested together with three levels of pruning. Both the parameters minimum support and complexity penalty affects the pruning so these were increased simultaneously to increase the pruning. Minimum support controls the number of cases required on each side of a split while complexity penalty affects the overall penalty applied to large trees, and together they control the total pruning.

The number of models was limited because it was not possible to work with the patient data directly, as explained later in this chapter. Assistance from authorized persons was needed in every step to enable evaluation with real data and therefore only a selection of models could be tested.

Quantitative evaluation, round two In round 2 an additional six models for each structure, twelve models in total, were created. The results of the first round influenced the choice of models for round two so that four slightly modified models based on the simpler algorithms were tested. Included were also two models based on the more advanced algorithms Neural Network and Logistic Regression. The models used are listed in Table 3.3.

<table>
<thead>
<tr>
<th>Model number</th>
<th>Algorithm</th>
<th>Parameter settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Naive Bayes</td>
<td>Maximum input attributes = ∞</td>
</tr>
<tr>
<td>12</td>
<td>Association Rules</td>
<td>Minimum probability = 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum importance = 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum support = 10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Maximum itemset size = ∞</td>
</tr>
<tr>
<td>13</td>
<td>Decision Trees</td>
<td>Maximum input attributes = ∞</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum support = 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complexity penalty = 10%</td>
</tr>
<tr>
<td>14</td>
<td>Decision Trees</td>
<td>Maximum input attributes = ∞</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum support = 5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complexity penalty = 20%</td>
</tr>
<tr>
<td>15</td>
<td>Neural Network</td>
<td>Default</td>
</tr>
<tr>
<td>16</td>
<td>Logistic Regression</td>
<td>Default</td>
</tr>
</tbody>
</table>

Table 3.3: The models used for each structure in round two of the quantitative evaluation. All algorithm parameters not specified were assigned default values.

For model number 11 there was no limit in the number of input attributes, because round one showed that an increased number of input attributes increased the performance for the Naive Bayes algorithm. Model number 12 has several parameters changed because the models based on the Association Rules algorithm in round one detected no infections at all. With
these changes it was believed that more rules would be included which would increase the predictive performance. The models based on Decision Trees in round one showed that harder pruning resulted in worse performance, so models 13 and 14 in round two used two lower levels of pruning together with an unlimited number of possible input attributes. Models 15 and 16 were based on the more advanced algorithms Neural Network and Logistic Regression, respectively. Default valued parameters were used just to get an indication of their performance compared to the simpler algorithms.

3.2.4 Model evaluation

The predictive performance of the developed data mining models, see Table 3.2 and 3.3, were evaluated quantitatively using 5-fold cross validation. This means that for each model the data was divided into five partitions which were all used for testing once. \(k=5\) was used instead of \(k=10\) to reduce the cross validation execution time. Cross validation is computationally expensive and therefore time consuming because the model needs to be both trained and tested \(k\) times with different partitions of the data. Also, in this study several models needed to be tested directly after each other which increases the total execution time even more. Time was an important factor because the evaluation had to be performed in a web application to be able to use real data, as explained later in this chapter. Therefore the execution time had to be reasonable to increase the user friendliness of the web application. It is unreasonable to wait for hours for a web page to load the results after a button was clicked. A solution to this was to decrease \(k\) which decreased the execution time. The disadvantage of a decreased \(k\) is that it may lead to a less accurate estimation of the prediction performance. This was not considered a critical drawback because this study was aiming at getting an indication of what kind of models would perform better rather than an as exact performance estimate as possible. The accuracy of the performance estimate using \(k=5\) was considered to be good enough for this purpose.

Cross validation was performed for each model using each of the six infection outcomes as the predictable attribute state. The state “No infection” was excluded because the purpose was to predict the infections. When running the cross validation each model-infection combination resulted in five different classification matrices, one for each of the five partitions used for testing. The number of True Positives, False Positives, True Negatives and False Negatives in the classification matrices were averaged over the five partitions by summation and division by five. This resulted in a single averaged classification matrix for each of the model-infection combinations.

Out of all care events in Sweden it is known that only approximately 8-9% result in a HAI, see Section 2.1. Because there are six different infections outcome in this study that together make up all the cases of HAIs,
3.2. DATA MINING CYCLE

each of the infections can be considered relatively rare. The most common evaluation measure used in evaluation of a data mining model is accuracy, see Table 2.2, which measures the overall classification performance of the model. However, the accuracy is not a good measure for datasets containing rare events because it puts too much weight on prediction of the majority class, in this case the outcome “No infection” [26]. As an example imagine that the output attribute state to predict is “Pneumonia” and that “Pneumonia” is represented in only 1% of the cases in the dataset while “No infection” make up 95%. If a model would classify all the cases as “No infection”, it would achieve a classification accuracy of 95%, even though it did not classify any of the pneumonia cases correctly.

There are other evaluation measures than accuracy that produces more reliable estimates of the performance when mining for rare events [26]. In this study, the detection of the minority states (the HAIs) is more important than detection of the majority state (“No infection”). This is because the purpose was to try to identify the patients with increased risk of incurring a HAI and the patterns describing these relationships. The measures used were therefore recall and precision, which only take into account the performance of predicting the minority states, and their harmonic mean, F-measure, see Section 2.4.5. These measures are useful when the majority and minority state are unequally important [26]. There are also other more advanced measures that could have been suitable for this study, such as the receiver operating characteristics (ROC) and the area under curve (AUC) associated with it [26], but these were not considered for time reasons.

This study is dealing with so called multi-state classification, meaning that there are more than two states of the output attribute. As a result, there is a need for a method to calculate the average performance over the six different states, i.e. all infection outcomes except “No infection”. There are two approaches to how to calculate the evaluation measures recall, precision and F-measure from the averaged classification matrix of each infection outcome and model described above: micro and macro averaging. The formulas for both averaging methods are listed in Table 3.4.

In micro averaging, the number of TPs, FPs, TNs and FNs is summarized for all of the infection outcomes and then the evaluation measures are calculated using these summarized numbers with the regular formulas in Table 2.2. This will favour bigger states, i.e. infection outcomes with more cases, because the number of TPs, FPs, TNs and FNs are higher and influences the calculated measures more. In macro averaging the evaluation measures are calculated individually for each infection outcome and then averaged over the states by summation and division by six. This results in equal impact from all the states independent of size. [27] For this study macro averaging was used because measures unbiased by state size were desirable.
CHAPTER 3. METHOD

<table>
<thead>
<tr>
<th>Measure</th>
<th>Micro-averaging formula</th>
<th>Macro-averaging formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>$\frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} TP_i + FN_i}$</td>
<td>$\frac{\frac{1}{l} \sum_{i=1}^{l} TP_i}{\frac{1}{l} \sum_{i=1}^{l} TP_i + FN_i}$</td>
</tr>
<tr>
<td>Precision</td>
<td>$\frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} TP_i + FP_i}$</td>
<td>$\frac{\frac{1}{l} \sum_{i=1}^{l} TP_i}{\frac{1}{l} \sum_{i=1}^{l} TP_i + FP_i}$</td>
</tr>
<tr>
<td>F-measure</td>
<td>$\frac{\sum_{i=1}^{l} 2TP_i}{\sum_{i=1}^{l} 2TP_i + FP_i + FN_i}$</td>
<td>$\frac{\frac{1}{l} \sum_{i=1}^{l} 2TP_i}{\frac{1}{l} \sum_{i=1}^{l} 2TP_i + FP_i + FN_i}$</td>
</tr>
</tbody>
</table>

Table 3.4: Formulas for micro and macro averaging of measures in multi-state classification [27]. $l$ indicate the number of states of the predictable attribute. Here $l = 6$ because there are six different infection outcomes included.

3.3 Pattern extraction

Based on the results of the cross validation round one and two, four of the models in Table 3.2 and 3.3 were chosen for the pattern extraction, two for each data mining structure. These are listed in Table 3.5.

<table>
<thead>
<tr>
<th>Data mining structure</th>
<th>Model number</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>AllClinics</td>
<td>11</td>
<td>Naive Bayes</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>Association Rules</td>
</tr>
<tr>
<td>WomensHealthClinic</td>
<td>1</td>
<td>Naive Bayes</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>Association Rules</td>
</tr>
</tbody>
</table>

Table 3.5: The models used for pattern extraction.

Two models based on the Naive Bayes algorithm were chosen because that algorithm showed the best predictive performance in the quantitative evaluation, which will be discussed later in this report. All models based on the Naive Bayes algorithm performed approximately equally well which made the exact choice of models less important. Also, two models based on the Association Rules were chosen because of their potential of finding interesting patterns that are more easily interpreted compared to other models. Because Association Rules is mainly a descriptive algorithm, it could generate interesting patterns even though it proved to be unsuitable for prediction.

3.4 Web application development

Nowhere in the personal data processor agreement described in Section 2.3 is the rights and obligations in a situation like this thesis described. According to a lawyer specialized in this field the agreement was not enough to allow
the author of this thesis to work directly with the patient data to evaluate the models.

Instead a web application which enabled indirect use of the patient data was developed for this purpose. Data mining structures and models were created on an Analysis Services server and populated with the real patient data from a database server using the views described in Section 3.2.2. Creation and population were performed by authorized Mawell personnel upon instruction by the author of this thesis. An Uppsala County Council staff member, i.e. a representative of the healthcare provider and the owner of the patient data, then ran the website and sent the results back. That way unauthorized persons were never in contact with the patient data but the models could be evaluated. The results of the data mining are patterns of a population as a whole and can never be traced down to individual patients.

The web application was built as an extension to the Infektionsverktyget web site but had a secret URL that only selected persons knew of and also had a login functionality. However, in its current state the data mining web application is not a real part of Infektionsverktyget and is only used for this thesis.

3.4.1 Environment

The web application was developed in Microsoft Visual Studio 2012 using the ASP.NET 4.5 framework which enables creation of dynamic web pages. The code behind was written in C#, an object oriented language similar to Java even though it is based on C++.

This environment was chosen for the data mining web application to match the rest of Infektionsverktyget, see Section 2.2. It is also an advantage that both the web application development environment and the data mining environment are Microsoft products because it makes them easier to integrate.

3.4.2 Login

Two different web pages were developed; CrossValidation for the quantitative evaluation of prediction performance and PatternExtraction for extraction of generated patterns. These pages can only be reached via a page called Default, which is launched automatically upon navigation to the specific URL, where an authentication check is performed. Before the Default page is launched the user must have inserted her or his SITHS card (with an associated SITHS certificate) in the computer card reader. When launching the Default page with an inserted SITHS card the unique ID of the user, known as the HSAID, is automatically extracted from the SITHS certificate. This HSAID is compared to the HSAIDs of the users that are authorized to log in to the page. If they are equal the user is logged in and redirected to
CHAPTER 3. METHOD

either the CrossValidation or the PatternExtraction page, depending on the current purpose. To avoid direct launches of the CrossValidation and PatternExtraction pages without a preceding login, which could pose a security threat, a session ID is used. When the user has been correctly authenticated during login via the Default page, the session ID is set to “true” and when the CrossValidation or PatternExtraction page is loaded the state of the session ID is checked. If it is not “true”, meaning that the user did not login, the user is sent to another page saying that the authentication failed.

3.4.3 Analysis Services connection

The mining structures and mining models were located on an Analysis Services server. To be able to evaluate the models a connection to the Analysis Services server had to be established. All communication with the Analysis Services server was performed using an API (Application Programming Interface) called ADOMD.NET, which is a .NET dedicated API developed specifically for Analysis Services [11, pp. 499-501]. An AdomdConnection was established using a connection string specifying the data source and catalog of the Analysis Services server. When the connection had been opened several AdomdCommands, in this case DMX queries, were created and executed. Execution of an AdomdCommand will result in an AdomdDataReader which was used to read the result of the query row by row. Most of the AdomdCommands that were used to perform cross validation and extract model patterns were SQL Server system stored procedures, i.e. predefined scripts that the system uses itself but also can be used by others. SQL Server Profiler was used to discover which system stored procedures SQL Server Management Studio uses when a user performs cross validation or browses a model in the SQL Server Management Studio interface. This had to be done to be able to perform the cross validation and model browsing in the web application instead of SQL Server Management Studio. Profiler is a program for tracing communication to and from a SQL Server instance and can be used to see what commands are being called.

3.4.4 CrossValidation page

After login and redirection to the CrossValidation page the web site looks as in Figure 3.3. The model names of the check box lists were changed between round one and two of the quantitative evaluation of predictive performance. In the drop down list the user selects the infection outcome to predict and then indicates which models to evaluate in the check box lists.

When the “Load Results”-button is clicked, the ticked models are added to a list and for every model in this list cross validation is performed with the chosen infection outcome as the predictable output state. The cross validation is performed on the Analysis Services server with a predefined
3.4. WEB APPLICATION DEVELOPMENT

Data Mining Web Application

Select infection outcome to predict and models to use. Click button to load cross validation results.

- All clinics, Naive Bayes, input 100
- All clinics, Naive Bayes, input 300
- All clinics, Association Rules, min probability 20%
- All clinics, Association Rules, min probability 40%
- All clinics, Decision Trees, input 100, min support 10, penalty 25%
- All clinics, Decision Trees, input 100, min support 50, penalty 50%
- All clinics, Decision Trees, input 300, min support 50, penalty 50%
- All clinics, Decision Trees, input 100, min support 50, penalty 75%
- All clinics, Decision Trees, input 300, min support 50, penalty 75%

Figure 3.3: Screenshot of the CrossValidation page after login, as used in round one.

For each partition of the cross validation the query result includes the number of TP, FP, TN and FN. For each of these an object of the type ResultNode is created. Because there are five partitions this results in 20 ResultNodes for each model. From this list, the average number of FP, FP, TN and FN over the five partitions is calculated for each model and new ResultNodes are created with the calculated numbers and added to a new list. The new list is then connected as a datasource to a gridview web control which dynamically displays the content of a list of objects in the form of a table.

The UML (Unified Modeling Language) diagram of the CrossValidation page is displayed as Figure 3.4.

3.4.5 PatternExtraction page

When the user has been logged in and redirected to the PatternExtraction page the web site looks as in Figure 3.5. The user chooses data mining model in the drop down list and clicks the button to load the patterns of the chosen model.

The contents displayed depend on the data mining algorithm of the chosen model. For the Association Rules algorithm, two gridviews containing association rules are displayed. One of them displays rules without any filter applied and the other one displays rules with a filter that only includes rules containing an infection outcome that is not “No infection”. Because infections are relatively rare the rules including them were suspected to become lost when no filter was used. The system stored procedure System.GetRules was used with slightly different parameter settings depending on the filter.
applied. The query results are turned into ARRuleNodes and displayed in the gridviews.

For the Naive Bayes algorithm the attribute characteristics of all infection outcomes except “No infection” are displayed in six separate gridviews. The attribute characteristics are retrieved via six subsequent calls using the system stored procedure System.GetAttributeCharacteristics with slightly
3.4. WEB APPLICATION DEVELOPMENT

different parameter settings depending on infection outcome. The result of each query is turned into a list of NBCharNodes that are displayed in a gridview.

The UML diagram of the PatternExtraction page is displayed as Figure 3.6.

Figure 3.6: UML class diagram of the code behind the PatternExtraction page.
Chapter 4

Result

4.1 Metadata

Table 4.1 contains the total number of care events in the data used for each data mining structure. Table 4.2 lists, for each of the data mining structures, the number of care events that resulted in each outcome of antibiotic treated HAI.

<table>
<thead>
<tr>
<th>Data mining structure</th>
<th>Number of care events</th>
</tr>
</thead>
<tbody>
<tr>
<td>All clinics</td>
<td>141 167</td>
</tr>
<tr>
<td>Women’s health clinic</td>
<td>26 536</td>
</tr>
</tbody>
</table>

Table 4.1: The total number of care events included in each data mining structure.

According to Table 4.2, sepsis was only detected in less than 0.02% of the care events in all clinics and even less in women’s health clinic. The unrealistically low figure was found to be caused by use of an outdated Snomed CT code which resulted in that not all sepsis cases were found. For this reason the influence of the performance of predicting sepsis was eliminated from all the following results.

4.2 Quantitative evaluation

Table 4.3 and 4.4 contain the macro-averaged performance measures generated in both rounds of cross validation for all clinics and women’s health clinic.

There were no results acquired for model 6-10, all based on the Decision Trees algorithm, for women’s health clinic, represented by hyphens in Table 4.4. The reason for this was that performing cross validation on those models resulted in an error, probably caused by a combination of a smaller dataset and harder pruning. There was no error during cross validating of the same models with more data available (from all clinics, models 6-10 in Table 4.3)
4.3. EXTRACTED PATTERNS

<table>
<thead>
<tr>
<th>Infection outcome</th>
<th>All clinics</th>
<th>Women’s health clinic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of cases</td>
<td>Proportion [%]</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>725</td>
<td>0.514</td>
</tr>
<tr>
<td>Urinary tract infection</td>
<td>1853</td>
<td>1.31</td>
</tr>
<tr>
<td>Clostridium difficile infection</td>
<td>525</td>
<td>0.372</td>
</tr>
<tr>
<td>Superficial postoperative infection</td>
<td>2357</td>
<td>1.67</td>
</tr>
<tr>
<td>Deep postoperative infection</td>
<td>1313</td>
<td>0.930</td>
</tr>
<tr>
<td>Sepsis</td>
<td>27</td>
<td>0.0191</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>6800</strong></td>
<td><strong>4.82</strong></td>
</tr>
</tbody>
</table>

**Table 4.2:** The number of cases of each infection outcome. The proportion columns indicate the proportion of each infection outcome out of all the care events.

<table>
<thead>
<tr>
<th>Model number</th>
<th>Recall [%]</th>
<th>Precision [%]</th>
<th>F-measure [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23.5</td>
<td>8.93</td>
<td>10.2</td>
</tr>
<tr>
<td>2</td>
<td>28.0</td>
<td>7.47</td>
<td>9.68</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0.0108</td>
<td>1.43</td>
<td>0.0214</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0.200</td>
<td>8.95</td>
<td>0.390</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>28.3</td>
<td>7.51</td>
<td>9.92</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>1.10</td>
<td>40.2</td>
<td>2.14</td>
</tr>
<tr>
<td>14</td>
<td>0.677</td>
<td>28.9</td>
<td>1.32</td>
</tr>
<tr>
<td>15</td>
<td>0.407</td>
<td>13.1</td>
<td>0.789</td>
</tr>
<tr>
<td>16</td>
<td>0.561</td>
<td>14.4</td>
<td>1.07</td>
</tr>
</tbody>
</table>

**Table 4.3:** Results of cross validation for all clinics. Model numbers refer to Table 3.2 and 3.3.

or with the same amount of data but less pruning (models 13-14 in Table 4.4) which motivates this as a probable cause of the error.

### 4.3 Extracted patterns

Patterns generated by the models listed in Table 3.5, based on the Association Rules and Naive Bayes algorithms, were extracted. There was an enormous amount of patterns extracted and therefore only a few examples,
CHAPTER 4. RESULT

<table>
<thead>
<tr>
<th>Model number</th>
<th>Recall [%]</th>
<th>Precision [%]</th>
<th>F-measure [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.6</td>
<td>2.79</td>
<td>4.53</td>
</tr>
<tr>
<td>2</td>
<td>22.5</td>
<td>5.63</td>
<td>8.43</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>19.8</td>
<td>7.26</td>
<td>9.79</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>7.22</td>
<td>40.1</td>
<td>11.4</td>
</tr>
<tr>
<td>14</td>
<td>0.116</td>
<td>2.22</td>
<td>0.220</td>
</tr>
<tr>
<td>15</td>
<td>0.835</td>
<td>20.3</td>
<td>1.57</td>
</tr>
<tr>
<td>16</td>
<td>2.20</td>
<td>11.1</td>
<td>3.63</td>
</tr>
</tbody>
</table>

Table 4.4: Results of cross validation for women’s health clinic. Model numbers refer to Table 3.2 and 3.3

that were considered the most interesting, are listed below.

Action and diagnosis terms are registered in Infektionsverktyget in Swedish. For convenience they are translated to English in the tables below. Diagnosis terms are translated using the International version of ICD-10.

4.3.1 Association Rules algorithm

Models based on the Association Rules algorithm resulted in patterns in the form of rules with associated importance, probability and support. The rules generated were identical both with and without the filter described in Section 3.4.5. Examples of rules generated with Association Rules for all clinics and women’s health clinic are displayed in Table 4.5 and 4.6, respectively.

4.3.2 Naive Bayes algorithm

Patterns extracted with the Naive Bayes algorithm were in the form of attribute characteristics of each infection outcome, with an associated frequency.

For both all clinics and women’s health clinic, no attribute characteristics influencing pneumonia were found. For the remaining four infection outcomes, examples of attribute characteristics influencing each of them are listed in Table 4.7 and 4.8 for all clinics and women’s health clinic, respectively.
### Table 4.5: Examples of rules extracted for all clinics.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Support</th>
<th>Probability</th>
<th>Importance</th>
</tr>
</thead>
</table>
| Diagnoses(Other disorders of the skin and subcutaneous tissue) = Existing,  
Diagnoses(Intestinal infectious diseases) = Existing ⇒ Infection = Clostridium difficile infection | 20      | 0.222       | 1.80       |
| Actions(Operations on the stomach and duodenum) = Existing,  
Diagnoses(Influenza and pneumonia) = Existing ⇒ Infection = Pneumonia | 11      | 0.262       | 1.73       |
| Diagnoses(Malignant neoplasms of bone and articular cartilage) = Existing,  
Age = 26-48 ⇒ Infection = Deep postoperative infection | 13      | 0.448       | 1.69       |
| Diagnoses(Lung diseases due to external agents) = Existing,  
Actions(Surgeries on the stomach and duodenum) = Existing ⇒ Infection = Pneumonia | 10      | 0.233       | 1.68       |
| Actions(Reoperations on the chest wall, pleura, mediastinum, diaphragm, trachea, bronchi and lungs) = Existing,  
Actions(Reoperations on the heart and intrathoracic vessels) = Existing ⇒ Infection = Superficial postoperative infection | 19      | 0.792       | 1.67       |
| Diagnoses(Influenza and pneumonia) = Existing,  
Diagnoses(Falls) = Existing ⇒ Infection = Pneumonia | 46      | 0.131       | 1.44       |
### Table 4.6: Examples of rules extracted for women’s health clinic.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Support</th>
<th>Probability</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnoses(Surgical operation and other surgical procedures as the cause of abnormal reaction of the patient, or of later complication, without mention of misadventure at the time of the procedure) = Existing, Actions(Bilateral salpingo-oophorectomy) = Existing, Actions(Hysterectomy) = Existing ⇒ Infection = Superficial postoperative infection</td>
<td>10</td>
<td>0.556</td>
<td>1.76</td>
</tr>
<tr>
<td>Diagnoses(Unintentional cut, puncture, perforation or haemorrhage during surgical and medical care) = Existing, Diagnoses(Leiomyoma of uterus) = Existing, Age &gt;= 37 ⇒ Infection = Superficial postoperative infection</td>
<td>14</td>
<td>0.467</td>
<td>1.69</td>
</tr>
<tr>
<td>Diagnoses(Malignant neoplasm of ovary) = Existing, Age &gt;= 37 ⇒ Infection = Urinary tract infection</td>
<td>14</td>
<td>0.110</td>
<td>1.42</td>
</tr>
<tr>
<td>Actions(Abdominal caesarean section on isthmus) = Existing, Actions(Oxytocin infusion for stimulation of obstetric labor) = Existing, Actions(Catheterization of urinary bladder) = Existing, Age = 20-31 ⇒ Infection = Deep postoperative infection</td>
<td>65</td>
<td>0.015</td>
<td>1.23</td>
</tr>
</tbody>
</table>
### Table 4.7: Examples of attribute characteristics extracted for all clinics.

<table>
<thead>
<tr>
<th>Influences</th>
<th>Attribute</th>
<th>State</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTI</td>
<td>Gender</td>
<td>Female</td>
<td>0.645</td>
</tr>
<tr>
<td>UTI</td>
<td>Age</td>
<td>&gt;=78</td>
<td>0.506</td>
</tr>
<tr>
<td>UTI</td>
<td>Diagnoses(Hypertensive diseases)</td>
<td>Existing</td>
<td>0.442</td>
</tr>
<tr>
<td>UTI</td>
<td>Gender</td>
<td>Male</td>
<td>0.355</td>
</tr>
<tr>
<td>UTI</td>
<td>Diagnoses(Persons with potential health hazards related to family and personal history and certain conditions influencing health status)</td>
<td>Existing</td>
<td>0.354</td>
</tr>
<tr>
<td>UTI</td>
<td>Diagnoses(Other diseases of urinary system)</td>
<td>Existing</td>
<td>0.308</td>
</tr>
<tr>
<td>UTI</td>
<td>Age</td>
<td>66-78</td>
<td>0.255</td>
</tr>
<tr>
<td>CDI</td>
<td>Gender</td>
<td>Female</td>
<td>0.531</td>
</tr>
<tr>
<td>CDI</td>
<td>Age</td>
<td>&gt;=78</td>
<td>0.512</td>
</tr>
<tr>
<td>CDI</td>
<td>Gender</td>
<td>Male</td>
<td>0.469</td>
</tr>
<tr>
<td>CDI</td>
<td>Diagnoses(Hypertensive diseases)</td>
<td>Existing</td>
<td>0.345</td>
</tr>
<tr>
<td>CDI</td>
<td>Diagnoses(Other forms of heart disease)</td>
<td>Existing</td>
<td>0.331</td>
</tr>
<tr>
<td>CDI</td>
<td>Age</td>
<td>66-78</td>
<td>0.261</td>
</tr>
<tr>
<td>CDI</td>
<td>Diagnoses(Intestinal infectious diseases)</td>
<td>Existing</td>
<td>0.261</td>
</tr>
<tr>
<td>SPI</td>
<td>Gender</td>
<td>Female</td>
<td>0.532</td>
</tr>
<tr>
<td>SPI</td>
<td>Gender</td>
<td>Male</td>
<td>0.468</td>
</tr>
<tr>
<td>SPI</td>
<td>Diagnoses(Complications of surgical and medical care, not elsewhere classified)</td>
<td>Existing</td>
<td>0.329</td>
</tr>
<tr>
<td>SPI</td>
<td>Age</td>
<td>48-66</td>
<td>0.279</td>
</tr>
<tr>
<td>SPI</td>
<td>Age</td>
<td>66-78</td>
<td>0.223</td>
</tr>
<tr>
<td>SPI</td>
<td>Age</td>
<td>&gt;=78</td>
<td>0.202</td>
</tr>
<tr>
<td>DPI</td>
<td>Gender</td>
<td>Female</td>
<td>0.518</td>
</tr>
<tr>
<td>DPI</td>
<td>Age</td>
<td>66-78</td>
<td>0.278</td>
</tr>
<tr>
<td>DPI</td>
<td>Age</td>
<td>48-66</td>
<td>0.257</td>
</tr>
<tr>
<td>DPI</td>
<td>Age</td>
<td>&gt;=78</td>
<td>0.249</td>
</tr>
<tr>
<td>DPI</td>
<td>Diagnoses(Hypertensive diseases)</td>
<td>Existing</td>
<td>0.219</td>
</tr>
<tr>
<td>DPI</td>
<td>Diagnoses(Complications of surgical and medical care, not elsewhere classified)</td>
<td>Existing</td>
<td>0.218</td>
</tr>
<tr>
<td>Influences</td>
<td>Attribute</td>
<td>State</td>
<td>Frequency</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------------------------------------------------------------</td>
<td>------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>UTI</td>
<td>Gender</td>
<td>Female</td>
<td>1.00</td>
</tr>
<tr>
<td>UTI</td>
<td>Age &gt; = 37</td>
<td></td>
<td>0.557</td>
</tr>
<tr>
<td>UTI</td>
<td>Actions(Catheterization of urinary bladder)</td>
<td>Existing</td>
<td>0.496</td>
</tr>
<tr>
<td>UTI</td>
<td>Age 31-37</td>
<td></td>
<td>0.244</td>
</tr>
<tr>
<td>CDI</td>
<td>Gender</td>
<td>Female</td>
<td>1.00</td>
</tr>
<tr>
<td>CDI</td>
<td>Actions(Catheterization of urinary bladder)</td>
<td>Existing</td>
<td>0.600</td>
</tr>
<tr>
<td>CDI</td>
<td>Age 1-20</td>
<td></td>
<td>0.6</td>
</tr>
<tr>
<td>CDI</td>
<td>Actions(Clinical examination in anaesthesia)</td>
<td>Existing</td>
<td>0.200</td>
</tr>
<tr>
<td>CDI</td>
<td>Actions(Lumbar epidural block with catheter)</td>
<td>Existing</td>
<td>0.200</td>
</tr>
<tr>
<td>CDI</td>
<td>Diagnoses(Other puerperal infections)</td>
<td>Existing</td>
<td>0.200</td>
</tr>
<tr>
<td>CDI</td>
<td>Age 20-31</td>
<td></td>
<td>0.200</td>
</tr>
<tr>
<td>CDI</td>
<td>Diagnoses(Single spontaneous delivery)</td>
<td>Existing</td>
<td>0.200</td>
</tr>
<tr>
<td>SPI</td>
<td>Gender</td>
<td>Female</td>
<td>0.996</td>
</tr>
<tr>
<td>SPI</td>
<td>Age &gt; = 37</td>
<td></td>
<td>0.800</td>
</tr>
<tr>
<td>SPI</td>
<td>Actions(Catheterization of urinary bladder)</td>
<td>Existing</td>
<td>0.411</td>
</tr>
<tr>
<td>DPI</td>
<td>Gender</td>
<td>Female</td>
<td>1.00</td>
</tr>
<tr>
<td>DPI</td>
<td>Age &gt; = 37</td>
<td></td>
<td>0.624</td>
</tr>
<tr>
<td>DPI</td>
<td>Actions(Catheterization of urinary bladder)</td>
<td>Existing</td>
<td>0.457</td>
</tr>
<tr>
<td>DPI</td>
<td>Age 31-37</td>
<td></td>
<td>0.202</td>
</tr>
</tbody>
</table>

Table 4.8: Examples of attribute characteristics extracted for women’s health clinic.
Chapter 5

Discussion

As explained in Section 3.4 it was not possible to work directly with the patient data which added some complication to the study. A certain amount of time had to be spent on development of the web application which limited the time available for the other parts of the study. Also, the evaluation of models was substantially slower and more inflexible than it could have been working directly with the patient data.

5.1 Metadata

According to Table 4.2 the proportion of HAIs found with the time frames stated in Section 3.2.2 was 4.82% for all clinics and 2.20% for women’s health clinic. One explanation that these numbers are lower than the point prevalence numbers presented in Section 2.1 is the erroneous selection of sepsis cases explained in Section 4.1. Another reason is that this study only was concerned with cases of HAIs that were treated with antibiotics because these are the infections that are registered in Infektionsverktyget. This will lower the proportion of HAIs found because not all HAIs are treated with antibiotics. A third contributor to the lower numbers is that it is possible that not even all HAIs treated with antibiotics were detected. In this study only infections debuting more than 48 hours after hospital admission are considered associated to the care event. Although in the definition of a HAI, see Section 2.1, it is stated that also infections debuting within 48 hours can be healthcare-associated if they can be traced to a medical action. On the other hand, it is also possible that care events and infections that have no actual relation have been connected. This might have happened if for instance a patient has frequent visits to the hospital and the time frames of each care event overlap each other.

The reason why the proportion of detected infections was lower for women’s health clinic is likely that this clinic is mainly dealing with pregnancies and births. These patients are usually not affected by the factors
5.2 QUANTITATIVE EVALUATION

contributing to an increased susceptibility of infections, presented in Section 2.1.

5.2 Quantitative evaluation

It was not possible to retrieve any cross validation results for model 6-10 for women’s health clinic, as described in Section 4.2. However, this was not considered a major drawback to the study because the same models performed poorly for all clinics and it is reasonable to believe that they would behave similarly for women’s health clinic.

It can be argued which of the evaluation measures recall and precision is the most important or whether they are equally important and emphasis should be on their harmonic mean, F-measure. In this study, the cost of decreased recall, i.e. increased false negative error, is that patients with increased risk of incurring a HAI are not detected as intended. This may lead to prolonged hospital stays, increased cost and unnecessary human suffer. On the other hand, the cost of decreased precision, i.e. increased false positive error, is that safety precautions are taken although the patient is not in increased danger of incurring a HAI. This may lead to unnecessary treatments and worry. However, in this study a higher recall is considered more important than a higher precision, because the whole purpose is to find risk patients. It is better if some patients are incorrectly treated as risk patients than if actual risk patients go undetected. Ideally both recall and precision would be high, but there is often a trade-off between them.

In this study, Naive Bayes was the algorithm that performed best with respect to recall but with this came a precision of less than 10%. An increased number of maximum input attributes seemed to be beneficial to some extent. These tendencies were the same for both all clinics and women’s health clinic even though the performance was somewhat better for all clinics. Although it was the best, it has a lot of room for improvement because the numbers are not great for neither recall nor precision.

The Decision Trees algorithm performed better with less pruning both for all clinics and women’s health clinic. The reason for this was that the infections are relatively rare in the dataset and are easily pruned away from the resulting tree. With less pruning, more cases of infection are left in the tree and they can be used to classify new cases of data, which increases the predictive performance. Although less pruning resulted in better performance, there is always a certain degree of pruning needed to avoid overtraining of the model. The Decision Trees models using the least amount of pruning had a high precision leading to an increased F-measure, but the recall was considered poor.

The Logistic Regression and Neural Network algorithms did not perform well even though they should be able to identify complex relation-
ships between inputs and outputs. This indicates that the simplicity of the other algorithms was not the main reason for their relatively poor performance, because even the more advanced algorithms performed equally bad or even worse. The overall biggest problem is considered to be the imbalanced dataset with respect to infectious states and the non-infectious state.

It was proven that the Association Rules algorithm was useless for classification of infections, because it classified all cases as the majority state; “No infection”. However, this was not surprising because the algorithm is mainly designed for descriptive purposes. The reason for the bad performance is believed to be the large variation and the large number of states in the data. This makes it hard for the algorithm to find rules including an infection that can be applied to the new cases of data and instead all cases are classified as “No infection”.

5.3 Extracted patterns

The patterns generated by the algorithms Association Rules and Naive Bayes were all easily interpreted and understood if the meaning of importance, frequency etc. was known. However, the rule form of the Association Rules patterns is believed to be the easiest to absorb for a person that is not familiar with data mining.

One problem that was not caused by the algorithms themselves but by the preprocessing is that some of the grouped terms are rather vague and can include a wide variety of diagnoses and actions. An example of such a term is “Persons with potential health hazards related to family and personal history and certain conditions influencing health status”, which needs a manual lookup to reveal what diagnoses are hidden behind it. Although somewhat of a problem, a certain degree of grouping is needed, as discussed in Section 3.2.2.

5.3.1 Association Rules

Many of the rules included trivial relationships, but also more interesting correlations. An example is the second rule of Table 4.5, “Actions(Operations on the stomach and duodenum) = Existing, Diagnoses(Influenza and pneumonia) = Existing ⇒ Infection = Pneumonia”. The diagnosis “Influenza and pneumonia” is the same as the infection outcome, and is obviously not an interesting risk factor. The action “Operations on the stomach and duodenum” on the other hand is not trivially related to the infection outcome and can be clinically valid. This according to IT consultant and M.D. Rikard Lövström, who is also a board member of The Board for Pharmaceuticals, IT and Medical Technology of Sveriges Läkarförbund, the Swedish Medical Association. It is possible that actions in the airways, such as gastroscopies, can cause pneumonia.
There were also other rules that were found to have clinical relevance. Rule number three of Table 4.5 can possibly be explained by an increased susceptibility to infections during chemotherapy, says Lövström. The surgeries included in rule number 5 of Table 4.5 and number 1 of Table 4.6 usually include large incisions that can be infected postoperatively. The Caesarean section included in rule number 5 of Table 4.6 is a large intervention and it is not hard to understand why it may be linked to DPIs.

Rules that were not believed to include actual risk factors also emerged. Rule number 6 of Table 4.5 is one example, in which the diagnosis “Falls” is included. It is likely that it is not the actual fall that is the risk factor, but rather high age (which is a risk factor common of both fall accidents and pneumonia) or the hospital visit following the accident. This shows that the rules generated cannot be taken as true without being manually checked and potentially investigated by professionals, e.g. epidemiologists or physicians.

5.3.2 Naive Bayes

In the current form of the attribute characteristics, it is hard to determine whether attribute-state pairs scoring high frequencies are doing so because they are great influencers or just because they are commonly occurring in the patient data. Studying the attribute characteristics influencing the DPIs in Table 4.7, both “Gender = Female” and “Gender = Male” have the highest frequencies of all with 0.518 and 0.482, respectively. In the overall dataset for all clinics these numbers are assumed to be 0.5 and 0.5. With numbers close to that these attribute-state pairs cannot be considered related to the outcome of DPIs even though they top the frequency list. It is also misleading when comparing frequencies of two-state attributes with multi-state attributes. Instead of studying the absolute frequencies it would be better to compare them to the frequencies in the whole dataset.

Because of this defect, it is hard to discuss the clinical validity of the attribute characteristics of Table 4.7 and 4.6 involving a diagnosis or action without knowing their frequencies in the whole dataset. As mentioned above the assumed gender distribution for all clinics is 0.5/0.5 and the frequency of each age interval is known to be 0.2 (five intervals with equally many cases). With this in mind, Table 4.7 states that being a female and/or of high age are risk factors of UTIs, because of their increased frequencies as compared to the whole dataset. These risk factors are not new discoveries but known medical facts that proves that clinically valid relationships can be generated also with the Naive Bayes algorithm, although somewhat harder to interpret directly.

50
5.4 Improvements and future work

During this study a number of possible improvements for further studies have been identified. The first and most obvious is to use the correct Snomed CT code for sepsis, i.e. the code that it is registered under in Infektionsverktyget.

A second improvement would be to include all HAIs independently of whether they have been treated with antibiotics or not. Achieving this would require involvement of more information in the database, for instance the diagnoses and laboratory test results could be searched for infections. However, doing so would lead to a new issue; determining whether the infections found are healthcare-associated and should be included. This was not a problem using the PrescriptionCause table because there the infections are registered as either healthcare or community acquired.

Thirdly, removing the redundant diagnoses could be an improvement to both the quality of the extracted patterns and the testing of the predictive performance. Redundant are those diagnoses that indicate the same types of infections that are present as infection outcomes in the data mining. Removing those would result in less trivial patterns such as “Influenza and pneumonia = Existing ⇒ Infection = Pneumonia” which is uninteresting because the pneumonia diagnosis cannot be a risk factor of pneumonia, they are the same. In prediction the algorithms would not be able to “cheat”, i.e. use the infection diagnoses to predict the outcomes. A disadvantage with removing redundant diagnoses could be that also potentially interesting patterns, such as the made-up rule “Influenza and pneumonia = Existing ⇒ Infection = Clostridium difficile infection”, are lost. However, no rules including an infection diagnoses leading to another type of infection were found in this study, so it is considered a minor future problem.

One of the biggest problems identified in the quantitative evaluation was the imbalanced dataset with respect to infection outcomes, which gave rise to rather poor predictive performance of the models. There are a number of different possible solutions to be tried, as suggested by M. Maalouf [26]. One that would be relatively easy to implement is under- and/or oversampling, in which the model is trained on a dataset with a modified proportion of infection outcomes and tested on a dataset with original proportions. It is not clear whether this would be enough but it can be used as a starting point.

Future studies should be focused on looking over and implementing the areas of improvement introduced above as well as on investigating the legal aspects of gathering of more patient data. More patient data is needed to even out human introduced differences in the data and to distinguish actual risk factors. In a smaller dataset differences may arise if a prescriber within a specific field of medicine is more prone to prescribe antibiotics or if prescribers tend to classify the infections differently. Another aspect of future studies could be to study rules generated by Association Rules that have a
lower support and/or probability, as suggested by Brosette et al. [20] and Ma et al. [21]. This would reveal less frequently occurring but potentially even more interesting relationships. Rikard Lövström says that with more source data available it could be really interesting for epidemiologists and researchers to review and try to interpret the extracted patterns. Further Lövström comments that the possibility of discovering risk factors previously unaware of, is interesting and valuable.

During the course of this project some ideas regarding how the infection outcome prediction can be used in practice in the future have arisen. The purpose is to identify patients with increased risk of incurring a HAI so that these patients can get extra attention in prevention of infections. The foundation is a trained data mining model that represents the clinical reality well and is able to predict infection outcomes with as high recall and precision as possible. Input data, i.e. gender, age, previous diagnoses and actions, on a patient could be fed either manually or automatically to the model. With a prediction join between the trained model and the new data the probabilities of each of the seven infection outcomes, including “No infection”, are returned. These probabilities can then be compared to normal probabilities of each infection outcome and significant increases can be detected. The advantage is that not only the predicted state, i.e. the infection outcome scoring the highest probability, is considered. Increases in probability also among the less probable outcomes can be detected and more potential risk patients can be identified. This implies that a model would not need to be 100% accurate in terms of recall or precision when predicting a specific outcome, but could be useful anyway to detect risk increases. As an example, it is the most probable that a patient will not incur a HAI and a data mining model may therefore score the outcome “No infection” the highest probability. Although not the most probable outcome, the patient can have an increased risk of incurring a specific HAI, which will be detected if also less probable outcomes are considered. The identification of risk patients could either be performed by nurses and physicians upon request or be incorporated into the medical record system for automatic identification and notification.
Chapter 6

Conclusion

Data mining models based on five different data mining algorithms were developed to investigate if they are suitable to describe and predict relationships between gender, age, previous diagnoses and actions as well as antibiotics-treated HAI outcome. A web application was developed to overcome legal restrictions. The conclusions that were reached are presented below.

- Models based on the Naive Bayes algorithm had the highest recall and performed somewhat better for all clinics versus only women’s health clinic. Models based on the Decision Trees algorithm with low pruning had the highest precision, but the recall was poor. No model was considered to have achieved a sufficient predictive performance.

- Extracted patterns revealed some clinically valid, some trivial and some probably clinically invalid relationships. A professional is needed to sort them out. Patterns generated by the Association Rules algorithm were favoured over those generated by the Naive Bayes algorithm.

- Several areas of improvement were identified. Data mining in Infektionsverktyget has the possibility of becoming an asset to both research and clinical work if these areas are handled and more patient data is gathered.
Bibliography


Appendix A

ICD-10-SE hierarchy

Figure A.1: Extract from the ICD-10-SE hierarchy tree.
Appendix B

KVÅ hierarchy

Figure B.1: Extract from the KVÅ hierarchy tree.