Knowledge Discovery and Anomalies

—

towards a Dynamic Decision-making Model for Medical Informatics

By

Heidi Arndt

Thesis presented in fulfilment of the requirements for the degree of Doctor of Philosophy in the Faculty of Arts and Social Sciences at Stellenbosch University

Supervisor: Prof J Kinghorn

March 2018

DECLARATION:

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2018

Opsomming

Wêreldwyd het gesondheidsorg 'n groot kommer geword vir die moderne samelewing wat uitgedaag word om gehalte versorging toeganklik en bekostigbaar vir almal te maak. Met 'n tanende wêreld ekonomie, toenemende mediese koste en 'n beter ingeligte kliëntebasis, is regerings sowel as gesondheidsorg-organisasies onder toenemende druk om 'n produk te lewer wat fokus op gehalte versorging, deursigtige koste en 'n uitstekende pasiënt ervaring. Dit vereis goed ingeligte en doelgerigte operasies en besluitneming deur gesondheidsorg-organisasies wat weer druk plaas op inligtingstelsels van die gesondheidsorg sektor.

In 'n omvattende literatuurstudie is bevind dat gesondheidsorg-organisasies organisasies is wat bestaan uit 'n wye verskeidenheid subsisteme in ’n komplekse omgewing. Daarbenewens is daar eienaardighede wat die ontwikkeling van gesondheid-inligtingstelsels uitdaag. Bisosiatiewe kennis ontdekking, wat die kreatiewe ontdekking van voorheen onbekende inligting uit onverenigbare domeine is, is ontwikkel as 'n alternatiewe instrument om die behoefte aan besluitneming-ondersteuning in die gesondheidsorg-sektor aan te spreek. Daar is verder bevind dat inligting-netwerke 'n nuttige manier is om data te integreer vanaf onverenigbare domeine. Laastens is gereelde patroon-ontginning geïdentifiseer as die masjienleer-instrument vir die ontginning van bisosiatiewe assosiasies binne inligting-netwerke.

‘n Kennis-ontdekking-raamwerk vir data-intensiewe navorsing gefokus op die veld van biomediese informatika is in hierdie studie ontwikkel. Binne hierdie raamwerk word data as ‘n geïntegreerde, heterogene inligting-netwerk en masjienleer-algoritmes voorgestel, met die uitdruklike doel om inter-konneksies binne hierdie strukture te vind wat tot bisosiatiewe kennis-ontdekking kan lei. Hierdie raamwerk is verder ontwikkel tot 'n kennis-ontdekking-proses model vir bisosiatiewe kennis-ontdekking, gefokus op die gesondheidsorg-sektor.

Die kennis-ontdekking-model vir bisosiatiewe kennis ontdekking is dan toegepas in ‘n gevallestudie wat gebruik gemaak het van die “Nationwide Inpatient Sample” data wat deel uitmaak van die “Healthcare Cost and Utilization Project.” Die gevallestudie het die konstruksie van onverenigbare domeine en hul integrasie in 'n heterogene inligting-netwerk suksesvol gedemonstreer. Verder het dit die toepassing van gereelde patroon-ontdekking-algoritmes gedemonstreer om sub-grafieke uit die gekonstrueerde inligting-netwerk te onttrek. Dit is gevolg deur die opbou van die onttrokke sub-grafieke as konsep-grafieke met die doel om die resultate te visualiseer vir verdere interpretasie.

Die volgende gevolgtrekkings is gemaak uit die resultate van die navorsing:

— Die voorgestelde verkennende data-ontginningsmetode wat gebruik maak van bisosiatiewe kennis-ontdekking het onverwagte, potensieel interessante verwantskappe binne die gekonstrueerde inligting-netwerk openbaar.

— Die modellering van gesondheidsorg data as 'n inligting-netwerk het visuele insig in die struktuur van die data toegelaat, wat die opsporing van nuwe insigte ondersteun het wat andersins nie geopenbaar sou wees nie.

— En het bewys dat organisasies in komplekse omgewings suksesvol ontleed kan word in ryk lae van abstraksie en die integrasie van hierdie lae kan geoutomatiseer word deur berekening.

Summary

Worldwide healthcare has become a major concern for modern society, which is challenged to make quality care accessible and affordable to all. With a slowing world economy, rapidly rising medical costs and a better-informed customer base, governments and healthcare organisations are under pressure to deliver a product that focuses on quality care, transparent costs and an excellent patient experience. This requires well-informed and nimble operating and decision-making by healthcare organisations, putting pressure on the discipline of informatics within systems.

In a comprehensive literature survey, it was found that healthcare organisations are organisations made up of a wide variety of subsystems operating in a complex environment. In addition, there are individualities that challenge the development of health information systems. Bisociative knowledge discovery, which is the creative discovery of previously unknown information from habitually incompatible domains, was introduced as an alternative tool to address the need for decision support in the healthcare sector. It was further found that information networks are a useful way to integrate data from habitually incompatible domains. Lastly, frequent pattern mining was identified as the machine learning tool for mining bisociations within information networks.

A knowledge discovery framework for data-intensive research focusing on the field of biomedical informatics was developed in this study. Within this framework, data are represented as integrated, heterogeneous information networks, and machine learning algorithms are applied to the data with the explicit purpose of finding interconnectedness within these structures that can lead to bisociative knowledge discoveries. This framework was further developed into a knowledge discovery process model for bisociative knowledge discovery with a focus on the healthcare sector.

The knowledge discovery process model for bisociative knowledge discovery was then applied in a case study which made use of the Nationwide Inpatient Sample data that forms part of the Healthcare Cost and Utilization Project. The case study successfully demonstrated the construction of habitually incompatible domains and their integration into a heterogeneous information network. Furthermore, it demonstrated the application of frequent pattern mining algorithms to extract subgraphs from the constructed information network. This was followed by the constructing of the extracted subgraphs as concept graphs with the purpose of visualising the results for further interpretation.

At the end of this research it was concluded that:

- The proposed explorative data mining method using bisociative knowledge discovery revealed unexpected, potentially interesting relationships within the constructed information network.
- Modelling data from the healthcare sector as an information network allowed visual insights into the structure of the data, which supported the detection of novel insights that otherwise would not have been revealed.
- Organisations operating in a complex environment can be successfully unpacked into rich layers of abstraction and the integration of these layers can be automated through computing.

Acknowledgements

I would hereby like to express my sincere thanks and gratitude towards:

- Wikus van Niekerk for his encouragement and interest.
- Prof J Kinghorn for his guidance and assistance.
- The financial assistance of the National Research Foundation (NRF) towards this study is hereby acknowledged.
- SNOMED International for making available SNOMED CT.
- Healthcare Cost and Utilization Project for making available the nationwide inpatient data.
- Opinions expressed and conclusions drawn in this dissertation are those of the author and are not necessarily to be attributed to the NRF, SNOMED International, or the Healthcare Cost and Utilization Project.

Contents

Chapter 1 Introduction 1

1.1. Research problem 4

1.1.1. Research Question 8

1.2. Research aim and objectives 9

1.3. Research approach 10

1.4. Chapter layout 12

Chapter 2 Literature review 13

2.1 Biomedical informatics 14

2.1.1 The science of biomedical informatics 16

2.1.2 Clinical informatics, a discipline within biomedical informatics 23

2.1.3 The uniqueness of the healthcare sector 25

2.2 Knowledge discovery from data 27

2.2.1 Koestler's concept of bisociation 28

2.2.2 Bisociative knowledge discovery 29

2.3 Information networks 34

2.3.1 Ontologies 36

2.3.2 Bisociative information networks 38

2.3.3 Homogeneous and heterogeneous information networks 41

2.4 Pattern recognition and machine learning 43

2.4.1 Mining heterogeneous information networks 45

2.5 Summary 49

Chapter 3 Knowledge discovery model for biomedical informatics 52

3.1 Scientific method 53

3.2 Scientific method for data-rich environments 54

3.2.1 Computational thinking 55

3.3 Modelling approach for data-centric research 58

3.3.1 Unified knowledge discovery framework 59

3.4 Knowledge discovery process model 61

3.4.1 Data preparation stage 68

3.4.2 Data mining stage 70

3.4.3 Knowledge evaluation stage 71

3.5 Summary 72

Chapter 4 Knowledge discovery in the healthcare sector, a worked case study 74

4.1 Understanding the healthcare problem domain 75

4.2 Understanding the data 77

4.2.1 Data selection 78

4.2.2 Data description 89

4.3 Data preparation 101

4.3.1 Feature selection 102

4.3.2 Data cleaning 105

4.3.3 Domain construction 112

4.3.4 Network integration 125

4.4 Data mining 130

4.4.1 Frequent pattern mining 130

4.4.2 Concept graph construction 137

4.5 Knowledge evaluation 146
4.5.1 Domain bridging concept discovery 146
4.5.2 Domain bridging subgraph discovery 149
4.6 Summary 151
Chapter 5 Summary and conclusion 154
5.1 Summary 154
5.2 Conclusions 155
5.3 Recommendations and future work 159
References 162
Appendix A 172
Appendix B 178

List of Figures

Figure 1: Research approach 11
Figure 2: Data, information and knowledge 19
Figure 3: Disciplines within Biomedical informatics 23
Figure 4: Illustration of Koestler's concept of bisociation (Dubitzky et al., 2012) 29
Figure 5: Atrial fibrillation as defined by SNOMED CT 37
Figure 6: *K*-partite graph structure of a healthcare network 40
Figure 7: Schema of a healthcare network 43
Figure 8: Unified knowledge discovery framework 60
Figure 9: DMKD process model for bisociative knowledge discovery in the healthcare sector 66
Figure 10: Anomaly distinction (adapted from Aggarwal, 2013) 76
Figure 11: SNOMED CT diagram for the solid organ transplant (procedure) concept 79
Figure 12: SNOMED CT diagram for transplant of lung procedure concept 81
Figure 13: SNOMED concept ids grouped by number of observations 86
Figure 14: Data selection and description process flow 88
Figure 15.1: Frequency distributions for transplant of kidney procedure 93
Figure 15.2: Frequency distributions for transplant of liver procedure 93
Figure 15.3: Frequency distributions for transplant of kidney procedure 94
Figure 15.4: Frequency distributions for transplant of liver procedure 94
Figure 16.1: Boxplots: Total charges with outliers 96
Figure 16.2: Boxplots: Length of stay with outliers 97
Figure 17.1: Scatter plot: Transplant of kidney (procedure) 99
Figure 17.2: Scatter plot: Transplant of liver (procedure) 100

Figure 18: Avg total charges (per procedure) within age categories 106
Figure 19: Cost & LOS per procedure type (SAS) 108
Figure 20: LOS category distribution per procedure type 110
Figure 21: Cost category distribution per procedure type 111
Figure 22: Cost domain model 113
Figure 23: Subgraph of the cost domain: heart transplants (Neo4J) 115
Figure 24: Patient domain model 116
Figure 25: Subgraph of the patient domain: Keck Hospital of USC (Neo4J) 117
Figure 26: Clinical domain model 118
Figure 27: Subgraph of the clinical domain: heart & lung transplants (Neo4J) 119
Figure 28: Cost network schema 121
Figure 29: Cost *k*-partite graph structure 121
Figure 30: Patient network schema 123
Figure 31: Patient *bi*–partite graph structure 123
Figure 32: Clinical network schema 124
Figure 33: Clinical *bi*-partite graph structure 124
Figure 34: Network integration workflow (KNIME) 126
Figure 35: Integrated, heterogeneous information network 128
Figure 36: Liver anomaly transplant procedure concept graph extraction workflow (KNIME) 138
Figure 37: Cypher query (Liver anomaly transplant procedure concept graph) 139
Figure 38: Liver anomaly concept graph 140
Figure 39: Kidney anomaly concept graph 141
Figure 40: Double lung anomaly concept graph 142
Figure 41: Single lung anomaly concept graph 143
Figure 42: Heart anomaly concept graph 144
Figure 43: Cost anomaly liver transplant concept 147

Figure 44: Domain bridging subgraph 150
Figure 45: Computational model for biomedical informatics 152

List of Tables

Table 1: Brachman's KDD process elements in the context of Mintzberg's three phases (Mintsberg, 1973)..... 5
Table 2: Data mining tasks for heterogeneous information network..... 47
Table 3: Comparison of three KD process models..... 63
Table 4: KDDNuggets poll results (Piatetsky, 2014)..... 64
Table 5: SNOMED CT refset for transplants..... 83
Table 6: Data extraction summary..... 89
Table 7: Distribution summary..... 91
Table 8: Pearson's correlation coefficient for transplant procedures..... 101
Table 9: NIS core data selected attributes..... 102
Table 10: NIS diagnosis data selected indicators..... 103
Table 11: Selected SNOMED CT concepts..... 104
Table 12: Missing values: TOTCHG..... 105
Table 13: Adjacency list of single lung transplant observations..... 131
Table 14: JIM algorithm execution values..... 133
Table 15: Extracted concept graphs..... 135

List of Figures: Appendix A

Figure A.1: Heart transplant procedure 172
Figure A.2: Double-lung transplant procedure 172
Figure A.3: Single-lung transplant procedure 173
Figure A.4: Heart transplant procedure 173
Figure A.5: Double-lung transplant procedure 174
Figure A.6: Single-lung transplant procedure 174
Figure A.7: Transplantation of heart (procedure) 176
Figure A.8: Double-lung transplant (procedure) 176
Figure A.9: Single lung transplant (procedure) 177

Chapter 1

Introduction

Healthcare is a major concern for modern society worldwide. As society is challenged to make quality medical care accessible and affordable to all, healthcare is evolving into a contentious political and economic issue (Yang & Hwang, 2006). On a fiscal level, countries are investigating various ways of dealing with an ageing population and curbing escalating healthcare costs. At the same time, on a political level, countries are trying to meet the expectations of a better-informed public who demand the best, most technologically advanced medical treatments combined with high quality services.

Population ageing is a worldwide phenomenon, and refers to the increasing proportion of older people in the population. The Population Division of the Department of Economic and Social Affairs of the United Nations publishes extensively on the change of population age structures within society. Its most recent report *World Population Ageing, 2015* (United Nations, 2015) states a 48% global increase in people aged 60 years or older over a five-year period, from 2010 to 2015. An additional 56% increase is predicted over the next 15 years. However, by definition, an ageing population is not determined by the actual growth of the number of older people in society, but rather by the percentage growth of older people in proportion to the total population growth. In 2010, 9.9% of the world's population was aged 60 or older; by 2015, 12.3% were in this group and by 2030 it should reach 16.5%. A similar trend exists in South Africa, where the

proportion of older people in 2015 was 7.7% of the total population, and is expected to reach a predicted 10.5% by 2030. Globally, the fundamental question that arises is: What will the impact of this ageing phenomenon be on healthcare systems worldwide?

The World Health Organisation reported, in the *World Report on Ageing and Health* (World Health Organization, 2015), that the impact of population ageing on healthcare systems is not yet clear. What is clear though, is that the disabilities caused by chronic diseases are strongly associated with age. This implies that within an ageing population there will be a higher number of people with disabilities and result in more people in need of care, consequently a major cost driver of healthcare in the future (United Nations, 2015).

On a political level, healthcare is one of the largest industries in the world and mostly funded by governments. The healthcare industry contributes globally, as reported by the World Bank *(Health expenditure, total (% of GDP)*, 2015), close to 10% of the gross domestic product in the world. Worldwide, most of these healthcare operations are funded by governments. With a slowing world economy, rapidly rising costs and a better-informed customer base, governments, as well as healthcare organisations, are under pressure to deliver a product that focuses on quality care, cost transparency and an excellent patient experience (Morris, 2016). To accomplish this, the healthcare industry is moving away from its traditional fee-for-service model, driven by volume with not much cost consideration, to a more value-based payment model driven by cost and quality care. Cost is now taking a central role in the healthcare industry. Although there is usually a negative connotation to cost, cost often increases due to the introduction of new medical treatments, which has a positive impact on the health and/or recovery of patients. Hence, in the healthcare industry, there is a fine balance between the management of costs while simultaneously provide quality care.

The effects of these two factors, an ageing population and the demand for quality care, are such that the operating and decision-making activities of healthcare organisations are constantly being challenged. Decisions on how to adapt procedures and medical interventions in order to deliver higher quality care at lower costs are of critical importance. This is, of course, not dissimilar to the pressures experienced in many other business and government fields.

One of the consequences of the above developments is the pressure put on the discipline of informatics. For the purpose of this study, specifically the pressures put on the area of biomedical informatics will be investigated, which occurs in two ways:

- Firstly, there is the challenge to reconceptualise and reconfigure the decision support systems that underlie all decision-making processes of organisations operating in a complex environment, such as typical healthcare organisations. This study will refer to organisations operating in a complex environment as complex organisations.
- Secondly, there is the challenge of interpreting and presenting the information generated by a reconfigured and more sophisticated decision support system in ways that convey decision-supporting meaning to non-technical, participants in the organisational flow of activities.

The remainder of this chapter provides an introductory overview of the thesis. Section 1.1 introduces the research problem and states the research question; Section 1.2 states the research aim and objectives and Section 1.3 discusses the research approach. Lastly, Section 1.4 presents the outline of the remainder of the dissertation.

1.1. Research problem

Biomedical informatics is defined as:

> *"... the scientific field that deals with medical information, data and knowledge-their storage, retrieval and optimal use for problem-solving and decision-making"* (Shortliffe & Blois, 2006) .

It is a multidisciplinary field which addresses a number of fundamental and application-related research problems (Greenes & Shortliffe, 1990). As such, biomedical informatics explores the adoption, development, and application of theory, methods and tools to support the following (VUMC, 2002):

- coding, storage, retrieval, and transmission of data,
- knowledge discovery, and
- decision making for treatments and preventive interventions, as well as for health management.

Since healthcare, as an organisational activity, lies at the intersection of almost all dimensions of human society and existence, and also draws on virtually all scientific fields, the nature of the data generated in the process is highly multi-dimensional and complex. This alone poses major problems for data management. It becomes even more complex when such data are to be utilised for knowledge discovery and decision making.

An important discipline that supports organisational decision making is knowledge discovery from data (KDD). KDD focuses on the extraction of knowledge from data repositories, i.e. databases and data warehouses, enabling these repositories to support organisations' decision-making processes. KDD can be defined as

> *"the non-trivial extraction of implicit, previously unknown and potentially useful information from data"* (Adriaans & Zantinge, 1996).

However, this definition only focuses on the features of the resultant information. Since it does not address the complexity of real-world instances of extraction, organisation, and presentation of the discovered information, there is an additional need to define KDD as a process. Brachman (1996) did exactly this, as:

> *"(The KDD process is) a knowledge-intensive task consisting of complex interactions, protracted over time, between a human and a (large) database, supported by a heterogeneous suite of tools".*

This highlights the iterative nature of the KDD process, the integral role of decision making as part of the process, as well as the utilisation of knowledge discovery tools. Following the work of Bendoly (2003), for the purpose of this study, Brachman is augmented with Mintzberg's classical three-phase decision-making process model (Mintzberg, 1973), to serve as the basis for framing a KDD process, illustrated in Table 1.

Table 1. Brachman's KDD process elements in the context of Mintzberg's three phases (Mintzberg, 1973)

Brachman's KDD process elements	Mintzberg's three phases	KDD Process
Task discovery, data discovery, data cleaning	Identification phase	Domain identification
Model development, model specification (data analysis), model fitting (data analysis)	Development phase	Strategy development/application
Model evaluation (data analysis), model refinement (data analysis), output generation	Evaluation (selection) phase	Results evaluation

In summary, the KDD process can be discussed in terms of the interaction of three phases:

- domain identification,
- strategy development/application, and
- results evaluation.

Intelligent data analysis occurs within the strategy development/application phase. Hand (Berthold & Hand, 1999) describes intelligent data analysis as the

> *"… repeated application of techniques, as one attempts to tease out the structure, to understand what is going on. To refine the questions that the researchers are seeking to answer requires painstaking care and, above all, intelligence. It is a carefully planned and considered process of deciding what will be most useful and revealing"* (Berthold & Hand, 1999).

This is where data mining tools and techniques are applied to discover useful information.

Recent literature surveys indicated that the KDD process has been successfully applied to different fields of human endeavour (including marketing, banking, customer relationship management, engineering and various areas of natural science) for the purpose of knowledge discovery (Bellazzi & Zupan, 2008). It is for this reason that similar KDD process applications have been used to analyse medical data; however, in the medical field despite initial high hopes, the KDD process has only delivered limited positive results.

Where such KDD processes and applications have been performed, they have only discovered simple relationships in data and have not yet

demonstrated the ability to discover intricate, hidden relationships in complex medical data as they have in data from other, less complex organisations. (Patel, *et al*., 2009). Bellazzi *et al* in his paper *Predictive data mining in clinical medicine: Current issues and guidelines* (2008) argues that current medical data mining, and more specifically clinical data mining, only deals with 'bed-side' problems. These data mining techniques are models that only forecast patient outcomes given a set of diagnostic data. However, the goal of predictive mining tasks in the clinical field is to derive models that use patient-specific information to predict outcomes of clinical interest, and thereby can support medical staff in their clinical decision making.

Compared to data mining in the fields of marketing, banking, customer relationship management and engineering, clinical medicine and the resultant data have several distinguishing features making data analysis in this area much more difficult and complex.

Firstly, one of these distinguishing features is that medicine operates in a safety critical context (Cios & Moore, 2002). In clinical medicine, every decision taken by medical staff needs to be clearly motivated, hence the importance of the interpretability of data analysis. Methods that offer explanations and models which are interpretable and allow domain experts to inspect the inner workings of the model will always be preferred in these safety critical applications.

Secondly, clinical data originates from multiple, unrelated, heterogeneous sources. The integration of the data from these diverse sources is of utmost importance and hence the need for standards for data presentation and coding, such as the Systemised Nomenclature of Human and Veterinary Medicine (SNOMED). However, although these standards help with ensuring the uniformity of data sets, they have yet to become integral to the data mining and analysis of clinical data.

Thirdly, lack of integration of genetic and clinical data to support therapeutic decision making. This is experienced in genomic medicine, which

deals with the analyses of gene expression data to diagnose diseases and obtain a prognosis. Most of these genomic studies publish predictive gene lists (vant' Veer *et al*., 2002), (Pomeroy *et al*., 2002). These lists vary from study to study, due to data selection bias and a lack of robustness in the analysis (Bellazzi & Zupan, 2008). To improve the accuracy, overall robustness, and clinical relevance of genomic medicine, there is a need for the integration of gene and clinical data, as discussed by Nevins (Nevins *et al*., 2003) and Futshick (Futschik, Sullivan, & Kasabov, 2003).

Lastly, there is a need for an interactive, explorative interface available to domain experts, who are not computer scientists. This interface should allow users, with seamless support, to interactively discover and formulate new hypotheses.

During this study, the KDD process was adapted for the successful application within the field of biomedical informatics, by taking into consideration the

- complexity of the environment in which an organisation operates, and
- shortcomings of existing KDD processes, that fail to address the unique requirements of a clinical environment.

1.1.1. Research Question

Based on the research problem outlined above, the primary research question for this study was formulated as:

> *Will a model, dedicated to the field of biomedical informatics and incorporating the complexity of healthcare organisations, address the lacum?*

1.2. Research aim and objectives

Against the background described above, the point of departure for this study is the present reality of the lack of adequate knowledge discovery process models specifically in biomedical informatics.

- This research project hypothesises that this failure is due to an inadequate recognition of the complexity involved in the environment of the healthcare sector with the result being that the models are incapable of including all of the relevant factors.
- As a transfer of existing KDD process models from other fields cannot directly solve the problem, the construction of a plausible model for complex environments is necessary, incorporating insights from recent advances in the fields of data modelling, machine learning and computational thinking.
- It is further hypothesised that a model, which conforms to more sophisticated KDD processing, to deal with the complexities, will address the gap in the healthcare sector.

The research objective was to develop and motivate a KDD process model dedicated to the field of biomedical informatics (BI), as well as demonstrate its practical application within the BI context. The constructed model comprises methods based on particular theories, as introduced in Chapter 2. In Chapter 3, a methodology was then developed, by which the model was deployed. Lastly, the model was applied to a case study, as discussed in Chapter 4, to demonstrate that the methodology functions within the BI context.

1.3. Research approach

This study was conducted in a data-rich, scientific environment. The underlying epistemology supporting this research was that of data-intensive science, and took on an exploratory approach to data-centric research. Within this approach, information emerges from data, compared to the more traditional confirmatory approach, during which hypothesised patterns, expected from the data, are exposed and examined.

For the purpose of this study, the scientific method was adapted in such a way to include methods and technologies that support data-centric research. This included the advances in data modelling, specifically modelling data as information networks. In addition, a recent development in the field of computer science, namely computational thinking, was incorporated. Lastly, the guidelines for the practice of reproducible research for data-intensive science were incorporated, specifically to make this study computationally reproducible.

The developed methodology was then applied to a case study, which is the most common research method used in information systems research. Yin (2009) defines case study research as follows:

> *"A case study is an empirical enquiry that:*
>
> – *Investigates a contemporary phenomenon in real-life context, especially when*
>
> – *The boundaries between phenomenon and context are not clearly evident"*

The phenomenon investigated by this study was: The development of a KDD process model, dedicated to the field of biomedical informatics, which incorporates the complexity of the sector, and addresses the gap in the field of KDD.

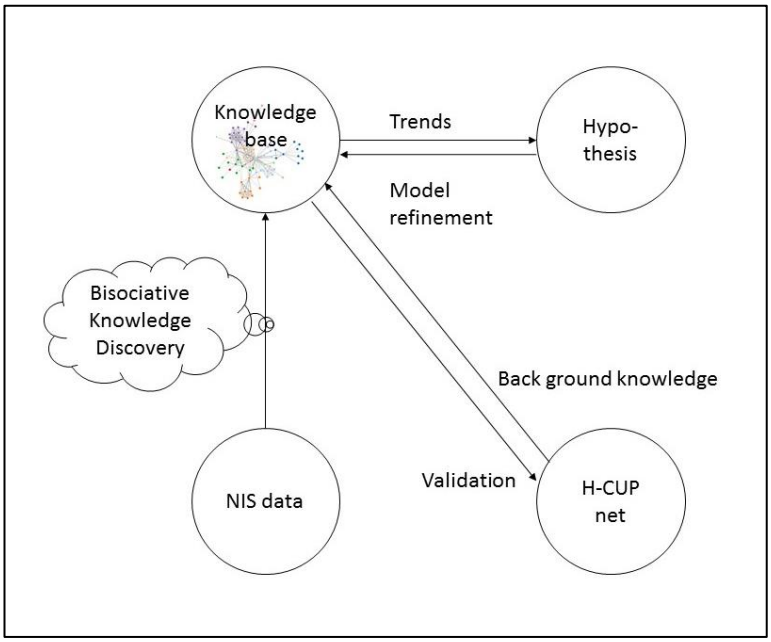

Figure 1: Research approach

The real-world domain analysed for this study was a Healthcare Cost and Utilisation System. This healthcare system was defined by sets of data contained in the National Inpatient Sample (NIS), of the Agency for Healthcare Research and Quality of the United States of America. Domains were identified and the data were modelled as an integrated heterogeneous information network. Knowledge was extracted from the information network by means of bisociative knowledge discovery. The knowledge was represented by means of concept graphs with their domain-bridging concepts and/or subgraphs.

H-CUPnet is an online query system which provides access to the health statistics on hospital inpatient data. For the purpose of this study the statistics in the H-CUPnet system were used to validate knowledge as presented by the concept graphs. The H-CUPnet system also supplied background knowledge of the NIS data, offering a better understanding of the data in the NIS data sets.

1.4. Chapter layout

This dissertation consists of two sections. The first part presents the theoretical basis for the development and application of a model dedicated to the field of biomedical informatics. The second part presents the experimental investigation into the applicability of the model developed in the first part, as applied to a real-world case study.

Chapter 2 lays the theoretical foundation for this study. In Chapter 3, the Data Mining and Knowledge Discovery process model for bisociative knowledge discovery, in the healthcare sector, is developed and defined. A case study was used for the experimental work and this is documented in Chapter 4. Chapter 5 summarises the main results of the dissertation and discusses areas of future work.

Chapter 2

Literature review

The theoretical basis of the research conducted in this study draws from recent advancements, methods and corresponding theories within the following scientific fields:

- Biomedical informatics
- Knowledge Discovery from Data (KDD) focusing on bisociative knowledge discovery
- Information networks as an abstract representation of the real-world focusing on heterogeneous information networks
- Pattern recognition and machine learning algorithms as the methods by which complex patterns are discovered in large data repositories

This chapter is organised as follows: Section 2.1 introduces the field of biomedical informatics and discusses the uniqueness of the healthcare sector. This is followed by Section 2.2 which introduces Koestler's concept of bisociation and discusses bisociative knowledge discovery. Section 2.3 describes the way in which information networks model real-world situations. Finally, Section 2.4 introduces the methods used for identifying patterns in these network models.

2.1 Biomedical informatics

Morris Collen (1986) documents the origin of the term 'medical informatics', citing from a written communication of J Anderson, at Kings College of Medicine (London), as follows:

> "*As you will see from the book on 'Education in Informatics of Health Personnel', we had been searching for some time before 1974 in the IFIP [International Federation for Information Processing] Technical Committee No. 4 to find a suitable term for the subject area. Professors Pages and Gremy of Paris were interested in at least two aspects being represented in the final term, namely, the French terms 'informatique' and 'automatique' that were used in France for medical information science or data processing. It was certain we had to find a new term for the book, and after much discussion we incorporated the words to form the name 'medical informatics'. We intended it to cover both the information and data parts as well as the controlling and automatic nature of data processing itself (*J. Anderson, MD, written communication, May 1986*)."*

As the Human Genome Project commenced during the 1990's, data analysis expanded into the field of basic biology. At this time, the term biomedical informatics started to circulate in literature. The same methods and processes that were identified within the field of medical informatics were now being applied to the broader field of biomedicine. For this reason, the term medical informatics was replaced by the term biomedical informatics (Kulikowski *et al*., 2012).

In his commentary *Medical informatics: Past, present, future* Reinhold Haux gives a modest definition of the field as

> *"… the discipline, dedicated to the systematic processing of data, information and knowledge in medicine and healthcare"* (Haux, 2010).

Furthermore, Shortliffe *et al* defines biomedical informatics as the

> *"scientific field that deals with biomedical information, data and knowledge — their storage, retrieval and optimal use for problem solving and decision making"* (Shortliffe & Blois, 2006).

Accordingly, biomedical informatics is a scientific field that focusses on the use of biomedical data, information and knowledge and the systematic processing thereof with the purpose to solve problems and support decision making within the biomedical environment.

One finds that the meaning of the terms: medical informatics, health informatics and biomedical informatics varies within and between different nations. In Europe the discipline is most often referred to as medical informatics, whereas in the United States the core discipline is referred to as biomedical informatics with different areas of application, of which medical informatics and health informatics are two (Haux, 2006), (Shortliffe & Blois, 2006).

Taking all the above into consideration, the American Medical Informatics Association (AMIA) identified the need to formally define the term biomedical informatics and in 2012 published a white paper (Kulikowski *et al*., 2012) in which the term was defined as:

> *"Biomedical informatics is the interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem*

solving, and decision making, driven by efforts to improve human health."

For the purpose of this study, the AMIA definition of biomedical informatics, as stated above, will be used.

2.1.1 The science of biomedical informatics

Bernstam *et al* (2010), in his article *What is biomedical informatics?* argues that to establish biomedical informatics as a scientific field, its definition must be grounded in theory. Considering the definition of biomedical informatics, as defined by the AMIA, one finds that data, information and knowledge are central to this field. Since these terms are often used interchangeably and literature lacks a consistent definition thereof, the first step in establishing biomedical informatics as a scientific field is to define the terms: data, information and knowledge.

Transfer of know-how occurs in a direct or indirect manner. The direct transfer of know-how is a face–to-face, person-to-person interaction. Sveiby interprets this direct transfer by using Polany's concept of tradition (Sveiby, 1996).

Polany (Sveiby, 1996) describes knowledge as a *process* i.e. a process-of–knowing. He defines this process as a personal one constructed by an individual within a specific social context. Polany then describes how this process-of-knowing can be used as a tool by a person for action or as a tool to acquire further knowledge. This process-of-knowing, according to Polany, is usually transferred via tradition which is defined as a system of values outside the individual. Hence, knowledge is personal and transferred in a direct manner from person to person via tradition. It is a slow process whereby the knowledge is mostly unarticulated and therefore difficult to distribute.

In addition, Polany describes knowledge also as an *object*, as something that can be articulated into information and then this information

becomes the source through which the knowledge is communicated. The challenge when articulating knowledge is to make the tacit knowledge explicit and distancing the individual from it.

Indirect transfer of knowledge occurs when this knowledge object i.e. information, is transferred via a specific medium, and not transferred by means of personal interactions. Sveiby interprets this indirect transfer of knowledge by using Information Theory (Sveiby, 1996).

Information is the source through which knowledge is communicated. During indirect transfer of knowledge, the information is articulated in, for example, the written word and in today's world, most often represented in an electronic format. Information theory describes this information as data with meaning. This form of knowledge transfer is a fast process whereby the information is easily duplicated and therefore has the ability to be mass distributed.

As this study concerns automating the process of meaning by making use of computation, it acknowledges the sociological interpretation of knowledge as an important field of study but concerns itself in this work mainly with the indirect transfer of knowledge, hence focussing on the information science and interpretation thereof.

In this realm, a number of studies have been conducted regarding the definition of data, information and knowledge including: Ackoff (1989), Rowley (2007), Dunn (2008) and Floridi (2008). In literature, it has been revealed that two main schools of thought exist on data: information and knowledge. The one school, the often cited Ackoff's knowledge hierarchy (Ackoff, 1989), versus the more recent definitions arising from the new school of thought, Philosophy of Information (Adriaans & van Benthem, 2008).

The first appearance of a data, information, knowledge hierarchy, argued by many, was in T.S. Eliot's poem "The Rock" (Elliot, 1934).

"Where is the life we have lost in living?

Where is the wisdom we have lost in knowledge?

Where is the knowledge we have lost in information?"

More recently, Ackoff (1989) described data, information, knowledge and wisdom in terms of the following hierarchy:

"Wisdom is located at the top of a hierarchy of types […] Descending from wisdom there are understanding, knowledge, information and, at the bottom, data. Each of these includes the categories that fall below it - for example there can be no wisdom without understanding and no understanding without knowledge."

This is often referred to in literature as the "knowledge hierarchy" and depicted in the following way:

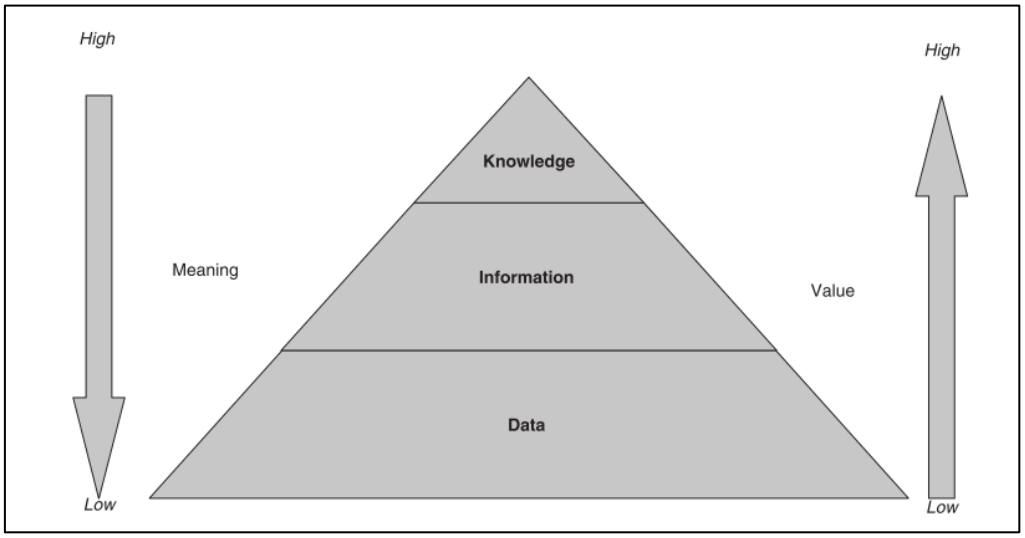

Figure 2: Data, information and knowledge

Within this knowledge hierarchy, Ackoff defines data as symbols that represent properties of objects, events and their environment. That is to say, they are products of observation. When taking into consideration Russell & Norvig's (1995) definition of an agent as:

> *"... anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors."*

It can be said that, data are symbols by which an agent perceives its environment. According to Dunn (2008) these symbols, on their own, have no context or meaning.

Information, on the other hand, is data that have been processed to be meaningful. Information is contained in descriptions as well as answers to questions that begin with words such as who, what, when and how (Ackoff, 1989). In other words, for data to be transformed into information, it has to be recorded in an interpreted way. In addition this information, i.e. data with meaning, allows an agent to act upon its environment. In this way, the data are made meaningful, valuable and relevant.

Within the context of Ackoff's knowledge hierarchy, knowledge is defined as know–how i.e. the application of data and/or information. Once an agent processes data and information and applies to it understanding, experience, skills, trust, values and beliefs, information is transformed into know-how i.e. knowledge, hence the hierarchy.

In the same way, the philosophic definitions of the terms: data, information and knowledge form a similar hierarchy. These definitions have recently been adopted by a new discipline, namely, the Philosophy of Information, and were published in the GDI (General Definition of Information) (Floridi, 2014). GDI defines data (singular: datum), information (semantic content) and knowledge as follows:

- A datum is a putative fact regarding some difference or lack of uniformity within some context.
- σ is an instance of information, understood as semantic content, if and only if:
  - σ consists of one or more datum;
  - the datum in σ are well-formed;
  - the well–formed datum in σ are meaningful.
    - Thus, information is well-formed, meaningful data.
- Knowledge is information that is true, justified and believed.

Similar to Ackoff's knowledge hierarchy these definitions also produce a natural hierarchy. A significant amount of data that is being produced lacks interpretation; as a result, it has no meaning. Therefore, there will always be more data than information.

Likewise, a significant amount of information that is being produced holds no truth, or lacks justification for why it is true, and as a result it is not knowledge. Therefore, there will always be more information than knowledge.

To continue, if information is meaningful data, depending on the environment in which one is working, it is sometimes more convenient to refer to data as the *syntactic part* of information, and meaning as the *semantic part*. That is to say, for data to be meaningful, it must be combined with other data and arranged in a systematic way, namely a syntax. Consequently if information is meaningful data, information can be interpreted as a syntax plus semantics.

Likewise, within a representational system, the formal methods i.e. the systematic rules used by computer programs, are responsible for the manipulation of the data within the system. Because of the data's relationship to these formal methods, it can be more convenient to refer to data as the *form* of the system and to the meaning as its *content*. The systematic rules that manipulate the data are dependent on the form (data) only, without any regard to its meaning. For this reason, it remains essentially a human task assuring that the input and output of these formal methods correctly preserve the meaning of the environment it represents (Bernstam *et al.*, 2010). Consequently, if information is meaningful data, information can also be interpreted as form plus content.

In conclusion, information, depending on the environment, can be defined in terms of

- data plus meaning,
- syntax plus semantics and/or
- form plus content.

To summarize, in Section 1.1 it has been stated that informatics is the science of information. In Section 2.1 information is defined as data with meaning. From this, one can conclude that informatics is the science of information where the object of study of this scientific field is data with meaning.

The object of study for this scientific field distinguishes informatics as a science from other related fields such as computer science, mathematics

and statistics. The object of study of computer science is the study of computation which is an important tool for informatics but it is neither necessary nor a sufficient condition for informatics. Similarly, mathematics and statistics relate to the formal study of abstract patterns and features of data, but not meaning. These related fields develop tools that are predominantly designed for data manipulation. The challenge within the field of informatics is to automate the processing of meaning with these tools that were predominantly designed for the purpose of manipulating data.

The level of complexity of this automated processing of meaning varies considerably within different fields of application. For example, within the banking industry transforming data (typically numbers) into information (typically account balances) is a mere manipulation of the display of numbers. For this specific application, the level of complexity of the automation process is relatively low, due to the narrow semantic gap that exists between the data and the information. If the problem is a form–based problem, or can easily be reduced to a form-based problem, as in the banking example, the automated processing of meaning can more easily be solved using data manipulation tools (Rowley, 2007).

Bernstam *et al* (2010) describes the science of biomedical informatics as:

> *"… the application of the science of information as data plus meaning to problems of biomedical interest."*

Within the field of biomedical informatics, biomedical concepts such as recognising a sick patient or defining a headache are hard to reduce to data and therefore are complicated to represent computationally. As a result, these problems are difficult to reduce to form-based problems. This fact contributes to the complexity of the automated processing of meaning within this field. Consequently, biomedical informatics employs methods and tools derived from a variety of fields including: information science, computer science,

cognitive science, business management and organisations management, statistics and biometrics, mathematics, artificial intelligence, operations research, economics, and basic and clinical health sciences. The enormous challenge is thus to develop tools to automate the processing of meaning specifically within the field of biomedical informatics.

2.1.2 Clinical informatics, a discipline within biomedical informatics

In Section 2.1, biomedical informatics (BI) is defined as an interdisciplinary field that studies the uses of biomedical data, information, and knowledge for decision making, driven by the desire to improve human health.

The AMIA Board white paper (2012) on the definition of biomedical informatics as a core scientific discipline illustrates this interconnectivity between BI and its major areas of applied research, as shown in Figure 3.

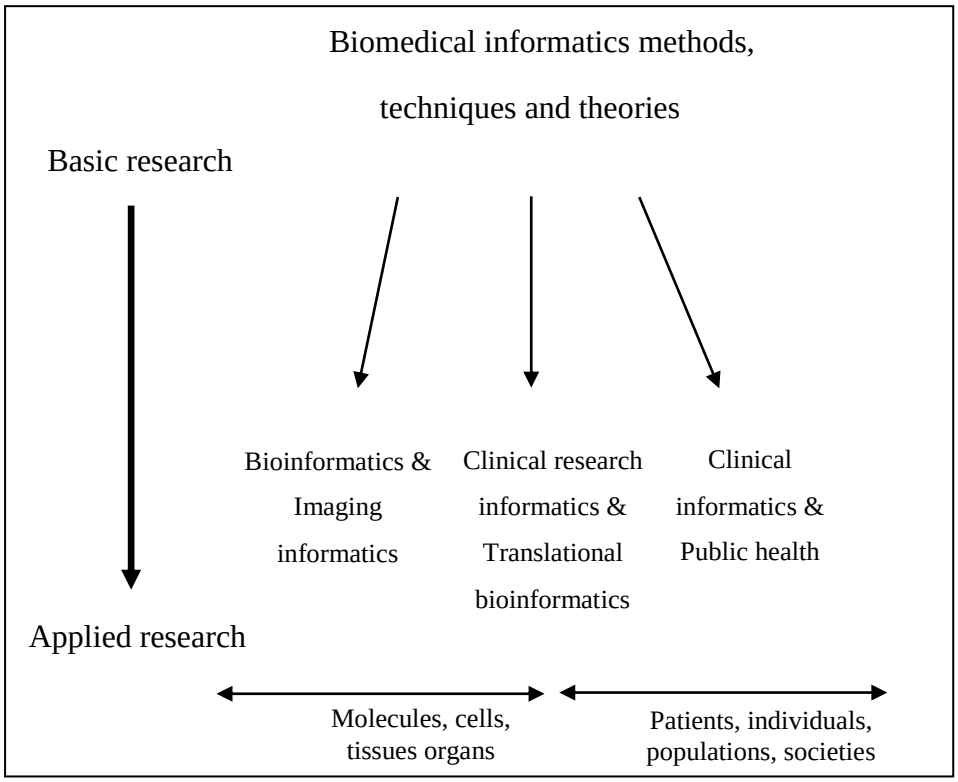

Figure 3: Disciplines within biomedical informatics

Biomedical informatics is the scientific discipline within which methods, techniques and theories are studied and developed. These methods are shared across the biomedical disciplines (application areas) including:

- Bioinformatics and imaging informatics,
- Clinical research informatics in translational bioinformatics, and
- Clinical informatics and public health informatics

from molecular to population levels and they define the core discipline of BI.

Research is driven by fundamental methodological issues that are usually identified within these application areas. These issues provide the research motivations with which the application areas interact with the core discipline. However, the core discipline stays identical regardless of the area of application that a specific research motivation is addressing.

Reed (in Gardner *et al*. 2009) defines clinical informatics as follows:

> "… *the subfield that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving, and decision making, driven by efforts to enhance individual and population health outcomes, improve patient care, and strengthens the clinician-patient relationship*".

AMIA defines clinical informatics as "… *the application of informatics and information technology to deliver healthcare services*" (AMIA, n.d.). Hence one can conclude that clinical informatics is an application area of BI that focusses on the improvement of healthcare services.

2.1.3 The uniqueness of the healthcare sector

Recall from Section 1.1 that from an informatics point of view, the healthcare sector constitutes perhaps the most challenging environment in which to construct successful information systems.

Healthcare organisations as well as their systems differ from other organisations in the following ways (Lorenzi & Riley, 2004):

i. Decision making is a systemic ingredient of the processes within a healthcare organisation. Even more so, of processes that depend on systems that support decision making regarding life and death situations. Mistakes made by these decision support systems can lead to serious harm. Therefore the *tolerance level for errors* differs significantly from, for example, financial systems which can far easier tolerate errors.

ii. These organisations have *complex as well as unique personnel structures*. Top management, unlike their counterparts in other industries, must sell their strategies to largely independent physicians. These physicians have the final authority and carry the responsibility on how to treat their patients. They are not employees of the hospital and in many cases earn far more than the institution's chief executive.

iii. The public's *image of healthcare organisations is declining* due to, among others factors, perfect outcomes for procedures are expected regardless of the risk, and the increase in the number of malpractice lawsuits within the industry.

iv. There is an *exponential rise in the cost* of healthcare compared to other industries mainly due to the explosion of

new technologies within the field, an aging population, rising patient expectations and legal liability issues.

v. In contrast with industry counterparts of its size, healthcare organisations operate in an oligopolistic economic environment.

vi. Lastly, the healthcare industry has a unique payment structure where the client has no idea what the final cost of procurement will be.

Healthcare organisations are complex organisations made up of a wide variety of subsystems. Many of these subsystems are similar to those in other organisations. However, embedded within are the peculiarities described above. As a result healthcare organisations challenge the development of health information systems.

Healthcare applications are technically complex with less mature software and hardware markets than other industries more often studied in information system research, such as manufacturing, airlines and financial institutions and services (Chiasson & Davidson, 2004). Unlike top-down hierarchical control structures found in many industries, the healthcare field has a dual administrative structure of medical personnel and administration, as mentioned before. The research of Kaplan (1994) and Anderson (1997) illustrated how these institutional structures increase the complexity of developing and implementing information systems among medical professionals.

Furthermore, healthcare is a complex mixture of for–profit and not-for–profit motives, as well as government, private not-for-profit, and private for-profit enterprises, with national differences due to regulatory and market structures. A trend towards privatisation is evident in hospitals and healthcare providers, hence the shift of focus to efficiency of service delivery and away from equity of and access to services as in the past (Scott, Ruef, Mendel, &

Caronna, 2000). All of these facts contribute to the uniqueness of the healthcare sector.

2.2 Knowledge discovery from data

Recall from Section 2.1.1 that data are mere symbols that represent properties of objects, events and their environment. They are products of observation. The technical advances in computational power, storage capacity, and inter-connectivity of computers have led to uncomprehendible quantities of digital data being generated on a daily basis, often referred to as the *data deluge*. This data, when recorded in a meaningful, interpreted way are transformed into information i.e. data with meaning. Once proven to be true, justified and believed, this information within a specific domain can be considered knowledge (Adriaans & van Benthem, 2008).

The scientific field of Knowledge Discovery from Data (KDD) focuses on the extraction of knowledge from data repositories i.e. knowledge bases, enabling these repositories to support decision-making processes. KDD can be defined as

> *"the non-trivial extraction of implicit, previously unknown and potentially useful information from data*" (Adriaans & Zantinge, 1996).

Modern KDD methods allow users to discover complex patterns of various types in large data repositories. Methodology frameworks, such as CRISP-DM, are adopted by industry for knowledge discovery projects. This allows the user to participate in a cycle of data preparation, data mining and knowledge evaluation. The underlying assumption being that the data to which these methods are applied originates from a single domain, hence finding associations in one domain. By domain, it is emphasized that the data under analysis represent the properties of objects, events and their environment pertaining to only one aspect.

In contrast to discovering patterns within a single domain, the term bisociations refers to finding relations across domains. Bisociation is a model of creativity proposed by Arthur Koestler in the 1960's in his book *The Act of Creation* (Koestler, 1964).

2.2.1 Koestler's concept of bisociation

"Creativity is the defeat of habit by originality"

— Arthur Koestler

Koestler refers to matrices and codes to explain the concept of bisociation. In Koestler's terminology a matrix refers to any habit, ability, skill, or any pattern of ordered behaviour that is governed by a code. A code is a set of fixed rules that govern the matrix. These rules can be expressed in terms of a simple mathematical equation which contains the essence of the pattern (Koestler, 1964). For example, when considering a chess game, the code will be the rule that governs the movement of the specific chess piece, whereas the matrix will be all the permissible moves of that specific chess piece when applying the rules.

Koestler argues that we construct matrices which shape our perceptions, thoughts and activities. Once we reach adulthood we have formed rigid, automated patterns of behaviour and thinking. Each governed by a code which consists of a fixed set of rules that we have either acquired or are innate. To escape from these atomised routines of thinking Koestler proposes that one needs to connect previous unconnected matrices of perception or experience, in a creative act which he defines as the act of bisociation. This leads to seeing familiar objects and events in a novel and revealing light.

Dubitzky *et al* (2012) illustrates this concept of bisociation as follows:

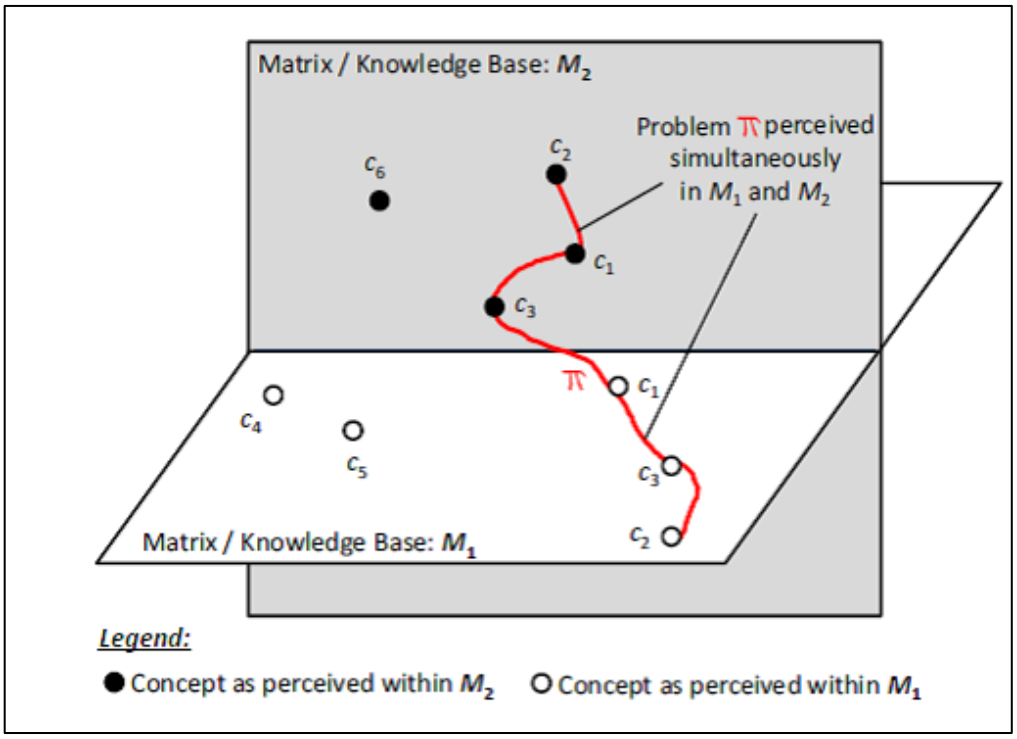

Figure 4: Illustration of Koestler's concept of bisociation (Dubitzky et al., 2012)

The diagram depicts two matrices of thought as two perpendicular planes M_1 and M_2 with six concepts c_1, $c_{2, \dots,}$ c_6. The concepts c_1, c_2, c_3, and c_6 are perceivable in M_2 and c_1, c_2, c_3, c_4, and c_5 are perceivable in M_1. The symbol π denotes a problem. The concepts c_1, c_2 and c_3 are associated with π and linked to one associative context, namely M_1 or M_2. Because c_1, c_2 and c_3 are perceivable in both matrixes it is possible to observe π from two frames of mind simultaneously. This implies that π is not only associated to M_1 or M_2 but also bisociated to M_1 and M_2.

2.2.2 Bisociative knowledge discovery

Transferring bisociation to a knowledge discovery scenario emphasises the creative discovery of previously unknown relations across domains. Where a domain is a collection of properties of a set of objects, events and their environment pertaining to only one aspect. Each domain is represented by its

own knowledge base. These knowledge bases, which are habitually incompatible, are connected during bisociative knowledge discovery in such a way that creative insights are generated. In order to discover these bisociations, the knowledge bases need to be explored in a creative way.

The term exploratory data analysis (EDA) was introduced for the first time by John Tukey (1977) in his book *Exploratory Data Analysis*. Tukey stated that EDA is mostly an unstructured, interactive search exercise through data with the goal to identify novel, potentially useful information without an initial hypothesis. Based on Tukey's work on EDA, Gossen defined the term bisociative data exploration as follows (Gossen, Nitsche, Haun, & Andreas, 2012):

> *"Exploration is bisociative, if the data set consists of two or more habitually incompatible domains (knowledge bases) and the user is presented unusual, but interesting domain-crossing connections with the aim of finding relevant and new relationships between those domains."*

These knowledge bases can be described in terms of domain theory in the following way:

A domain theory consists of all concepts relevant to a given domain, at a given point in time. These concepts (set of rules) govern the domain and form the basic units from which a domain theory is constructed (similar to Koestler's code that governs a matrix) thus:

> *A domain theory D_i defines a set of concepts that are associated with a particular domain i.*

A knowledge base is a subset of concepts from an underlying domain theory (Koestler's matrix concept is reflected in the definition of a knowledge base) therefore:

> *A knowledge base K_i is defined as a subset of a domain theory D_i, i.e., $K_i \subseteq D_i$.*

Recall from Section 2.1.1 that an agent is "*anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors*". A knowledge base is agent specific. Depending on whether the agent knows concepts in that domain or not, the agent-specific knowledge base may be empty or non-empty. The concepts belonging to the non-empty knowledge base are all selected from a single domain theory. This selection takes place in a very biased or habitual way based on how the agent perceives its environment. As a result, the knowledge base imposes a unique, biased perspective of the agent on that domain.

An agent has exactly one knowledge base per domain. These agent-specific knowledge bases are incompatible, since given a concrete problem only one agent–specific knowledge base will be active at any given point in time. Koestler refers to this phenomenon as habitually incompatible knowledge bases. Agents are equipped with the ability to detect patterns of bisociation by bringing together multiple knowledge bases simultaneously. Hence, in terms of domain theory, bisociation can be defined as follows (Dubitzky *et al*., 2012):

Let U denote the universe of discourse, which consists of all concepts.

Let $c \in U$ denote a concept in U.

Within the universe of discourse, a problem π is associated with the concepts $X \subset U$. Typically, in a concrete setting, a subset $P \subset X$ is used to describe and reason about π.
D_i denotes a domain theory which represents the total knowledge within a domain, hence

$$\cup D_i = U \text{ furthermore, } \exists i, j: D_i \cap D_j \neq \emptyset.$$

R denotes an intelligent agent which possesses exactly one knowledge base (empty or non-empty) per domain theory D_i.

$K_i^R \subset D_i$ denotes the knowledge base with respect to the intelligent agent R and domain theory D_i. Notice, an intelligent agent R has exactly one single knowledge base K_i^R per D_i.

$K^R = \cup_i K_i^R$ denotes the entire set of knowledge bases incorporated in the intelligent agent R. K^R represents the total knowledge that R has in all the domains.

This means that:

> *Two agent-specific knowledge bases K_i^R and K_j^R $(i \neq j)$ are said to be habitually incompatible if, at a given point in time t, there is no concept $c: c \in K_i^R \wedge c \in K_j^R$ that is active or perceived simultaneously in K_i^R and K_j^R.*

Therefore, an agent usually employs a single frame of mind (knowledge base) at a time to think about a problem.

Let π denote a concrete problem and let $X \subset U$ denote the concepts

associated with π. Also, let K_i^R and K_j^R denote two habitually incompatible agent-specific knowledge bases *(i ≠ j).*

> *Association occurs when elements of X are active or perceived in K_i^R at time t only.*

For example, at time t the concepts *A = {c1, c2, c3}* may be active in K_i^R. In this case we say that the concepts in *A* are associated (with each other).

> *Bisociation occurs when elements of X are active or perceived simultaneously in both K_i^R and K_j^R at a given point in time t.*

This refers to the situation where a problem is perceived simultaneously in two frames of mind (or matrices of thought). For example, at time *t* the concepts *B = {c1, c2, c3}* may be active or perceived simultaneously in K_i^R and K_j^R. In this case, one may say that the concepts in *B* are bisociated.

Thus, bisociative knowledge discovery is the creative discovery of previously unknown information, in particular relationships that were previously overlooked in-between data sets from habitually incompatible domains. Gossen, *et al* (2012) argues that graph structures are the most promising way to represent these data sets such that one can derive relationships between habitually incompatible domains.

The following section addresses the representation of habitually incompatible domains as a specific type of graph structure, namely an information network.

2.3 Information networks

We live in an interconnected world, which implies that the data or concepts that represent real-world situations interact with each other forming complex, interconnected networks. These interconnected networks are called information networks. The internet, social networks, biological networks, transportation systems, health systems, and electrical power grids are all examples of information networks (Han, 2009).

Information networks are used to integrate large volumes of data from different domains. A useful way of modelling these network structures is using graphs. For the purpose of this study, the definitions of information networks must be formalised. The formal definition that is used in this study is that of Sun & Han as defined in *Mining Heterogeneous Information Networks: Principles and Methodologies* (Sun & Han, 2012).

In mathematical literature, a network structure is modelled as a graph which consists of a collection of vertices joined by edges. Likewise, information networks are composed of information units that represent physical objects as well as intangible objects such as ideas or events. The link between these units is represented by relations which can be semantic or solely correlational by nature. Formerly, Han ((2009),(Sun & Han, 2012)) defines an information network as follows:

> *"An information network is defined as a directed graph* $G = (V, E)$ *with an object-type mapping function* $\tau: V \rightarrow A$ *and a relation-type mapping function* $\emptyset: E \rightarrow R$, *where each information unit* $v \in V$ *belongs to one particular object type* $\tau(v)\, \epsilon\, A$, *each relation* $e \in E$ *belongs to a particular relation type* $\emptyset(e) \in R$ *, and if two relations belong to the same relation type, the two relations share the same starting object type as well as the ending object type."*

In order to differentiate between different types of information networks, distinctions are made between the properties of their information units and their relations. These properties determine the expressiveness of the network and thereby its ability to model specific types of data e.g. ontologies versus experimental data.

The properties of information units are as follows (Kötter & Berthold, 2012):

- Basic information units have a label attached to them that identifies the object or concept it represents.
- Attributed information units are basic information units with additional attributes attached to them. These attributes carry non-semantical information about the unit, such as a user-readable label.
- Typed information units are basic information units that have a label attached which allows the agent to differentiate between the semantics regarding the information unit e.g. a patient vs a doctor.
- Lastly, an information unit can be hierarchical, this is when an information unit is used to condense part of a network by representing a sub-graph composed of any number of information units and their relations.

A relation represents the basic connection between two information units and is not required to carry any labels. However additional properties of a relation can be:

- Attributed and typed which are similar to that of information units.
- Weighted relations carry a label which represents the strength of the relation.

- Directed relations carry a label which explicitly models a relation that is only valid in one direction e.g. parent to child relationship
- Lastly, relations that exist between more than two information units can carry a multi-relation property (Kötter & Berthold, 2012).

These properties, belonging to the information units and their relations, allow one to differentiate between the prominent types of information networks. These include ontologies, semantic networks, topic maps, weighted networks and bisociative information networks (BisoNets). For the purpose of this study the focus lies on ontologies and BisoNets as described next.

2.3.1 Ontologies

Ontologies have been developed within the discipline of artificial intelligence to describe the static domain knowledge of knowledge-based systems with which it can facilitate knowledge sharing and reuse. According to Fensel (Fensel, 2001) ontologies provide an explicit conceptualisation i.e. meta-information that describes the semantics of the static knowledge belonging to a specific domain. This meta–information can be modelled as an information network made up of a controlled vocabulary of well-defined terms as well as the relations that exist between them (Musen *et al*., 2012). Since these models provide a shared and common understanding of a domain, it makes the sharing of knowledge between different data sources and systems possible.

Within public health, the first ontology, namely the International Classification of Diseases (ICD), was published in the late 1880's containing 200 ways of dying (Bodenreider, 2008). The ICD, which is used by the World Health Organisation to determine morbidity and mortality statistics, is now in its tenth edition containing 13 000 such concepts (Sarkar, 2010). The

development of biomedical ontologies has been a major area of emphasis since the 1980's. Within the discipline of clinical informatics, specific clinical terminologies such as the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) have been developed to represent clinical information associated with patient care. The National Library of Medicine developed the SNOMED CT ontology.

SNOMED CT uses the term "concept", which represents a unique clinical meaning, for describing an information unit. Within the SNOMED CT structure, there are nine top level concepts and these are used to name the main branches i.e. hierarchies of the network. Each branch contains similar concept types. Figure 5 depicts an information unit labelled "Atrial Fibrillation", within the SNOMED CT ontology's "Clinical finding" hierarchy, and its relations.

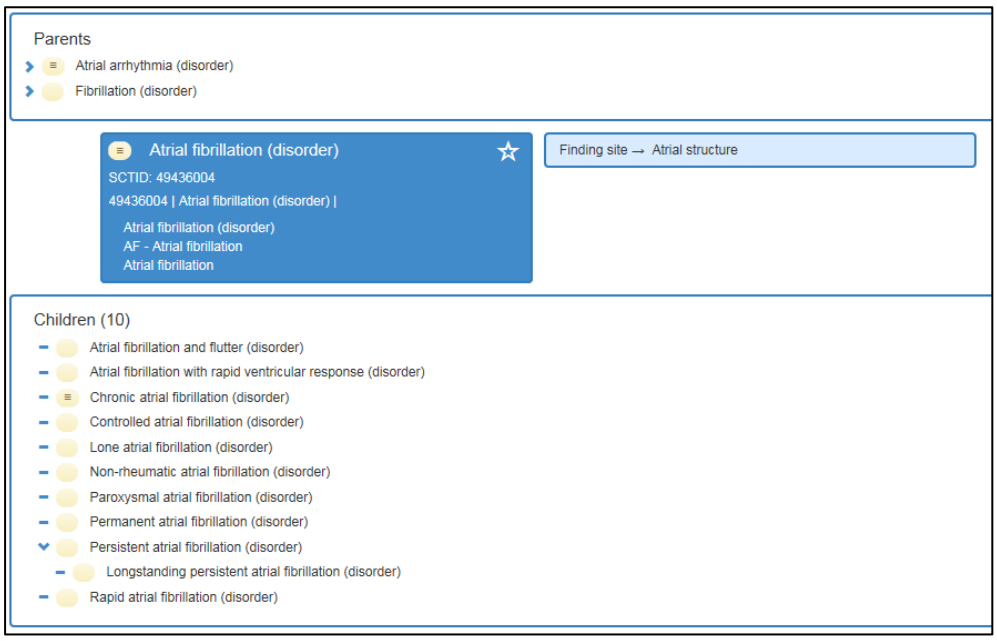

Figure 5: Atrial fibrillation as defined by SNOMED CT

As seen in the above example, a concept is uniquely identified by an identifying label i.e. 49436004. It also has a user-readable label, the fully specified name (FSN) attribute, "Atrial fibrillation" attached to it. As well as

a finding site attribute, "Atrial structure", which specifies the body site affected by the condition. Hence the concept is an attributed information unit.

The concept 49436004 participates in 12 | is a | relationships. Each relation specifies the information unit that 49436004 is connected to, as well as the direction of the relation. Parents vs children denotes the valid direction of the relation, for example "Parents" implies a broader relationship i.e. parent. "Children" a narrower relationship i.e. child. Figure 5 illustrates the two broader and ten narrower relationships. Each relation is of the same type, namely "related–to". All the parent and children concepts are of the same type namely, "disorder". This particular subset of the SNOMED CT ontology is therefore a directed homogeneous information network.

2.3.2 Bisociative information networks

Kötter *et al* (2012) in *From Information Networks to Bisociative Information Networks* introduces the bisociative information networks (BisoNets) model specifically for the modeling of data from diverse domains into an integrated information network which supports the discovery of bisociative knowledge. Kötter defines BisoNets as follows:

> A BisoNet $B = (V_1, \ldots, V_k, E, \lambda, \omega)$ is an attributed graph, where $V = U_{i \leq k} V_i$ represents the union of all vertex partitions and $k \geq 2$ denotes the number of existing partitions. Every vertex $v \in V$ represents a unit of information and can be a member of multiple partitions.
>
> The set of edges
>
> $$E = \{\{u, v\} : u \in V_i;\ v \in V_j\ ;\ j \neq i\}$$
>
> connects vertices from two different vertex partitions, whereas an edge $e = \{u, v\} \in E$ represents a connection

between the two vertices $u \in V_i$ and $v \in V_j$ where $i \neq j$ and $2 \leq i, j \leq k$.

The function $\lambda: V \rightarrow \sum^*$ assigns each vertex $v \in V$ a unique label from $\sum^*$. This allows for the identification of a vertex by its unique label.

The certainty of a relation is represented by the weight of an edge $e \in E$, which is assigned by the function $\omega: E \rightarrow [0, 1]$ and where a weight of 1 represents the highest certainty.

Accordingly, BisoNets consists of vertices which represent arbitrary units of information and their relations. BisoNets do not store any other additional data within its information units, except for its source reference. Consequently they have the ability to represent large amounts of data with the capability of source identification. Hence, adopting a *k*-partite graph structure as illustrated in Figure 6.

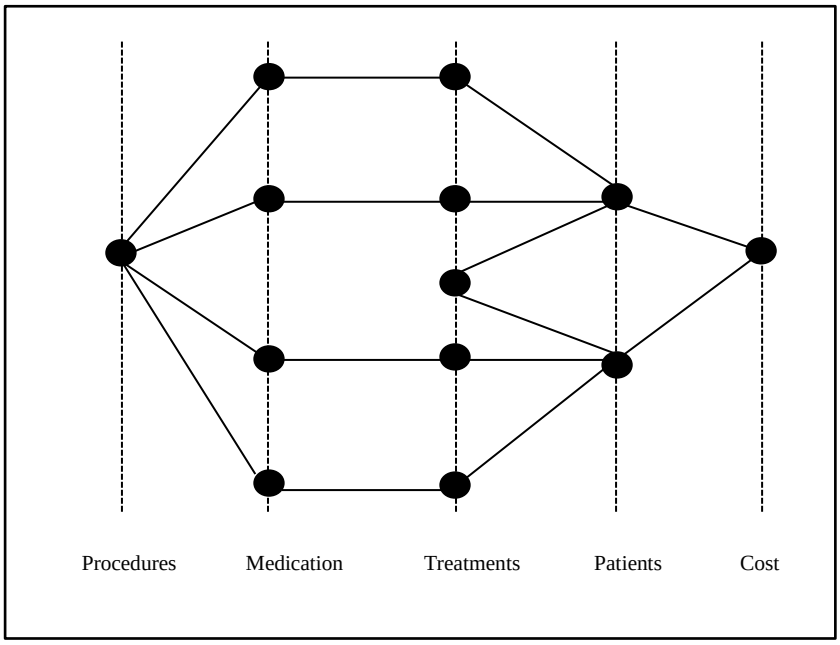

Figure 6: *K*-partite graph structure of a healthcare network

Vertices of the same type are grouped into vertex partitions, and since a vertex can play diverse roles, it can be assigned to more than one partition. The vertices of a partition act as an information unit or relation depending on the view of the agent. If one considers a healthcare network, in one view medication can describe the relationship between similar types of procedures. This view of the agent implies that procedures act as information units and medication as a relation. In a different view, when the procedures describe a basket of medication the procedures act as relations and the medication as information units. Hence the role of the vertex partition depends on the agent's view of the data.

Vertices of different partitions are connected by edges which lead to a *k*–partite graph structure. As a result BisoNets must consist of at least two partitions, one representing the information units and another describing the relationships between these units. The relations are weighted with the weight representing the certainty of the connection. BisoNets are integrated information networks that support attributed, typed and hierarchical

information units as well as attributed, typed, directed, multi relation and weighted relations.

A diverse domain like healthcare with heterogeneous data sources includes the concepts: patients, doctors, treatments, medication, hospitals and procedures. Treating these concepts as if they are of the same type leads to the loss of valuable semantic information. Similarly, when each concept is treated as a distinct type, important schema level information is lost. It is thus important to know that patients and doctors are of different types when compared to one another, but are of the same type (person) when compared to a different type, such as procedures. As a result, concepts must be modelled as typed, hierarchical information units and/or relations (Sun & Han, 2012).

Section 2.3.3 formalises the above discussion regarding the properties of different types of information networks.

2.3.3 Homogeneous and heterogeneous information networks

In Section 2.3, an information network is defined as:

> *"... a directed graph* $G = (V, E)$ *with an object type mapping function* $\tau: V \rightarrow A$ *and a relation type mapping function* $\emptyset: E \rightarrow R$*, where each information unit* $v \in V$ *belongs to one particular object type* $\tau(v) \in A$*, each relation* $e \in E$ *belongs to a particular relation type* $\emptyset(e) \in R$*, and if two relations belong to the same relation type, the two relations share the same starting object type as well as the ending object type."*

Han (2009) makes the following distinction between homogeneous and heterogeneous information networks:

> "*When the types of information units* $|A| > 1$ *or the types of relations* $|R| > 1$*, the network is called a heterogeneous information network; otherwise, it is a homogeneous information network.*"

As all the information units and all the relations belonging to an ontology are of the same type, an ontology is considered a homogeneous information network. BisoNets on the other hand are constructed from various types of information units and relations, therefore known as a heterogeneous information network. As a result, this makes them well suited as a tool for the modelling of heterogeneous data sources such as data belonging to the healthcare sector.

Due to the complexity of heterogeneous information networks, a meta-structure, namely a network schema, is used to describe the network. Han (2009) defines a network schema as follows:

> *"A network schema, denoted as* $T_G = (A, R)$*, is a meta template for a heterogeneous network* $G = (V, E)$ *with the object type mapping* $\tau: V \rightarrow A$ *and the relation mapping* $\emptyset: E \rightarrow R$*, which is a directed graph defined over object types* A*, with relations as relation types from* R*".*

This network schema specifies type constraints on the sets of information units as well as the relations between them. This structures these units and relations into multiple types, creating semi–structured heterogeneous information networks.

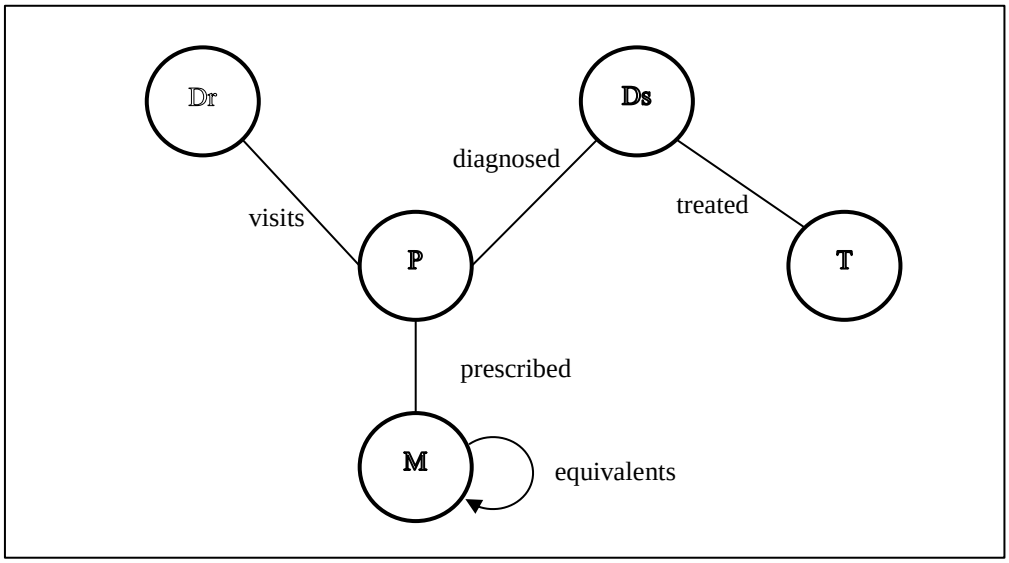

Figure 7: Schema of a healthcare network

As mentioned in Section 2.3.2 the healthcare sector lends itself to being modelled as a semi-structured heterogeneous information network. Typed information units such as *patients* (P), *doctors* (Dr), *diseases* (Ds), *medication* (M) and *treatments* (T) form part of this network as illustrated in Figure 7. For each patient, $p \in P$, there are links to a set of doctors. These links are of the relation type *visits*. Further, there are links to a set of diseases which belong to the relation type *diagnosed with*. And lastly, the network contains links to a set of medications that belong to the relation type *prescribed*. For each medication, $m \in M$, there are links, of the relation type *equivalents*, to a specific set of medication as illustrated by the arrow in Figure 7.

2.4 Pattern recognition and machine learning

Bishop (2006) describes pattern recognition as the automatic discovery of regularities in data through the use of machine learning algorithms. The field of machine learning is about the development of these algorithms as well as the activity by which these discovered patterns are acted upon; for instance, by classifying data into categories. Furthermore, the knowledge discovery

process that applies machine learning algorithms to data in search of regularities is commonly referred to as data mining. Mitchell (1997) describes the learning activity of an algorithm as follows:

> *"A computer program (algorithm) is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T, as measured by P, improves with experience."*

The field of machine learning concerns itself with the use of algorithms which have the ability to discover patterns in data (*E*) and describe these patterns in terms of a model (*T*). The model's ability to perform its task is measured against a performance measure (*P*). This is referred to as the learning phase. The trained model then executes its tasks (*T*) on unseen data; this step is known as generalisation.

During the learning phase, either supervised or unsupervised learning can occur depending on how the experiences (*E*) are presented to the algorithm. When the training data comprises of input values with their corresponding output values, the task is referred to as a supervised learning task, such as classification or regression tasks. However when the training data comprises of only input values, without their corresponding output values, the task is referred to as an unsupervised learning task, such as clustering groups of similar experiences.

Recall from Section 2.3.2 that the healthcare sector can be modelled as a semi-structured heterogeneous information network. The following section introduces models and algorithms used for mining heterogeneous information networks, drawing from the works of Sun (Sun & Han, 2012), Han (Han, 2009, Sun & Han, 2013) and Ji (Ji, 2012).

2.4.1 Mining heterogeneous information networks

Information in heterogeneous information networks (HIN) propagates across objects of various information unit types by means of links, made up of relations belonging to different relation types. Links indicate the interactions between the various information unit types and imply similarity or influence among the objects. Different links carry different semantics and differ in the strength of their ability to determine "influence" across the linked objects. According to Sun *et al* (Sun & Han, 2012), (Sun & Han, 2013) this principle of exploring the power of links lays the foundation of the methodologies for the mining tasks related to KDD from heterogeneous information networks. Sun *et al* (Sun & Han, 2012) proposes the following three new principles that can guide the KDD process for these type of information networks.

The first new principle is that objects in a network are interdependent, and knowledge can only be discovered by taking a holistic view of the information propagated through the network. The challenge is how to propagate information across heterogeneous types of objects and relations. Drawing from existing work done on homogeneous information networks, this includes the computation of ranking scores, similarity scores and clustering across heterogeneous objects and links.

Secondly, heterogeneous information networks are semi-structured by means of a network schema which provides a meta structure for the information network. This meta structure provides guidance for the search and knowledge discovery tasks on the network. Hence the second principle is that the meta structure of an information network provides a useful and powerful way for its exploration.

Thirdly, in a heterogeneous information network, multiple links may exist across different types of objects carrying slightly different semantic meanings. A certain weighted combination of different relation types might best fit a particular application. Hence the third principle is that the desired

relation type combination can automatically be selected by a mining task based on a user's guidance by means of appropriate weight allocation.

The mining tasks developed for heterogeneous information networks are designed according to these three principles:

- Clustering and classification of heterogeneous information networks,
- Meta-path-based similarity search and mining, and
- User — guided relation strength-aware mining

Table 2 shows a summarisation of the tasks.

Table 2: Data mining tasks for a heterogeneous information network

Principle	Task Name	Task Type	Learning	Algorithm	Aim
Clustering & Classification	Frequent pattern mining	Clustering	Unsupervised	JIM	To find graph fragments that are contained in many graphs.
	Hierarchical clustering	Clustering	Unsupervised	DistMatrix	To derive domains from common neighbourhoods within the graph structure.
Meta-path based mining	Meta-path based relationship prediction	Prediction	Supervised	Supervised relation Prediction framework	To extract topological features of a HIN, within a given past time interval and using this to predict relationship building for a future time interval. A relationship can be an immediate link or path instance following some meta-paths.
User-guided strength aware mining	User-guided clustering via meta paths	Clustering	Unsupervised	PathSelClus	The selection of an appropriate combination of weighted meta- paths for generating desired cluster results, by making use of user guidance.

2.4.1.1 Frequent pattern mining

Frequent patterns include item sets and/or substructures that repeat in data more frequently than a minimum threshold, specified by a user. When the structure of the data found within the frequent pattern is of a graph type structure, the frequent pattern is referred to as a subgraph or structural pattern. Frequent patterns play an important role in the field of KDD specifically for the discovery of interesting relationships within the data.

Agrawal (1994) first proposed frequent pattern mining and it was used for market basket analysis. However since then, frequent pattern mining has developed into a focused theme in data mining research (Han, Cheng, Xin, & Yan, 2007) and the three basic methods for these types of mining algorithms include:

- Apriori,
- FP-Growth, and
- Eclat.

The key difference between these methods is that Apriori and FP-Growth mine frequent patterns from observations represented in a horizontal data format. To clarify, every observation is assigned a unique identifier (ID). The data are then represented by a list of observation ID's and each ID is followed by the items linked to that observation. Zaki (1997) introduced the Eclat algorithm which makes use of vertical observation representation. That is to say, that the data are represented by a list of items and for each item, the ID's of the corresponding observations that contain the item are listed. Drawing from the work of Kötter *et al* (2012a) the Jaccard Item Set Mining (JIM) algorithm, which is an extended form of the Eclat algorithm, was used for the purpose of this study.

The Eclat algorithm generates cluster item sets by making use of equivalence class clustering. From each cluster, a frequent item set is

generated using bottom–up traversal (Zaki *et al.*, 1997). Because of the vertical observation representation used to cluster items, only one complete database scan is required which makes this a fast clustering algorithm with low memory utilisation. The Jaccard similarity coefficient is a statistic used for comparing the similarity and diversity of item sets. The JIM algorithm makes use of the Jaccard similarity coefficient rather than item set support to determine the threshold that the items in an item set must exceed (Segond & Borgelt, 2011).

2.5 Summary

In this chapter the four scientific fields which form the theoretical basis of this study were introduced. The first scientific field is biomedical informatics; the development of the field was outlined followed by the introduction of clinical informatics, one of the sub-field disciplines that this study focusses on. This was followed by a discussion of the uniqueness of the healthcare sector, specifically focussing on the complexity of its organisations and the impact of such complexity on their information systems.

The next scientific field introduced was Knowledge Discovery from Data (KDD). The principles of KDD were discussed and, more specifically, bisociative knowledge discovery explained, based on a model of creativity proposed by Arthur Koestler in the 1960's. It is argued that information networks are the most promising way to represent habitually incompatible domains and hence the third scientific field to be introduced was information networks.

Information networks were introduced as graph structures based on graph theory. The two most prominent types of information networks used within this study were further detailed, namely ontologies and BisoNets.

The last scientific field introduced in this chapter was machine learning and especially, frequent pattern mining as proposed by Agrawal in

1994. Frequent pattern mining as a tool for mining information networks was discussed.

In this study the following aspects will be addressed:

- **It is difficult to discover knowledge from data of complex organisations**.

  Modern KDD methods have had limited success in the field of biomedical informatics when compared to the success in other fields of human endeavour including marketing, banking, customer relationship management, engineering and various natural science fields. Bisociative knowledge discovery was introduced in this study as an alternative method for knowledge discovery from biomedical data.

- **The data of complex organisations needs to be modelled differently.**

  Due to the complexity of the world we live in, data or concepts that represent real-world situations interact with each other forming complex, interconnected networks. These interconnected networks are called information networks. For this study, bisociative information networks were introduced, specifically for modelling the data of diverse domains into an integrated information network, which supports the discovery of bisociative knowledge.

- **It is difficult to reduce the aspects describing complex organisations to data and hence, complicated to automate the processing of finding meaning within these organisations.**

  Healthcare organisations are complex. This is due to the fact that healthcare concepts such as recognising a sick patient or defining a headache are hard to reduce to data. This makes it complex to represent the information in a format that is computationally

useful. As a result, a wide semantic gap exists between the data and the information. Therefore, it is a challenge to develop tools that can automate the processing of meaning in the healthcare sector.

The methods and methodology developed and described in the next chapter were based on the theories discussed in this chapter.

Chapter 3

Knowledge discovery model for biomedical informatics

The fourth paradigm for scientific research was defined by Jim Gray (2007) in his presentation, *A Transformed Scientific Method*, to the Computer Science and Technology Board of the National Research Council in 2007. Gray argued that originally only experimental and theoretical sciences were recognised by scientists as the basic paradigms for explaining phenomena observed in nature. Soon these theoretical models grew too complicated to be solved analytically and a third paradigm, computational simulation, was introduced.

During the past few decades, simulations have been used successfully to model many complex phenomena such as the big bang theory, the human genome, predicting climate change and many more. However, simulations as well as the increase in data generated from the experimental sciences, are contributing to the data deluge the world is currently experiencing. This calls for the emergence of a fourth paradigm of scientific research, namely, data-intensive science, with new research methods and methodologies (Hey, Tansley, & Tolle, 2009), (Bell, Hey, & Szalay, 2009).

In this chapter, a model that addresses the challenges of data–intensive science, dedicated to the field of biomedical informatics, is proposed. Firstly, a framework that consists of the methods which are based on the theories

discussed in Chapter 2 is introduced. This is followed by the development of a process model within the proposed framework.

The chapter is organised as follows: Section 3.1 introduces the concept of data-intensive science and its impact on the existing scientific method. Section 3.2 discusses the adaption of the scientific method to accommodate data-centric research. This is followed by Section 3.3 which introduces the unified knowledge discovery framework, a modelling approach for data–centric research. Lastly, in Section 3.4 the unified framework is further developed into a data mining and knowledge discovery process model for bisociative knowledge discovery, and the stages of this process model, as pertaining to this study, are discussed.

3.1 Scientific method

The scientific method, as we know it, involves the stating of a hypothesis, conducting experiments, analysing the collectable evidence and drawing conclusions that either prove or reject the hypothesis. Chris Anderson (2008), Editor-in-Chief of Wired magazine authored the controversial article *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, in which he argues that the scientific method is becoming obsolete in the realm of data-intensive science.

According to the scientific method, if a correlation exists between two variables, no conclusion could simply be drawn between the two variables. Instead one has to understand the relationship between the variables and demonstrate this understanding through the construction and presentation of a scientific model. That is to say, a model is required to formalise the cause of any correlation. Therefore, according to the scientific method, correlation on its own does not imply causation and as a result data without a model is considered worthless.

To the contrary, with the emergence of data-intensive science, computing clusters and machine learning algorithms are used to discover

correlations and patterns, which scientists were unable to identify before, in vast amounts of data. That is to say, scientists are now able to analyse data without first hypothesising what the data might reveal. This implies that there is no need for scientists to identify testable models any longer to prove their hypothesis. Correlations and patterns discovered by machine learning algorithms in data-rich environments open up a new way of interpreting the world. This new approach offers a different method for conducting scientific research in a data-rich environment. Anderson concludes that

> *"Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all*" (C. Anderson, 2008).

In a response to Anderson's article, Massimo Pigliucci poses the question *"... if we stop looking for models and hypotheses, are we still really doing science?"* (Pigliucci, 2009). He argues that science is not just about finding patterns and that science advances only if it can provide an explanation for the patterns found. Hence, correlation must be explained to imply causation.

3.2 Scientific method for data-rich environments

The fourth paradigm of scientific research, as defined by Gray, recognises a need to establish a new research methodology for data-centric research. Vast amounts of data alone do not lead to data-intensive science, and yet the scientific method as we know it is incapable of addressing the complexity of the data-rich scientific environments in which research is conducted. Data-intensive science takes an exploratory approach to data–centric research where information emerges from the data, compared to the more traditional

confirmatory approach which examines hypothesised patterns expected from the data (Newman, Ellisman, & Orcutt, 2003), (Michener & Jones, 2012).

This exploratory approach necessitates the modelling of large amounts of heterogeneous data from multiple sources, followed by the analysis of the data models using techniques tailored for the discovering of complex patterns in heterogeneous data. The novel and surprising patterns discovered during the analysis provides valuable insights for concrete hypothesis about the underlying real-world domains modelled by the data. The results complement the scientific method of generating a hypothesis and experimental testing.

This requires the adaption of the scientific method to include methods and technologies required to perform data-centric research. Instrumental contributors to these new methods and technologies are (Buchan, Winn, & Bishop, 2009):

- the advances in data modelling focussing specifically on graph databases i.e. information networks as discussed in Section 2.3,
- the developments in the field of machine learning as discussed in Section 2.4, and
- recent developments in the field of computer science, namely, computational thinking, as discussed next.

3.2.1 Computational thinking

Computer science is the study of computation, the study of what can be computed and how to do so. The formal foundation of computer science lies within the scientific field of mathematics. The study of computation necessitates conceptualising a problem by thinking at multiple levels of abstraction. Computational thinking draws from this fundamental concept of computer science.

Computational thinking involves solving problems, designing systems and understanding human behaviour by decomposing complex phenomena into smaller fragments and solving each fragment by applying computational concepts. It is a form of analytical thinking that overlaps other fields, such as (Wing, 2006):

- mathematics — the methods for solving problems,
- engineering — the methodologies to approach, design and validate complex systems, and
- scientific thinking — the understanding of computability, intelligence and human behaviour.

Wing (2008), in the *Philosophical Transactions Series* of the Royal Society, defines abstractions as the mental tools used during computation, with which one conceptualises the decomposed components of complex problems. These abstractions form the essence of computational thinking and are extremely general. They have a tendency to be more complex than the numerical abstractions used within the field of mathematical sciences and differ in the following ways.

Firstly, a mathematical abstraction, such as integers, is structured and precise and adheres to specific well-defined algebraic properties. In contrast, computational thinking abstractions don't necessarily adhere to these rules. An algorithm for instance is an abstraction of a step-by-step procedure to convert specific input into required outputs. What does it mean to add two algorithms? This combining procedure will become an abstraction by itself and will require careful thought.

Secondly, computational thinking abstractions are implemented within the constraints of the real world and therefore should be concerned with boundary values determining the edge conditions of the abstraction as well as its failure conditions. Computational thinking involves thinking in terms of prevention, protection and recovery from worst-case scenarios. For

example what happens when a server is down, the disk is full, and/or the processor is out of memory?

And lastly, in a data-rich environment, it is critical to the abstraction process how well the abstraction reflects the problem it is conceptualising. Hence establishing which variables should be incorporated into the abstraction and which can be ignored form the foundation of computational thinking.

Wing (2008) explains that during a computational thinking exercise, within which a complex problem is being decomposed, an abstraction process is involved which establishes layers of abstraction. Within computing there will always be at least two of these layers. The layer of interest, as well as the layer above or below it. These layers interact with one another by means of an abstraction function. For example, when using an application programming interface (API) to develop a graphical user interface (GUI), the user is concerned with the functionality of the GUI and need not to know anything about the implementation of the API. Therefore the user's layer of interest is the GUI and the layer below it will be the API. However the developer is concerned with implementation of the API to ensure the required functionality is available to the user. Hence the developer's layer of interest is the API and the layer above it is the GUI. A well-defined abstraction function integrates these two layers of abstraction. This concept of working with multiple layers of abstractions and understanding the functions that integrate them makes it possible to build large, complex systems for solving complex problems.

Hence computational thinking involves defining solutions to problems in terms of multiple layers of abstractions, working with these different layers of abstractions and understanding the relationships between them. Accordingly, computing involves the automation of these layers of abstractions. This automation can take place by making use of a machine like

a computer, or a human, or a combination of the two depending on the task at hand.

By using smarter and more sophisticated abstractions, Wing (2006) proposes that computational thinking will not only help the development of more complex systems, but also support the analysis of large volumes of data. Computational thinking will be able to extract hidden knowledge from data by means of the development of abstractions to represent and process the data.

Taking these new developments into account led to the development of a knowledge discovery framework for the healthcare sector which included a computational model for biomedical informatics, as discussed next.

3.3 Modelling approach for data-centric research

Buchan *et al* (Buchan *et al*., 2009) proposes a unified modelling approach to data-centric research, specifically referring to the healthcare sector. The use of electronic health records (EHRs), scientific outputs and biotechnologies are mostly responsible for the rise in available healthcare data. Because of the increased use of standard terms and ontologies, especially in EHRs, healthcare data has gradually become more structured and as a result can be shared more easily. Initiatives like the NIH-supported Biomedical Informatics Research Network (BIRN) project is making significant contributions towards the ease of data collection and sharing within the biomedical research community (Newman et al., 2003), ("Biomedical Information research Network," n.d.). Nonetheless there is still very little available, in terms of healthcare models, that can be applied to this wealth of data in order to extract meaning.

Outputs from healthcare research have been growing exponentially and the hypothesis-driven scientific method has served it well. However, Buchan *et al* (2009) argues that this method is inadequate for reflecting the complexity of the healthcare sector. To demonstrate, during clinical drug trials, up to 80 per cent of possible cases could be excluded from the trials for

the reason that the conventional models are unable to accommodate more complex situations. For example, patients with multiple co-morbidities or patients that use multiple medications are excluded. Also the EHRs of patients from physicians outside the study, to whom the drug is not prescribed, could be used as a natural control group if the conventional models were capable of integrating the data of these patients.

Hence Buchan *et al* (2009) argues that data alone will not lead to data-intensive healthcare and a new methodological approach is required to address the complexities of the healthcare sector. Instrumental to this methodology is the use of machine learning, combined with computational thinking, for the development of computational frameworks that seek patterns in collections of healthcare data.

3.3.1 Unified knowledge discovery framework

Drawing from the work of Buchan *et al* (Buchan *et al*., 2009) the following knowledge discovery framework, that takes advantage of the wealth of data available and furthermore supports the complexity of the healthcare sector, is proposed as illustrated in the Figure 8.

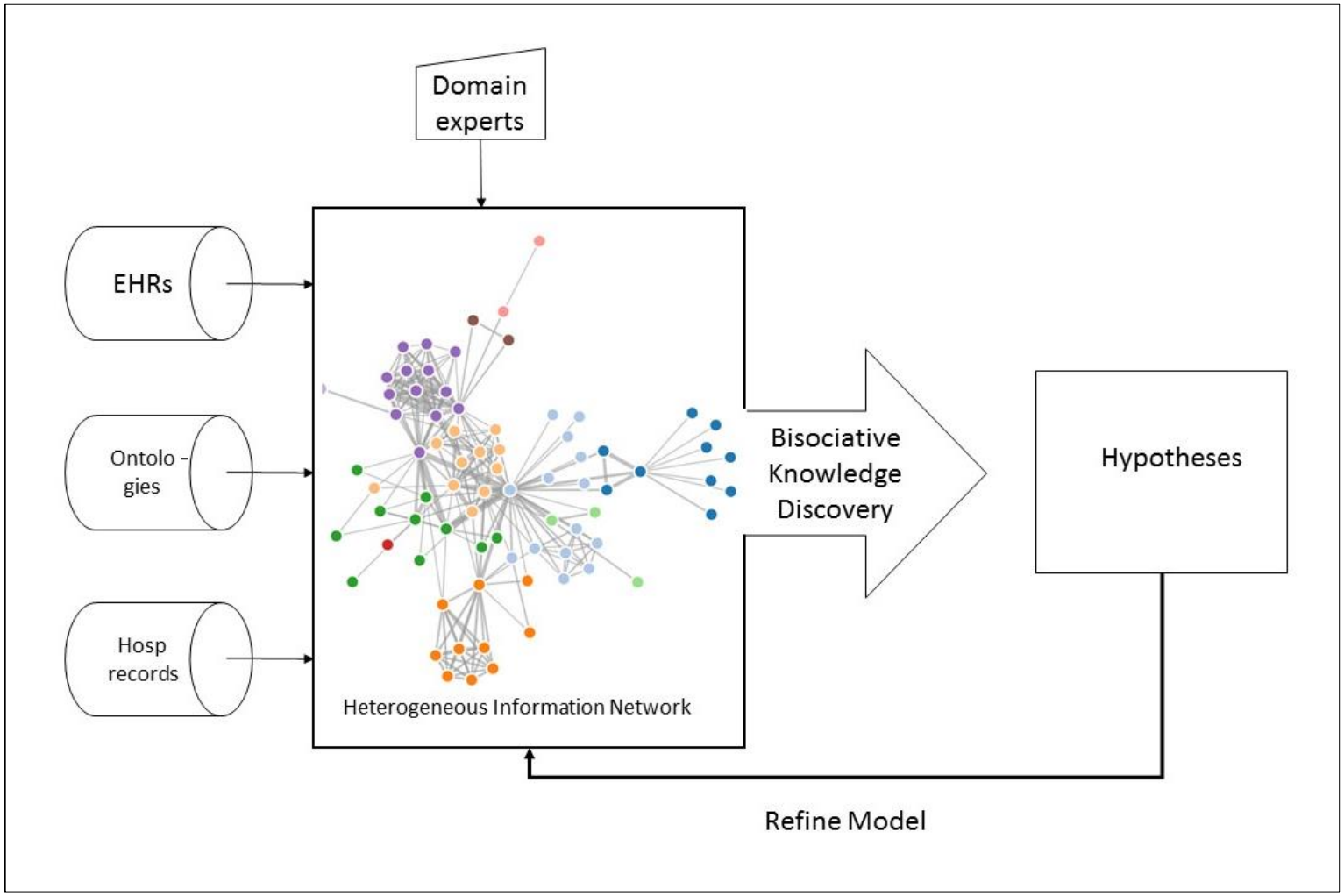

Figure 8: Unified knowledge discovery framework

Within this framework, data are represented as an integrated heterogeneous information network (IHIN). Data are sourced from diverse data repositories to form different domains. In the healthcare sector, this data will include the data available in EHRs, data collected from hospital records, ontologies, as well as domain knowledge from experts. The IHIN is the integration of these domains, each representing a knowledge base. Machine learning algorithms are applied to the IHIN with the purpose of finding domain-crossing relationships and this process is called bisociative knowledge discovery.

The extracted knowledge is analysed and then used to either refine the IHIN or generate a concrete hypothesis for further investigation. The development of a knowledge discovery process model based on the unified knowledge discovery framework is discussed next.

3.4 Knowledge discovery process model

The knowledge discovery process refers to the overall process of discovering knowledge from data. Fayyad *et al* (1996) defines it as

> *"... non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data"*

and proposes the following 5 stages for the so-called KDD process namely:

1. Selection — creating a target data set
2. Pre-processing — cleaning the target data set to obtain consistent data
3. Transformation — reducing dimensionality and finding useful features to represent the target data
4. Data mining — searching for patterns of interest
5. Interpretation / evaluation — interpreting and evaluating the discovered patterns

This is an iterative and interactive process.

A consortium composed of DaimlerChrysler, SPSS and the NCR developed the CRISP-DM process for knowledge discovery in 1999 to establish a standard for knowledge discovery projects. This was to prove to industry that knowledge discovery was mature enough to be adopted as a key part of their business processes (Pete *et al*., 2000). The CRISP-DM process consists of the following 6 stages which again are interactive and iterative:

1. Business understanding — convert the business needs into a data mining problem definition
2. Data understanding — data collection and initial exploration
3. Data preparation — construct a target data set from raw data
4. Modelling — search for patterns of interest
5. Evaluation — evaluate the discovered patterns
6. Deployment — present knowledge gained in a way that the customer can understand and use the insights

Cios *et al* (Cios & Kurgan, 2005), (Cios & Moore, 2002) introduced the Data Mining and Knowledge Discovery (DMKD) process model for knowledge discovery as a tool to enable industry to model real-world problems. The DMKD model is based on the CRISP-DM model, with the addition of more feedback mechanisms—which makes it more automated—as well as the capability to interact with other problem domains, such that the discovered knowledge may also be applied to these domains. The six stages of the DMKD model, which again are interactive and iterative, are as follows:

1. Understanding the problem domain — work with domain experts to define the problem
2. Understanding the data — collect data and verify their usefulness
3. Preparing the data — feature cleaning, data selection and network construction
4. Data mining — using data mining tools to extract knowledge
5. Evaluating the discovered knowledge — ensuring data owners can understand the results
6. Using the discovered knowledge — defining where and how the discovered knowledge will be used

Table 3 compares the 5-stage KDD process model with the 6-stage CRISP-DM process model and the 6-stage DMKD process model as described above.

Table 3: Comparison of three KD process models

KDD	CRISP-DM	DMKD
	1. Business understanding	1. Understanding the problem domain
1. Selection	2. Data understanding	2. Understanding the data
2. Pre–processing		
3. Transformation	3. Data preparation	3. Preparing the data
4. Data mining	4. Modelling	4. Data mining
5. Interpretation/ evaluation	5. Evaluation	5. Evaluating the discovered knowledge
	6. Deployment	6. Using the discovered knowledge

When comparing the three KDD process models, it can be concluded that CRISP–DM and DMKD are implementations of Fayyad's KDD model,

incorporating two additional stages: one that precedes and one that follows the KDD process. The preceding KDD stage for both CRISP-DM and DMKD is to work with the domain experts to determine the project goals and translate them into data mining problem definitions. The post-KDD model stage concerns the incorporation of the discovered knowledge into the problem domain. An opinion poll conducted in October 2014 by KDDNuggets probing *"What main methodology are you using for your analytics, data mining, or data science projects?"* showed that CRISP-DM is still the process model of choice. This is shown in Table 4 (Piatetsky, 2014).

Table 4: KDDNuggets poll results on the most commonly used methodology for data-intensive research (Piatetsky, 2014)

What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total] 2014 poll 2007 poll	
CRISP-DM (86)	43% 42%
My own (55)	27.5% 19%
SEMMA (17)	8.5% 13%
Other, not domain-specific (16)	8% 4%
KDD Process (15)	7.5% 7.3%
My organizations' (7)	3.5% 5.3%
A domain-specific methodology (4)	2% 4.7%
None (0)	0% 4.7%

The DMKD process model is similar to CRISP-DM and has been successfully applied to numerous problems within in the healthcare sector (Cios, Teresinska, Koniezna, Potocka, & Sharma, 2000), (Sacha, Cios, & Goodenday, 2000). The strength of the DMKD process model lies in its extensive feedback loops, as illustrated in Figure 9, hence it is the most

appropriate process model for the conducting of bisociative knowledge discovery in the healthcare sector.

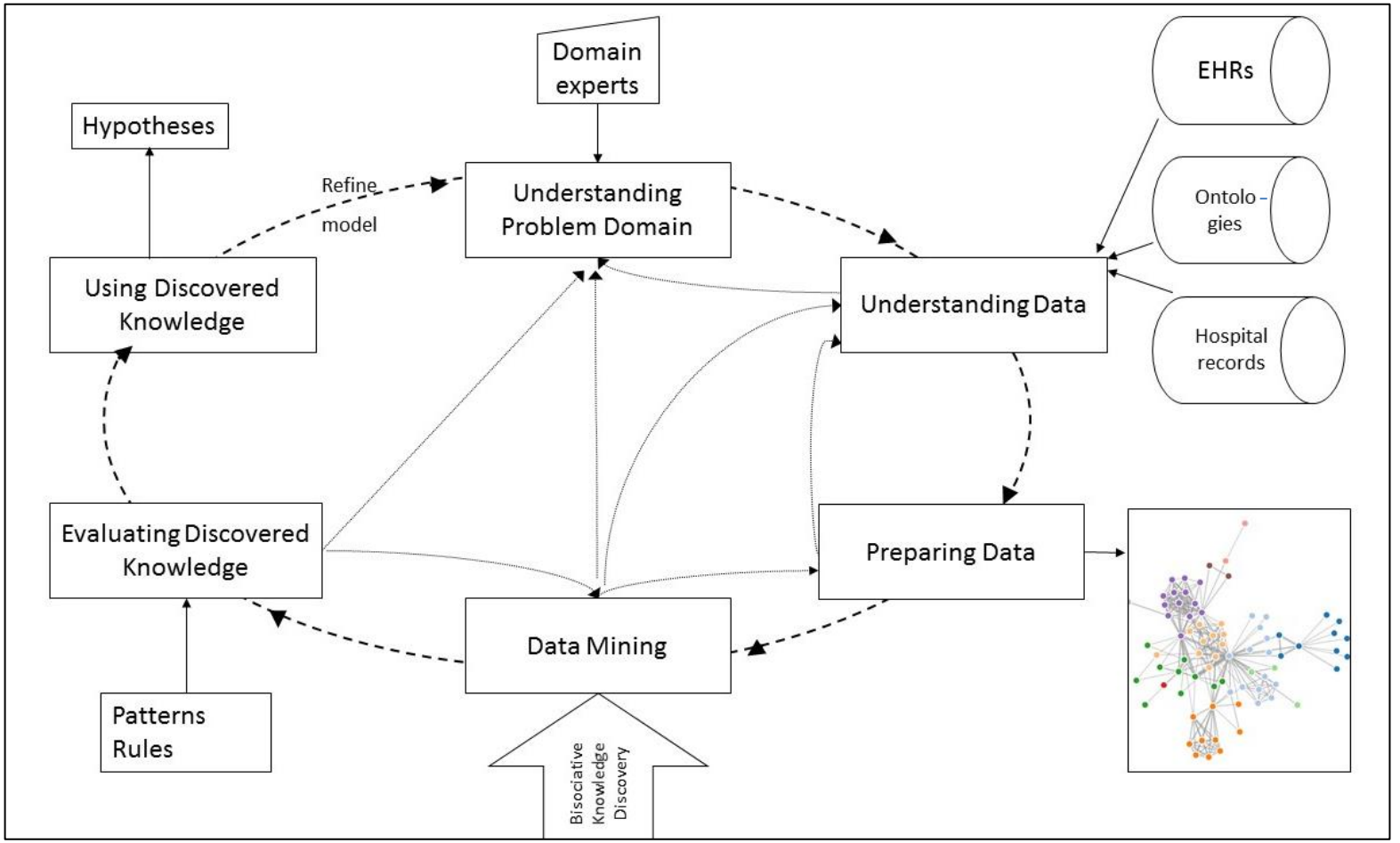

Figure 9: DMKD process model for bisociative knowledge discovery in the healthcare sector

Kitzes *et al* (2017) in *The Practice of Reproducible Research* argues that data-intensive science must strive to be computationally reproducible. Due to the advances in technology and hence the incorporation of computational methods into research methodologies, it has become necessary to make data-centric research more clear, transparent and organised. This can be accomplished by following the guidelines of computationally reproducible research.

Computational reproducibility entails

> *"... that when provided with identical source code, input data, software and computing environment configurations, that an independent party can exactly reproduce the results of the original work* (Kitzes *et al*., 2017)".

The three key practices are the following:

- the automation and provenance tracking,
- availability of the data and software, and
- open reporting of results.

Automation entails that the computational aspects of the data preparation, mining and evaluation steps are encoded in software and documented in such a way that these steps can be replicated automatically. Provenance tracking involves the documentation of the platform on which these automated steps are executed in such a way that the platform can be replicated.

A key component of computational reproducibility is the public availability of the data and software. This is accomplished by making use of data available on public-accessible databases and using open source software that is downloadable.

Lastly, it is important to provide sufficient details of the research in a widely accessible form. The choice of where to publish has a direct influence on the accessibility of the findings. Choosing an open access platform, which is online, allows wider access to articles as access is free and requires no subscription fee.

Kitzes *et al* (2017) further recommends a basic computational reproducible workflow which incorporates the three key practices in three stages namely: data acquisition; data processing and data analysis. These stages directly map to the data preparation, data mining and discovered knowledge evaluation stages of the DMKD process model, as used for this study. This study incorporates the key practices of reproducibility within the applicable stages of the DMKD process model and therefore produces computationally reproducible research within the realm of data-intensive science.

Next, the stages of the DMKD process model for bisociative knowledge discovery in the healthcare sector are discussed:

- data preparation stage
- data mining stage, and
- the discovered knowledge evaluation stage

3.4.1 Data preparation stage

Cios *et al*. (Cios *et al*., 2000) states that data preparation is the key step within the DMKD model on which the success of the entire knowledge discovery process depends. Literature reveals that this stage usually takes up most of the time in any KDD project, at least half of the effort is spent on data preparation. The subsequent outcome of this stage is a data model populated with relevant data that serves as the input to the following stage of data mining.

The main purpose of the data preparation stage is the construction of the integrated heterogeneous information network which consists of data sourced from diverse sources. This is accomplished by identifying the domains and representing each as a heterogeneous information network. Next, for the purpose of this study, a BisoNet, as introduced in Section 2.2.2, was the framework used for the unification of the loosely coupled domain data.

As mentioned in Section 2.3, graphs are the chosen way of modelling these information networks, with each vertex representing an information unit and the relations between these units represented by edges. Modelling these information networks is an abstracted activity during which the different features of each information network are denoted within a graph format. This allows for the structuring and manipulation of the features (Robinson, Webber, & Eifrem, 2015).

This study made use of GRAD, a generic database model especially designed for the advanced modelling and sophisticated analysis of graph data. GRAD was introduced in February, 2016 by Ghrab *et al*. in the paper *GRAD: On graph database modelling* (Ghrab, Vaisman, Zimányi, Romero, & Skhiti, 2016) as

> *"… a database model, as defined by Codd that consists of a set of (1) data structures, (2) integrity constraints, and (3) manipulation operators."*

Most leading graph databases are currently built on property graph models. However, these models are not complete data models as described by Codd and tend to be only data structure definitions. For more advanced data modelling and analysis, a semantically richer data structure is required with the ability to define integrity constraints and formulate manipulation operators. GRAD is a complete graph database model that includes advanced

data structures, has a set of rules that enforces the integrity of graph data, and a set of operators that enable analysis.

3.4.2 Data mining stage

As introduced in Section 2.4.1.1, frequent subgraph mining is closely related to frequent item set mining, with the difference being that the purpose of the frequent subgraph mining is to discover frequent subgraphs as an alternative to discovering frequent subsets. These discovered subgraphs, called concept graphs, represent a high level abstraction of the integrated heterogeneous information network, which in turn represents the loosely coupled domain data.

Drawing from the work of Kötter *et al* (Tobias & Berthold, 2012), Nagel *et al* (Nagel, Thiel, & Tobias, 2012) and Schmidt *et al* (Schmidt, Kranjc, Mozetic˘, Thompson, & Dubitzky, 2012) frequent subgraph mining, which involves the use of unsupervised, clustering, machine learning algorithms, was used during the datamining stage with the intent to discover subgraphs, and more specifically, concept graphs for the extraction of information from integrated heterogeneous information networks.

A concept graph can be defined as a bi-partite subgraph from an information network which consists of two vertex partitions of which one partition represents the concept members (information units) and the other the properties that these members share (represented by the direct neighbours of the information unit), called aspects. The concept graph describes a concept. This concept is represented by an information unit that is related to the vertices of the member set within the concept graph only and to no other information units in the information network. Concept graphs are based on the assumption that similar information units share more properties than dissimilar information units. For this reason, concept graphs represent dense subgraphs in an information network made up of two disjoint, but fully

connected, sets of vertices of which one represents the concept members and the other the aspects that the members share (Tobias & Berthold, 2012).

When the information network, from which the concept graph is extracted, has the properties of an integrated, heterogeneous information network, which implies that the information units originate from diverse domains, this *k*-partite graph can be defined as a BisoNet. Furthermore, in a BisoNet, the direct neighbours of an information unit represent the aspects of that unit. The more neighbours' units shared, the more similar they are and therefore it can be derived that concept graphs are dense subgraphs in a BisoNet. As a result, this allows the detection of domain bridging concepts and therefore has the potential of supporting creative thinking.

Recall from Section 3.1.3 that computational thinking involves solving problems by working with layers of abstractions, as well as the automation of these layers. This makes it possible to automate a layer of abstraction by detecting concept graphs and physically storing these concepts as bi-partite graphs, makes this a computational thinking exercise. Even more so, a discovered subgraph comprises information units. Each information unit represents either a concept member or an aspect of a specific domain. When these information units stem from different domains, the subgraph represents a domain bridging concept graph (Tobias & Berthold, 2012). Domain bridging concept graphs allow us to detect bridging concepts i.e. bisociations which, in turn, supports creative thinking.

3.4.3 Knowledge evaluation stage

A bisociation is a connection of two information units from diverse domains. The purpose of the evaluation stage is to identify these bridging concepts. The bridging concepts can exist within a single concept graph when the graph contains information units from diverse domains. Otherwise, the bridging concepts can be part of two overlapping concept graphs that describe concepts from diverse domains. During the evaluation stage, the novel, potentially

useful concept graphs are visualised with a network visualisation tool named Neo4J, for further analysis.

Neo4J is an open source, native graph database management system with a NOSQL data store, and uses native graph storage (also known as "index-free adjacency") to store and manage connected data. Cypher is Neo4J's expressive graph database query language that supports the basic CRUD methods namely create, read, update, and delete (Robinson *et al.*, 2015).

3.5 Summary

In this chapter, a knowledge discovery framework for data-intensive research dedicated to the field of biomedical informatics was developed, drawing from the work of Buchan. Within this framework, data are represented as integrated heterogeneous information networks. Machine learning algorithms are applied to the data with the purpose of finding domain crossing relationships to support bisociative knowledge discovery.

This framework was then further developed into a knowledge discovery process model for bisociative knowledge discovery specifically in the healthcare sector. Three stages of the process model, namely: the data preparation stage, the data mining stage, and the discovered knowledge evaluation stage, were then discussed in more detail.

In this study the following methods were introduced:

- **Explorative data analysis, using bisociative knowledge discovery, can be a useful approach to discover knowledge from the data of complex organisations.**

  Bisociative knowledge discovery is an exploratory approach which allows the discovery of knowledge from large amounts of heterogeneous data integrated from multiple sources. It allows for the analysis of the data using techniques tailored for the

discovering of novel patterns. In this study, bisociative knowledge discovery was used within a framework with the aim of discovering knowledge from the data of complex organisations.

– **The data of complex organisations can be modelled differently by using information networks as a novel approach.**

  In this framework, data of complex organisations are represented as an integrated, heterogeneous information network. Data are sourced from diverse data repositories to form habitually incompatible domains. The data modelling approach integrates these domains where each represents a different knowledge base as proposed in the DMKD process model, for bisociative knowledge discovery, in the healthcare sector.

– **Complex organisations can be unpacked into different layers of abstraction and the integration of these layers can be automated.**

  Computational thinking is a method by which problems can be solved and systems can be designed. Computational thinking involves the decomposition of a complex phenomenon into smaller parts and then solving each part by applying computational concepts. The solution to the complex problem is defined in terms of multiple layers of abstractions and the functions which integrate these layers.

The proposed knowledge discovery process model was applied to a case study as presented in the next chapter.

Chapter 4

Knowledge discovery in the healthcare sector, a worked case study

In Chapter 3, the DMKD process model for bisociative knowledge discovery, dedicated to the field of biomedical informatics, was introduced. This model incorporates methods and technologies required to perform data-centric research in a novel way. Firstly, this required that the data were modelled as a graph database, with an integrated heterogeneous information network structure, in comparison to the traditional row and column data model of relational databases. Secondly, data mining tasks, developed for heterogeneous information networks, were used to extract frequent subgraphs from the network structure. Lastly, computational thinking methods were applied by decomposing the complex phenomenon into multiple layers of abstraction followed by the automation of the integration of these layers.

This chapter presents the application of the DMKD process model to a case study making use of the Nationwide Inpatient Sample data, which forms part of the Healthcare Cost and Utilization Project, made available by the Agency for Healthcare Research and Quality of the United States of America (HCUP Central Distributor, 2015).

Chapter 4 is structured as follows: each section is dedicated to a different stage of the DMKD process model. Section 4.1 introduces the healthcare problem domain, then Section 4.2 discusses the data used for the purpose of this study in terms of the selection criteria and a description

thereof. This is followed by the most time consuming stage of the process model, the data preparation stage, presented in Section 4.3. This section includes feature selection, data cleaning, domain construction as well as the construction of the integrated heterogeneous information network. Section 4.4 presents the data mining stage, during which the actual knowledge discovery took place, followed by the evaluation of the discovered knowledge as discussed in Section 4.5.

4.1 Understanding the healthcare problem domain

This research took a data-centric point of view and considered anomaly detection in biomedical data. Anomaly detection refers to the problem of discovering patterns in data that do not conform to their expected behaviour (Chandola, Banerjee, & Kumar, 2009). These non-conforming patterns are of significant interest to data analysts and are known as anomalies or strong outliers. These two concepts, anomalies and outliers, are often used interchangeably in literature.

The concept of an outlier has been formally defined by Hawkins as

> *"... an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism"* (Aggarwal, 2013).

Hawkins' definition emphasises the important role of the generating process of patterns. These patterns are a reflection of observations collected about events, amongst other things. When a process behaves in an unusual way, outliers are produced. Hence, outliers contain important information regarding the abnormal characteristics of events and this may potentially lead to new and useful insights that are of interest to data analysts.

A process that adheres to a certain model produces patterns. Outliers are those patterns that deviate from the expected results of this model. These outliers are then classified according to their strength: a strong outlier is

named an anomaly and weak outliers are labelled as noise. The outlierness of a pattern can be quantifiably measured by outlier detection algorithms and are described in terms of its outlierness-score.

Typically, the outlierness-score of an anomaly will be greater than that of noise, as illustrated in Figure 10. Figure 10 demonstrates how noise forms a semantic boundary between normal patterns and true anomalies.

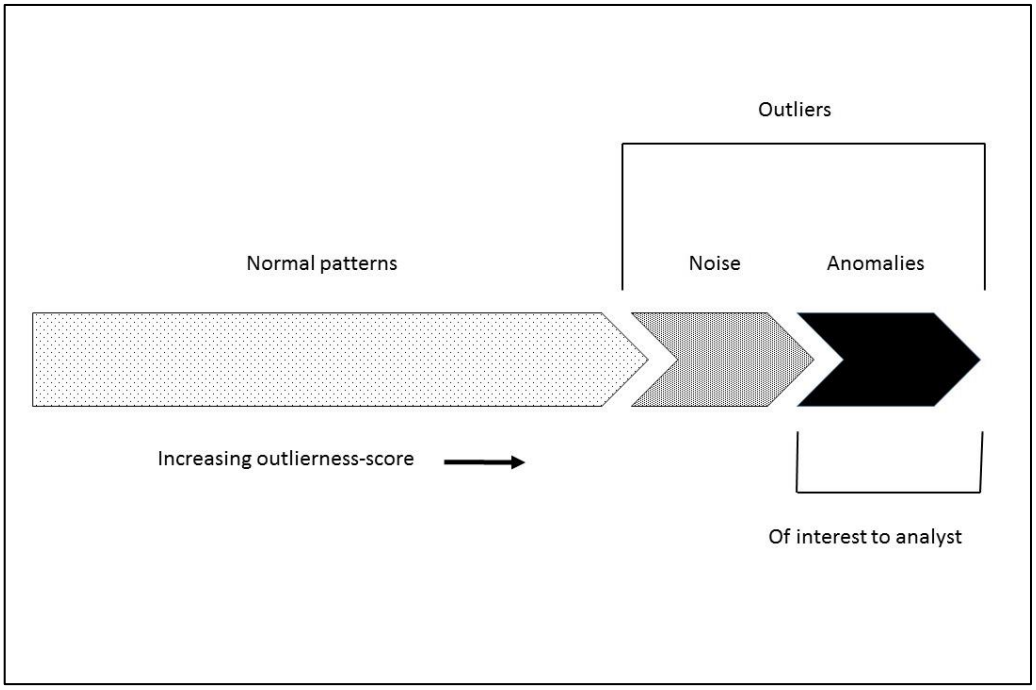

Figure 10: Anomaly distinction (adapted from Aggarwal, 2013)

Although the outlierness-score is an indication of the strength of an outlier, it is not the distinguishing factor. A subjective judgement by the analyst, determined by their interest for that particular situation, is the distinguishing factor which regulates the strength of an outlier i.e. the distinction between an anomaly and noise. Whereas an anomaly is of interest to the data analyst, noise is a mere hindrance.

The real-life relevance is central to anomaly detection. A specialised form of anomaly detection is novelty detection. Novelty detection aims to

detect previously unobserved (novel) patterns in data (Markou & Singh, 2003). A novel pattern will typically be incorporated into a KDD model after being detected, although not true for all anomalies as those without real-life relevance will be discarded. The real-world application for the purpose of this research, concerns the discovery of anomalies in biomedical data.

4.2 Understanding the data

In their position paper *Uniqueness of Medical Data Mining*, Cios and Moore (2002) discuss the major points of the unique nature of biomedical data. Specifically applicable to this research are the following (Cios & Moore, 2002):

- its impreciseness and
- the lack of formal structure of biomedical data.

Firstly, inherent error is embedded in clinical data due to the fact that a medical diagnosis by a healthcare practitioner is based on a synthesis of subjective human observations and objective test results that characterise the medical condition of a patient. It is a process during which a healthcare practitioner attempts to identify a disease or condition that best explains the patient's symptoms, signs and test results. Hence, by nature, biomedical data are imprecise, subjective and prone to error.

Secondly, a lack of formal structure exists into which this imprecise data can be organised. The field of medicine trails behind the other sciences in its ability to characterise its underlying data structures mathematically. This may be due to the multitude of anatomic locations and distinct diseases in its vocabulary, as well as the fact that healthcare is primarily a human care activity, and the justification to collect data is only for individual patient benefit and not for the purpose of analysis. However, the success of KDD largely depends on the ability of a system to tabulate equivalent concepts, hence it relies on the existence of a formal structure.

This study makes use of the Nationwide Inpatient Sample (NIS) data, which forms part of the Healthcare Cost and Utilization Project (HCUP), sponsored by the Agency for Healthcare Research and Quality of the United States of America. The NIS is the largest inpatient care database publically available on a yearly basis. The 2011 extract which was made available at the end of 2014 was used for the purpose of this study. It contains the data of at least 8 million all-payer, hospital stays from more than a 1 000 hospitals across the United States. The large sample size of the NIS supports the analysis of rare conditions and treatments such as specific types of cancer and organ transplants (HCUP Central Distributor, 2015). The data records include clinical information as well as resource use information typically found on patient discharge abstracts, which is discussed in detail in Section 4.3.

After consultation with experts, employed by a South African private hospital group, regarding which treatments have the most disperse cost structures, it was suggested to look into septicaemia and organ transplants. Interestingly enough, septicaemia was ranked as the most expensive condition treated in U.S. hospitals during 2011 (Torio & Andrews, 2013). However, for the purpose of this study, major procedures are of interest, hence the choice to focus on organ transplants.

4.2.1 Data selection

The SNOMED CT ontology, as introduced in Chapter 2.3.1, was the formal structure used for identifying all organ transplant procedures. Within the SNOMED CT concept model under the top level concept: 'Procedure,' all organ transplant procedures, which included solid organ and lung transplant procedures, were identified. Furthermore, the procedure hierarchy was used to identify all the related concepts that were linked by the | is a | relationship attribute, as illustrated in Figure 11 and Figure 12.

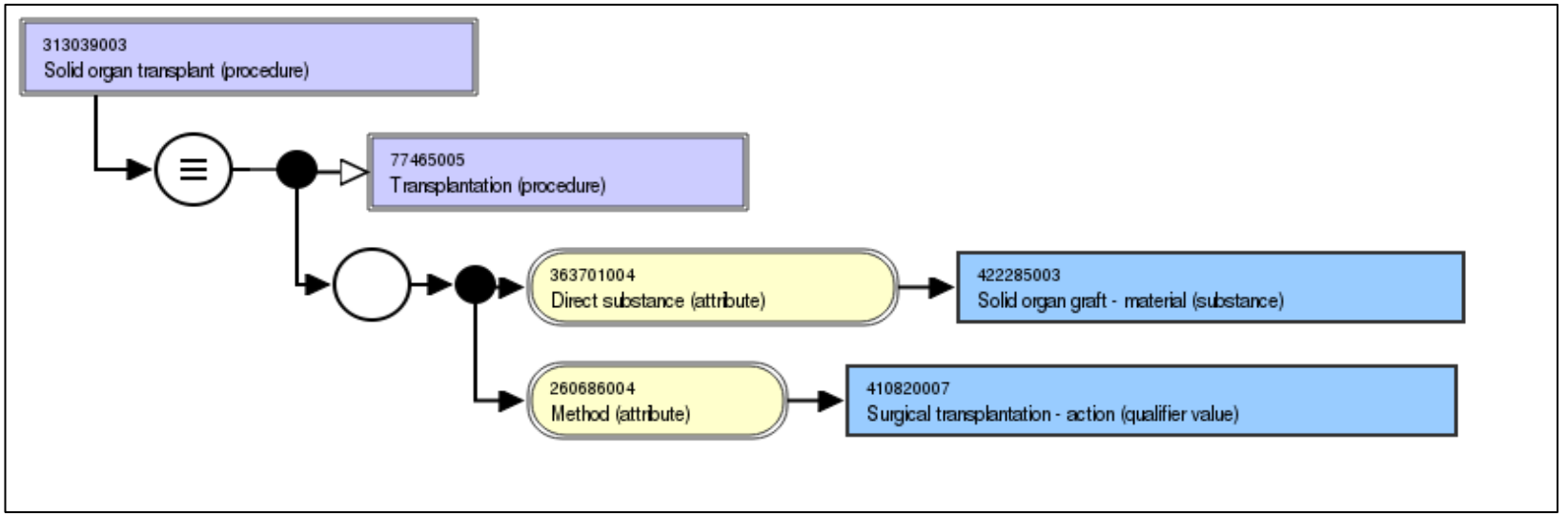

Figure 11: SNOMED CT diagram for the solid organ transplant (procedure) concept

The full definition of the supertype concept "solid organ transplant" can be described as follows:

A solid organ transplant (procedure)

- is a concept within the "procedure" concept model,
- is fully defined,
- has a unique concept identifier of 313039003,
- has a FSN of "Solid organ transplant" which is one of its readable label attribute values
- has a method attribute value of "Surgical transplantation - action" which represents the action to be performed to accomplish the procedure and
- relates to six subtype concepts within the procedure hierarchy:
    - Transplant of kidney
    - Transplant of heart
    - Transplant of liver
    - Transplant of pancreas
    - Transplant of spleen
    - Cadaveric renal transplant

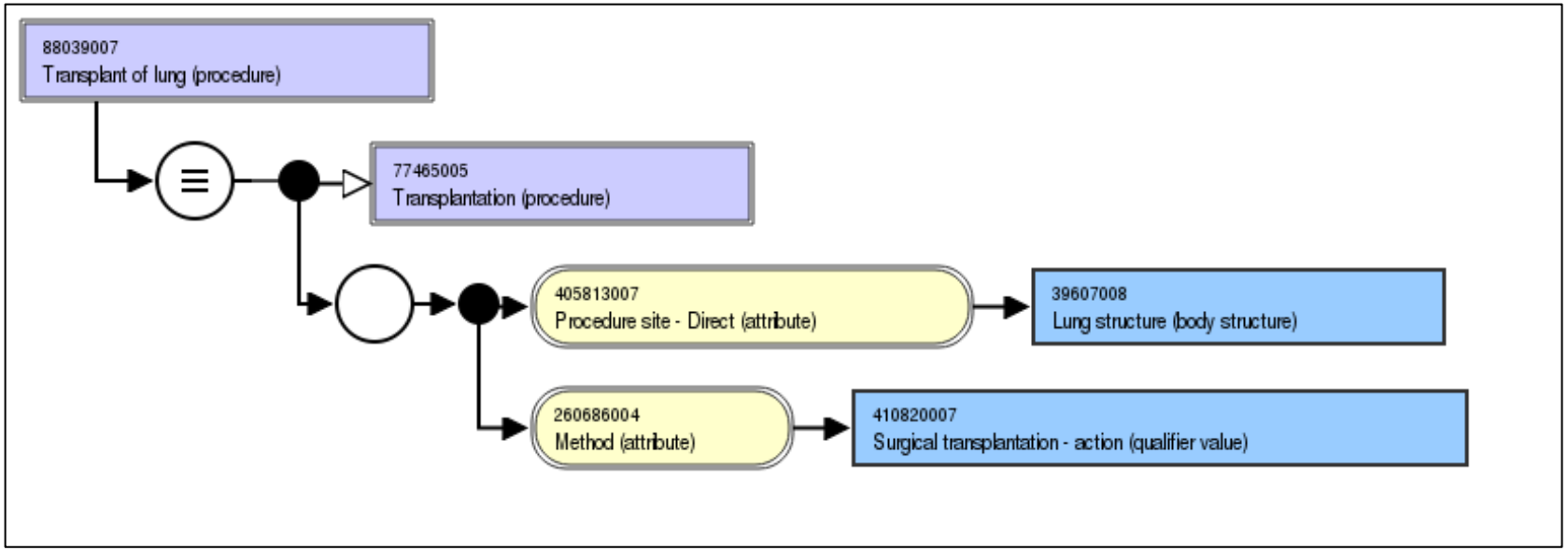

Figure 12: SNOMED CT diagram for transplant of lung (procedure) concept

The full definition of the supertype concept "transplant of lung" can be described as follows:

A transplant of lung (procedure)

- is a concept within the "procedure" concept model,
- is fully defined,
- has a unique concept identifier of 88039007,
- has a FSN of "Transplant of lung" which is one of its readable label attribute values,
- has a method attribute value of "Surgical transplantation - action" which represents the action to be performed to accomplish the procedure,
- has a procedure site attribute value of "Direct- Lung structure", and
- relates to eight subtype procedure concepts within the procedure hierarchy:
    - Allotransplant of heart and lung
    - Autotransplant of lung
    - Bilateral sequential single lung transplant
    - Double lung transplant
    - Heart-lung transplant with recipient cardiectomy-pneumonectomy
    - Reimplant of lung
    - Single lung transplant
    - Transplant of single lobe of lung

The two supertype concepts with all their subtypes were then extracted into a SNOMED CT Refset, as listed in Table 5, to be used during data selection.

.

Table 5: SNOMED CT Refset for transplants

SNOMED Concept FSN - Supertype	**SNOMED Concept FSN –** Subtype	**SNOMED Concept FSN –** Subtype	**SNOMED Concept FSN –** Subtype
Procedure on lung (procedure)	Transplant of lung (procedure)	Single lung transplant (procedure)	
		Double lung transplant (procedure)	
Solid organ transplant (procedure)	Transplant of liver (procedure)	Liver transplant with recipient hepatectomy (procedure)	
		Liver transplant without recipient hepatectomy (procedure)	
		Orthotropic liver transplant (procedure)	Orthotropic transplant of whole liver (procedure)
		Heterotopic liver transplant (procedure)	
		Replacement of previous liver transplant (procedure)	
	Transplant of heart (procedure)	Heart transplant with recipient cardiectomy (procedure)	Heart-lung transplant with recipient cardiectomy-pneumonectomy (procedure)
		Xenotransplant of heart (procedure)	
		Allotransplant of heart (procedure)	Allotransplant of heart and lung (procedure)
			Heterotopic allotransplant of heart (procedure)
			Orthotropic allotransplant of heart (procedure)
		Autotransplant of heart (procedure)	
	Transplant of spleen (procedure)	Method: surgical transplantation attribute	
	Transplant of pancreas (procedure)	Transplant of pancreas (procedure)	Homotransplant of pancreas (procedure)
	Transplant of kidney (procedure)	Autotransplant of kidney (procedure)	
		Donor renal transplantation (procedure)	Renal homotransplant excluding donor and recipient nephrectomy (procedure)
			Live donor renal transplant (procedure)
			Xenograft renal transplant (procedure)

However, the NIS data records were stamped with the ICD–9–CM procedure codes and not SNOMED CT concept identifiers. In order to extract the transplant records from the NIS data files, the ICD–9–CM procedure codes used within the NIS records had to be mapped to SNOMED CT concept identifiers as discussed next.

4.2.1.1 ICD_9_CM to SNOMED CT mapping

The United States National Library of Medicine develop and maintain a set of files known as the Unified Medical Language System (UMLS). The purpose of these files is to establish a platform through which different biomedical vocabularies can interact. The ICD-9-CM to SNOMED CT mapping files form part of this platform. A representative of SNOMED confirmed the validity of the mapping files and hence they were included as part of the study. The ICD-9-CM to SNOMED CT mapping consists of two files, namely,

- 1 To 1, and
- 1 To Many mapping file.

The 1-To-1 maps allow for a direct translation from an ICD-9-CM code to a SNOMED concept without the loss of any meaning. However, the SNOMED concepts are more granular than the ICD-9-CM codes therefore more than one concept will map to the same code. Consequently this necessitates the need for 1-To-Many mappings. The mapping process is summarised as follows:

NIS Core File	No of records
Total no of records	8 023 590
Total no of records with procedure codes	5 064 722

ICD-9-CM File	No of records
Total no of procedure codes	3 878
Procedure codes mapped to SNOMED	2 238
– 1 To 1 map	1 753
– 1 To Many map	485
Procedure codes referenced by NIS Core records	3 274
– with SNOMED mapping	2 048
– without SNOMED mapping	1 226

NIS Core mapped to SNOMED File	No of records
NIS Core records with SNOMED mapping	4 723 812
NIS Core records without SNOMED mapping	340 910
Total:	5 064 722

As illustrated, above 93.3% of the NIS Core File records, with ICD-9-CM procedure codes, were mapped to SNOMED concept identifiers.

The following mappings were manually added

ICD–9 Code	ICD Name	SNOMED Concept ID	SNOMED Concept FSN	No of records
50.59	Other transplant of liver	174426002	Transplant of liver (procedure)	1043
33.6	Combined heart–lung transplantat	174802006	Allotransplant of heart and lung (procedure)	8
11.69	Other corneal transplant	60656008	Corneal transplant (procedure)	14
			Total	1065

Once the mapping was completed, all records with a major operating room procedure code indicator were extracted from the "NIS Core records with SNOMED mapping" file. This produced a target data set, "NIS with SNOMED mapping", of 2 163 652 observations mapped to 1 705 SNOMED concept IDs. The distribution of these observations among the unique concept IDs are as illustrated below.

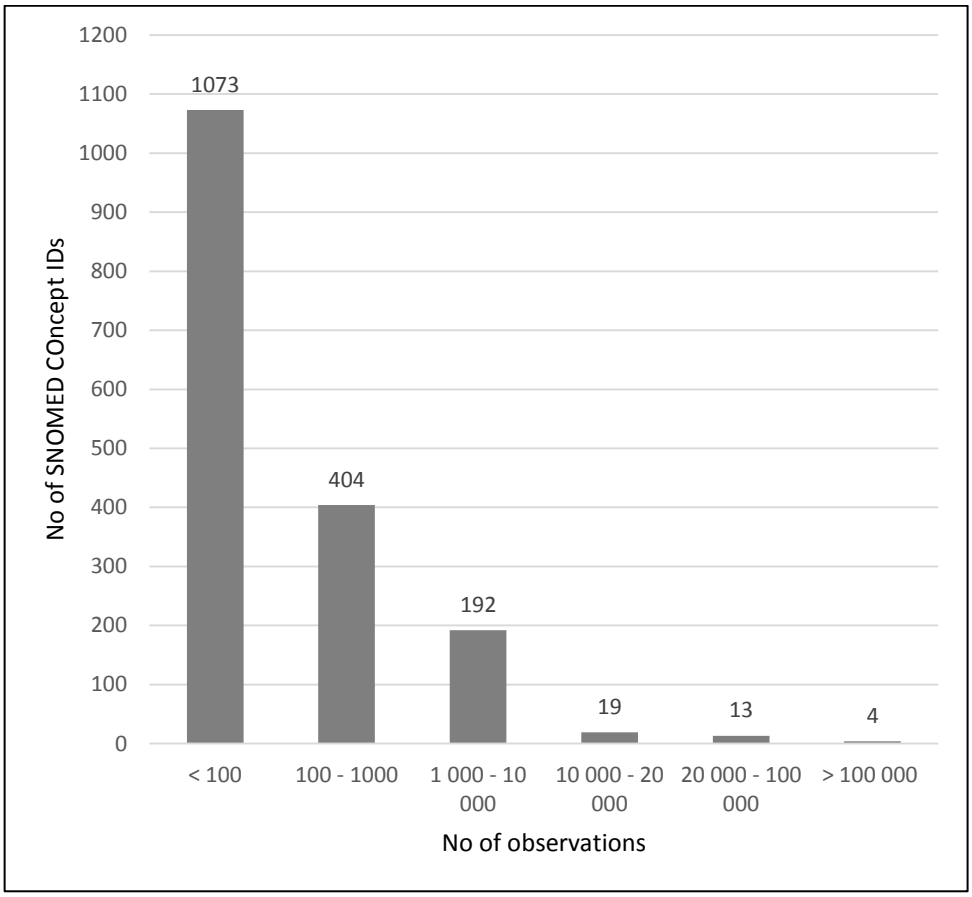

Figure 13: SNOMED Concept IDs grouped by number of observations

63% of the concept IDs have less than 100 observations, hence these major operating room procedures were sparsely populated. Taking into account the distribution of this sample set, one can conclude that densely populated concept IDs have more than 100 observations.

SAS software was used to select and describe the data sets. The target set, "NIS with SNOMED mapping" file, as well as the SNOMED CT Refset for

transplants were loaded into SAS. The following process flow was developed in SAS for extracting the transplant data — SAS Script

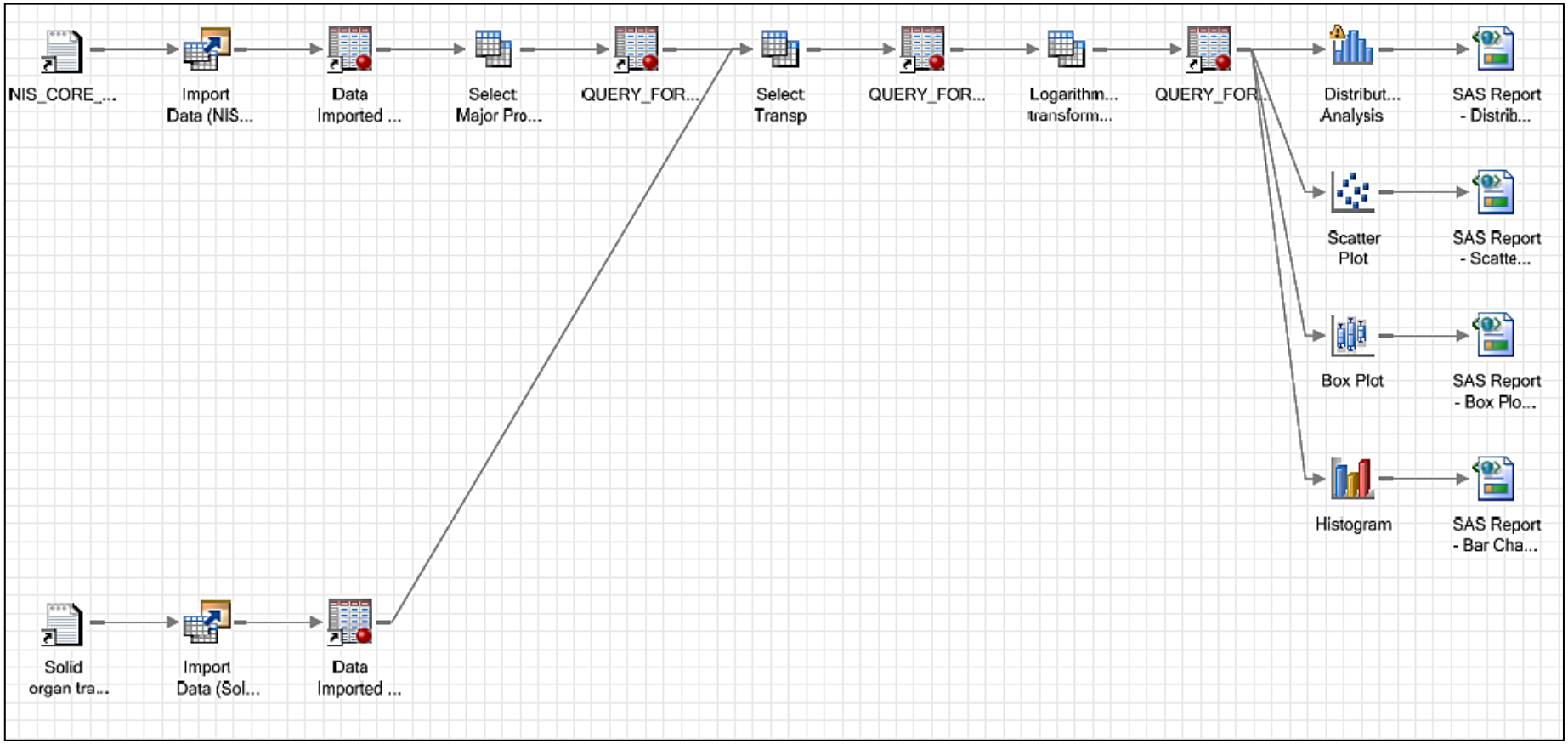

Figure 14: Data selection and description process flow

Table 6 below contains a summary of the extracted data.

Table 6: Data extraction summary

SNOMED_CID	SNOMED_FSN	Number of observations
70536003	Transplant of kidney (procedure)	3333
18027006	Transplant of liver (procedure)	1043
32413006	Transplant of heart (procedure)	409
232658009	Double lung transplant (procedure)	395
232657004	Single lung transplant (procedure)	129
62438007	Transplant of pancreas (procedure)	44
71947008	Homotransplant of pancreas (procedure)	40
174426002	Heterotopic liver transplant (procedure)	14
175899003	Autotransplant of kidney (procedure)	9
174802006	Allotransplant of heart and lung (procedure)	8
	Total	5424

The summary clearly indicates that some of the transplant procedures were sparsely populated in terms of the observation distribution amongst the procedure codes, as illustrated in Figure 13. Based on the distribution, procedure codes with an observation count of less than 100 were excluded for the purpose of this study.

The next step was to explore the selected data and construct a description of the data to be used during the knowledge discovery stage of this study.

4.2.2 Data description

Berthold in *The Guide to Intelligent Data Analysis* (Berthold, Borgelt, Hoppner, & Klawon, 2010) states that data selected for analysis should be described in terms of its distribution and correlations. There are three characteristics that summarise the distribution of a set of observations, namely:

- central tendency,
- shape, and
- variability.

For this study the central tendency of the distributions was determined by their respective mean and median values, as listed in Table 7. The “total charges” and “length of stay” attributes had the most significant influence on the total cost of each observation consequently they were used for the distribution analysis.

Table 7: Distribution summary

SNOMED_FSN		Mean	Median	Range	IQR
Transplant of kidney (procedure)	Total charges ($):	185 551	162 714	2 946 891	85 398
	Length of stay (days):	7.28	5	248	4
Transplant of liver (procedure)	Total charges ($):	471 490	304 059	3 695 379	311 004
	Length of stay (days):	21.32	12	271	16
Transplant of heart (procedure)	Total charges ($):	704 031	477 970	4 682 967	513 498
	Length of stay (days):	41.09	23	307	38
Double lung transplant (procedure)	Total charges ($):	594 304	464 794	3 033 849	423 287
	Length of stay (days):	26.55	20	164	18
Single lung transplant (procedure)	Total charges ($):	487 348	337 869	2 121 185	298 014
	Length of stay(days):	20.99	14	91	14

The mean and median values of both the “total charges” and “length of stay” attributes, of every one of the transplant procedures, were unequal. This implied an asymmetric distribution of the cost and length of stay attribute values for these observations.

A larger mean than median value as illustrated in Table 7, for both attributes, for all procedures, usually indicates a positively skewed shape. This shape was confirmed by generating frequency distributions for both the “total charges” and “length of stay” attributes for every one of the procedures. The frequency distributions were displayed as histograms and confirmed positive skewness as illustrated in Figure 15.1 to Figure 15.4. The histograms of the two most densely populated transplant procedures were included for illustration. (The frequency distributions of all the transplant procedures used for this study are included in the Appendix A.)

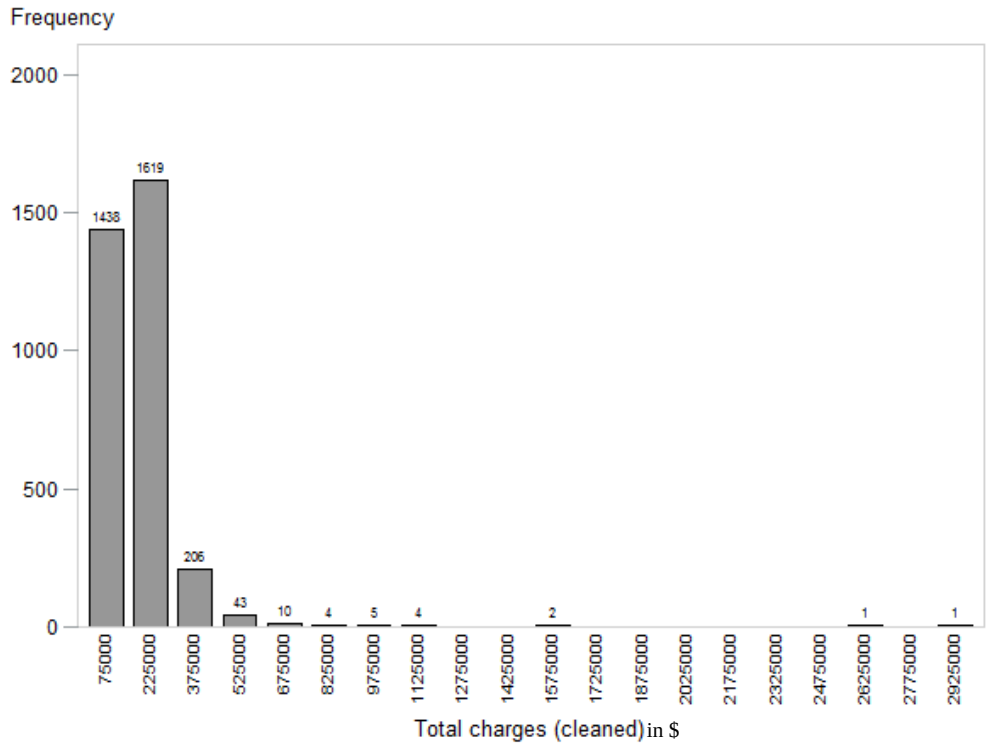

Figure 15.1: Frequency distributions for transplant of kidney procedure

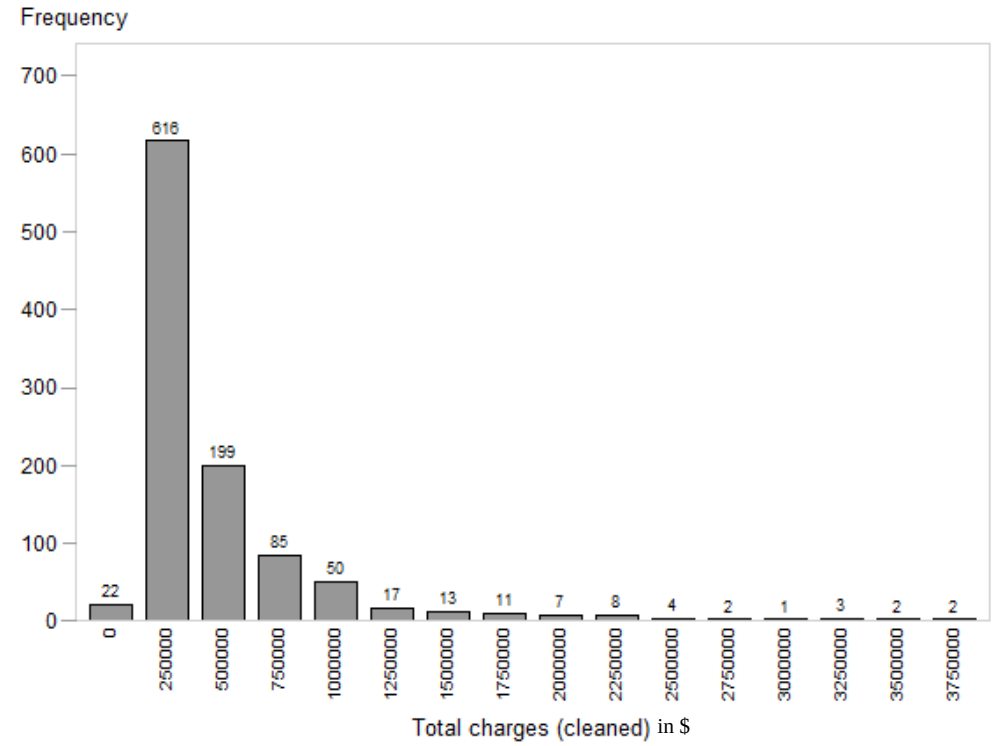

Figure 15.2: Frequency distributions for transplant of liver procedure

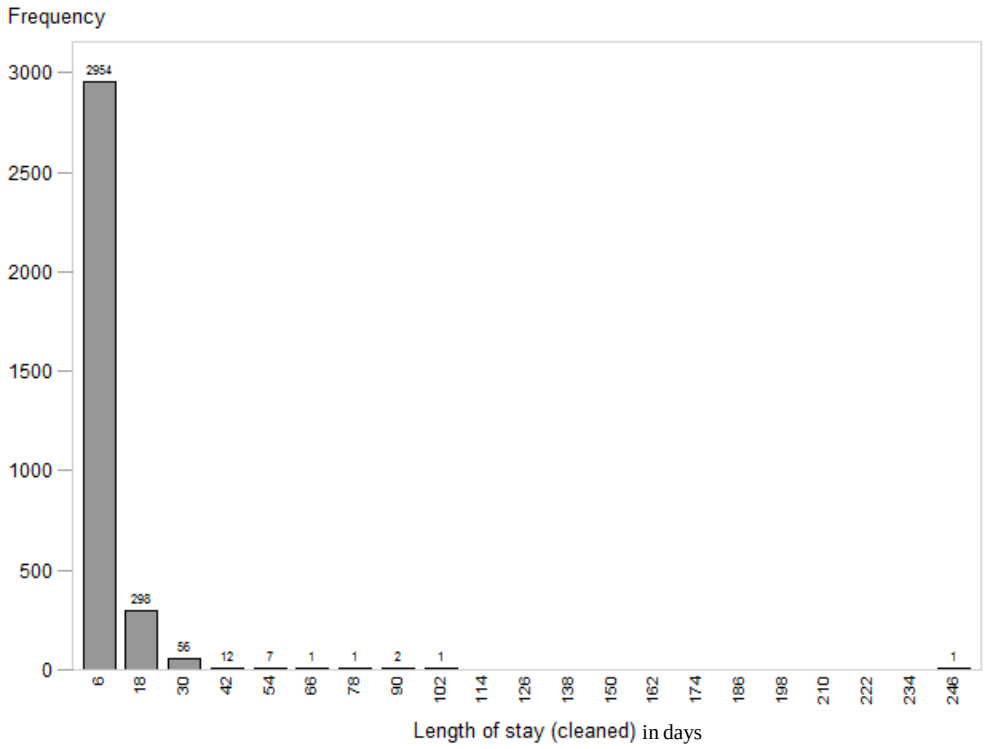

Figure 15.3: Frequency distributions for transplant of kidney procedure

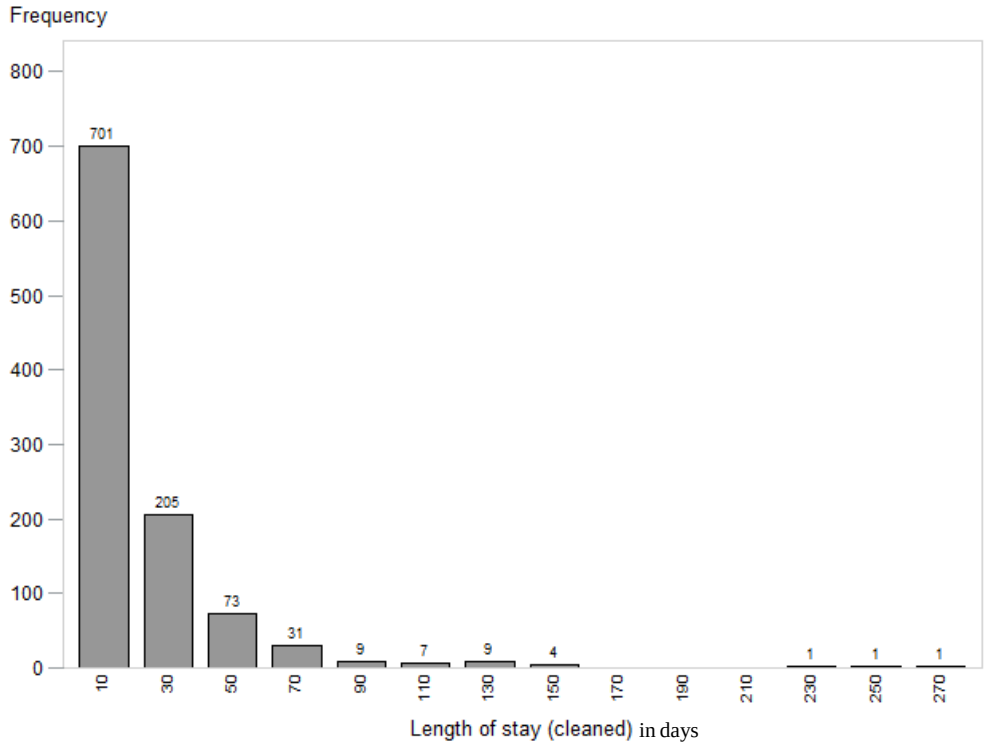

Figure 15.4: Frequency distributions for transplant of liver procedure

Lastly the distributions were described in terms of their variability i.e. how spread out the observations are. Indicators for variability are range, and inter quartile range (IQR) values as listed in Table 7. Variability however can be more effectively visualised by means of boxplots as illustrated in Figure 16.1 and Figure 16.2.

Figure 16.1 displays the "total charges" attribute boxplots for every one of the transplant procedures and reveals the following facts:

- The medians of all five procedures are closer to the bottom of the range than the top which implies that most procedure costs are towards the lower rather than higher end of the spread.
- Heart transplant procedures have the widest total cost range.
- Kidney transplant procedures have the narrowest midspread, which implies that 50% of these observations fall within the narrowest cost range compared to the other procedures.
- All five procedures have outliers which are marked by circles and lie outside 1.5 times the IQR value of the specific procedure.
- Positive skew is clearly visible in all five graphs.

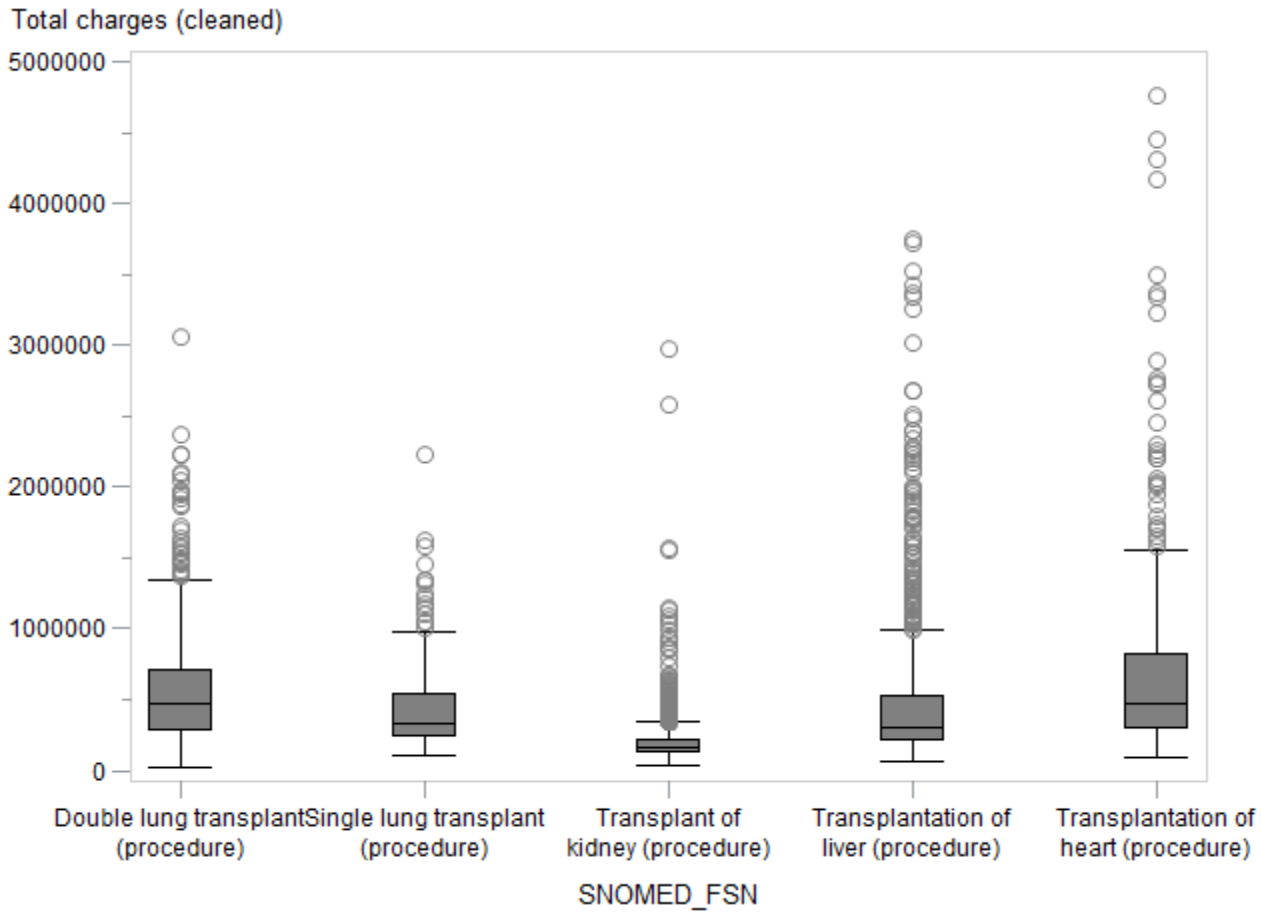

Figure 16.1: Boxplots: Total charges with outliers

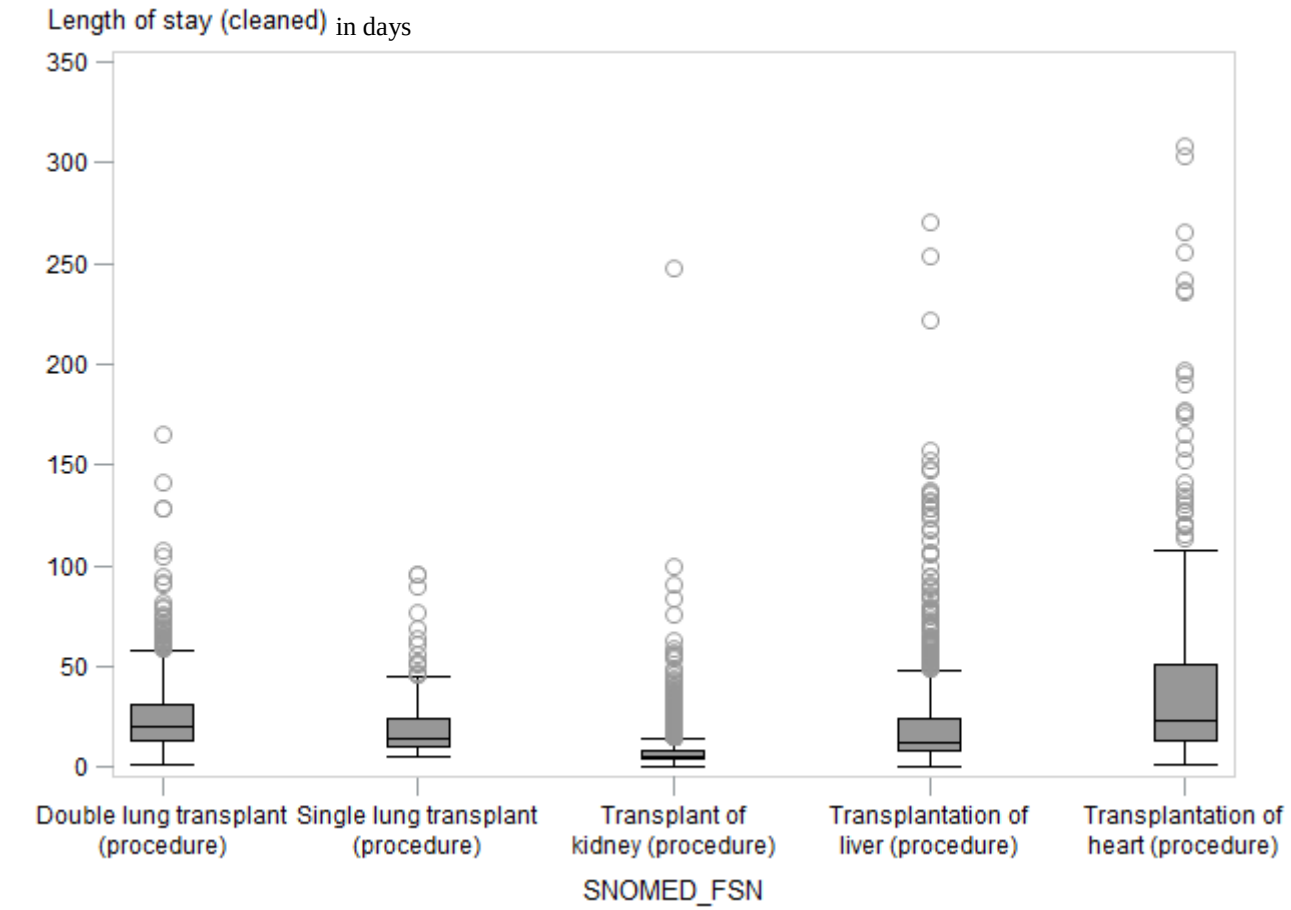

Figure 16.2 Boxplots: Length of stay with outliers

Figure 16.2 displays the “length of stay” attribute boxplots for every one of the transplant procedures. These plots are very similar to those of Figure 15.1 and 15.2. Once again, the heart transplant procedure has the widest range, the kidney transplant procedure has the smallest IQR value, and the medians are towards the lower end of the spread. All five procedures have outliers that are marked by circles and lie outside 1.5 times the IQR value of the specific procedure and positive skew is clearly visible in all five graphs. This similarity could imply a strong correlation between the values of the two attributes and was examined next.

Due to the positive skewness of the data, a log base 10 transformation was used to investigate the correlation between the two attributes, “total charges” against “length of stay”. This was done by means of scatter plots as displayed in Figure 17.1 and 17.2 (the scatter plots of all the transplant procedures used in this study are available in Appendix A).

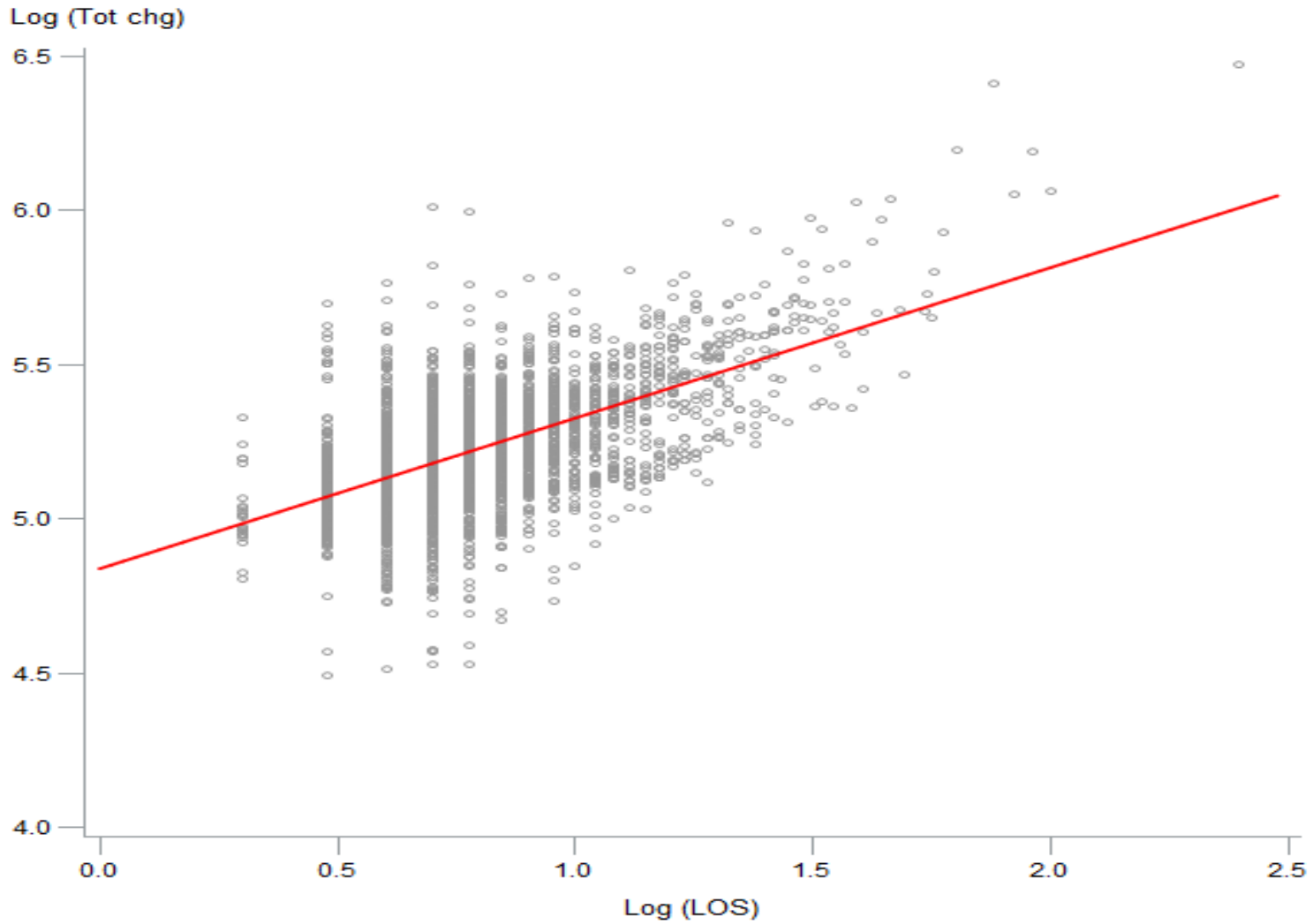

Figure 17.1: Scatter plot: Transplant of kidney (procedure)

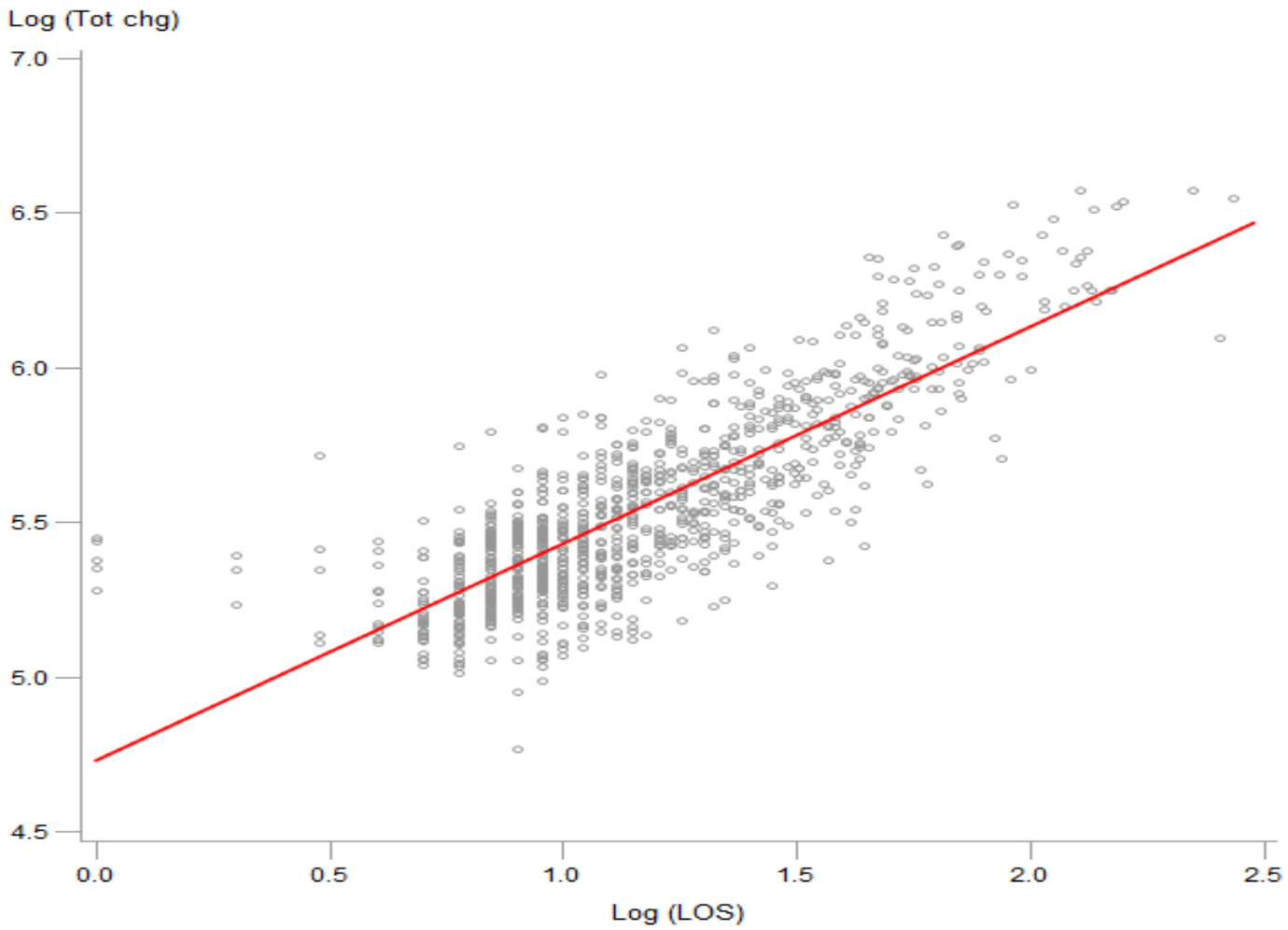

Figure 17.2: Scatter plot: Transplant of liver (procedure)

Table 8: Pearson's correlation coefficient for transplant procedures

Procedure	Double lung transplant (procedure)	Single lung transplant (procedure)	Transplant of kidney (procedure)	Transplant of liver (procedure)	Transplant of heart (procedure)
Attribute	Log (LOS)				
Log (Tot chg)	0.657	0.777	0.592	0.819	0.753

The scatter plots confirmed a strong linear association between the log of length of stay (Log (LOS)) and log of total charges (Log (Tot chg)) values. The coefficient values in Table 8 indicate that the sample of liver transplant procedures have the strongest positive correlation and that of the kidney transplant procedures the weakest.

This summarises the data understanding stage of the process model. The data understanding stage was followed by the preparation of the selected data as described in the following section.

4.3 Data preparation

The data preparation phase was divided into three steps. namely:

- feature selection,
- cleaning, and
- construction and integration.

First the features of the selected observations pertaining to this study were identified in the NIS data. Next the attribute values of these features were examined to determine the quality of the data. Based on these findings, certain decisions were made and actions were taken to address the data quality issues. This was followed by the construction of the integrated information network which comprises of the integration of the three constructed domain graph databases, namely the patient, cost and clinical graphs. The data preparation stage was an iterative process and only the summation of the final outcome of each step is presented next.

4.3.1 Feature selection

From the NIS Core data set the following analytic and demographic attributes were selected for each observation, as seen in Table 9.

Table 9: NIS Core data selected attributes

Attribute type	Attribute name	Attribute description	No missing values
Categorical	HOSPID	HCUP hospital identification	
	PATIENTID	Unique record number	
	ATYPE	Admission type	298
	ASOURCE	Admission source	4593
	FEMALE	Indicator of sex	
	DX1	Diagnosis 1 (ICD–9–CM)	
	PR1	Procedure 1 (ICD–9–CM)	
	PAY1	Primary expected payer	112
	ORPROC	Major operating room	
	DIED	Died during hospitalisation	
Numeric	LOS	Length of stay (cleaned)	1
	TOTCHG	Total charges (cleaned)	2
	AGE	Age in years at admission	
	NCHRONIC	Number of chronic conditions	

From the NIS diagnosis data set containing the observations' co-morbidities, the following indicators were selected as presented in Table: 10:

Table: 10 NIS diagnosis data selected indicators

Attribute type	Attribute name	Attribute description	Freq %
Categorical	APRDRG	All patient refined-diagnosis related group	
	Risk_Mortality	Risk of mortality (likelihood of dying)	
	APRDRG_Severity	Severity of Illness	
	CM_AIDS	Co-morbidity present: AIDS	0.2
	CM_ALCOHOL	Co-morbidity present: Alcohol abuse	4
	CM_ANEMDEF	Co-morbidity present: Deficiency anaemias	42
	CM_ARTH	Co-morbidity present: Rheumatoid arthritis/collagen vascular diseases	3
	CM_BLDLOSS	Co-morbidity present: Chronic blood loss anaemia	1
	CM_CHF	Co-morbidity present: Congestive heart failure	5
	CM_CHRNLUNG	Co-morbidity present: Chronic pulmonary disease	9
	CM_COAG	Co-morbidity present: Coagulopathy	22
	CM_DEPRESS	Co-morbidity present: Depression	8
	CM_DM	Co-morbidity present: Diabetes, uncomplicated	15
	CM_DMCX	Co-morbidity present: Diabetes with chronic complications	13
	CM_DRUG	Co-morbidity present: Drug abuse	1
	CM_HTN_C	Co-morbidity present: Hypertension (combine uncomplicated and complicated)	20
	CM_HYPOTHY	Co-morbidity present: Hypothyroidism	10
	CM_LIVER	Co-morbidity present: Liver disease	19
	CM_LYMPH	Co-morbidity present: Lymphoma	0.2
	CM_LYTES	Co-morbidity present: Fluid and electrolyte disorders	44
	CM_METS	Co-morbidity present: Metastatic cancer	0.1
	CM_NEURO	Co-morbidity present: Other neurological disorders	5
	CM_OBESE	Co-morbidity present: Obesity	12
	CM_PARA	Co-morbidity present: Paralysis	1
	CM_PERIVASC	Co-morbidity present: Peripheral vascular disorders	5
	CM_PSYCH	Co-morbidity present: Psychoses	2
	CM_PULMCIRC	Co-morbidity present: Pulmonary circulation disorders	6
	CM_RENLFAIL	Co-morbidity present: Renal failure	10
	CM_TUMOR	Co-morbidity present: Solid tumour without metastasis	5
	CM_VALVE	Co-morbidity present: Valvular disease	3
	CM_WGHTLOSS	Co-morbidity present: Weight loss	7

From the SNOMED CT ontology, the following concepts and attributes, as seen in Table 11, were selected. These participated in the defining relationships of the five transplant concepts selected for this study.

Table 11: Selected SNOMED CT concepts

	Concept hierarchy	**Concept description**
Concept	Procedure	Double lung transplant
		Kidney operation
		Renal replacement
		Single lung transplant
		Solid organ transplant
		Thorax transplant
		Transplant of kidney
		Transplant of lung
		Transplant of heart
		Transplant of liver
	Body structure	Heart structure
		Kidney structure
		Left lung structure
		Liver structure
		Lung structure
		Right lung structure
		Thoracic structure
	Substance	Heart graft - material
		Kidney graft - material
		Liver graft - material
	Qualifier value	Surgical transplantation - action
Relation	**Relation type**	**Destination concept hierarchy**
	Direct substance	Substance
	Is a	Procedure
	Method	Qualifier value
	Procedure site	Body structure

4.3.2 Data cleaning

During the data cleaning stage, the missing values were resolved by:

- Eliminating the ATYPE and ASOURCE features from the observation records.
- Setting the 112 missing PAY1 values to the mode value of "1".
- Removing one observation that had a missing LOS and TOTCHG value.
- Setting the TOTCGH value of an observation which had a missing TOTCHG value, but had a LOS value, to the average TOTCHG value of similar observations as calculated in Table 12 below.

Table 12: Missing values: TOTCHG

SNOMED FSN	Length of stay (days)	Age (years)	Total charges ($)
Transplantation of heart (procedure)	100	11	1 134 140
Transplantation of heart (procedure)	100	28	652 371
Transplantation of heart (procedure)	102	0	797 355
Transplantation of heart (procedure)	102	42	1 590 923
Transplantation of heart (procedure)	*103*	*43*	*?*
Transplantation of heart (procedure)	104	0	1 106 972
Transplantation of heart (procedure)	104	39	1 221 830
Transplantation of heart (procedure)	105	0	1 532 601
		Average	**1 148 027**

The CM_ULCER co-morbidity was removed from the selected features as it was too scarcely populated and had a frequency of only one observation, as indicated in Table: 10.

Where required, during the data cleaning stage, attribute types were transformed from numerical to categorical values. This involved the following selected features namely: age, cost, length of stay and the number of chronic diseases, as discussed next.

The age feature was changed from a numerical to a categorical attribute type. The age groups, as defined by HCUP's summary statistics, were used to categorise the observations. These were 0 – 17; 18 – 44; 45 – 64 and 65+; however, for the purpose of this study, year 0 (birth) was put into its own category as it had such an exceptional high cost value per observation, as displayed in the Figure 18.

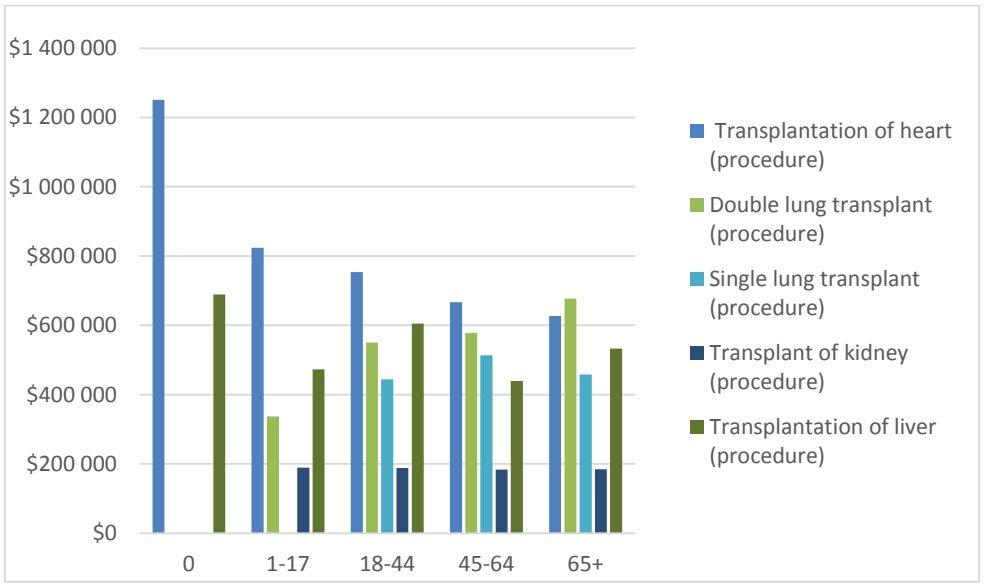

Figure 18: Avg total charges (per procedure) within age categories

Two additional categorical attributes were added to the data, namely, cost and length of stay (LOS) indicators. The cost category for each observation indicates to which cost group the observation belongs; similarly

the length of stay (LOS) category indicates to which length of stay group the observation belongs. For each indicator the following five categories were defined:

Category	Description
1	Below
2	Middle Half
3	Above
4	Outlier
5	Anomaly

The boxplots generated for each procedure type, as illustrated in Figure 18, were used to derive the category boundaries for each indicator, procedure type combination. The upper inner fence (whiskers) and upper outer fence values were determined by adding 1.5 * IQR and 2 * IQR respectively to each upper hinge value. Hence for each indicator, procedure type combination, the category boundaries were determined individually.

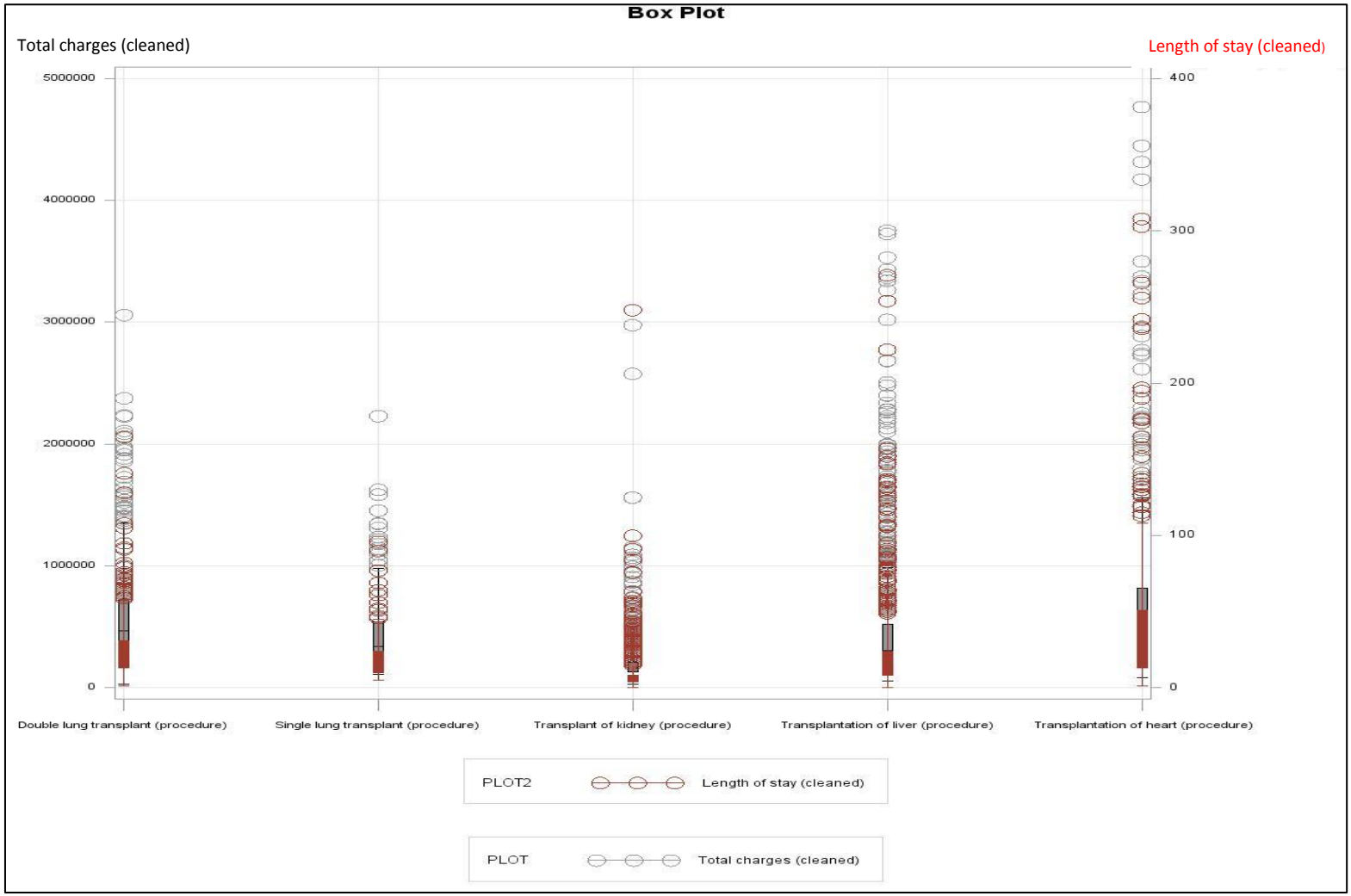

Figure 19: Cost and LOS per procedure type (SAS)

The lower hinge values of each boxplot, pertaining to the indicator, procedure type combination were used as the upper boundary values for Category 1. Similarly, the lower and upper hinge values determined the boundaries for Category 2, the upper hinge and upper inner fence values determined the boundaries for Category 3, the upper inner fence and upper outer fence values determined the boundaries for Category 4 and lastly the Category 5 lower boundary values were determined by the upper outer fence values. The distribution of the observations within the cost and LOS categories are illustrated in Figure 20 and Figure 21.

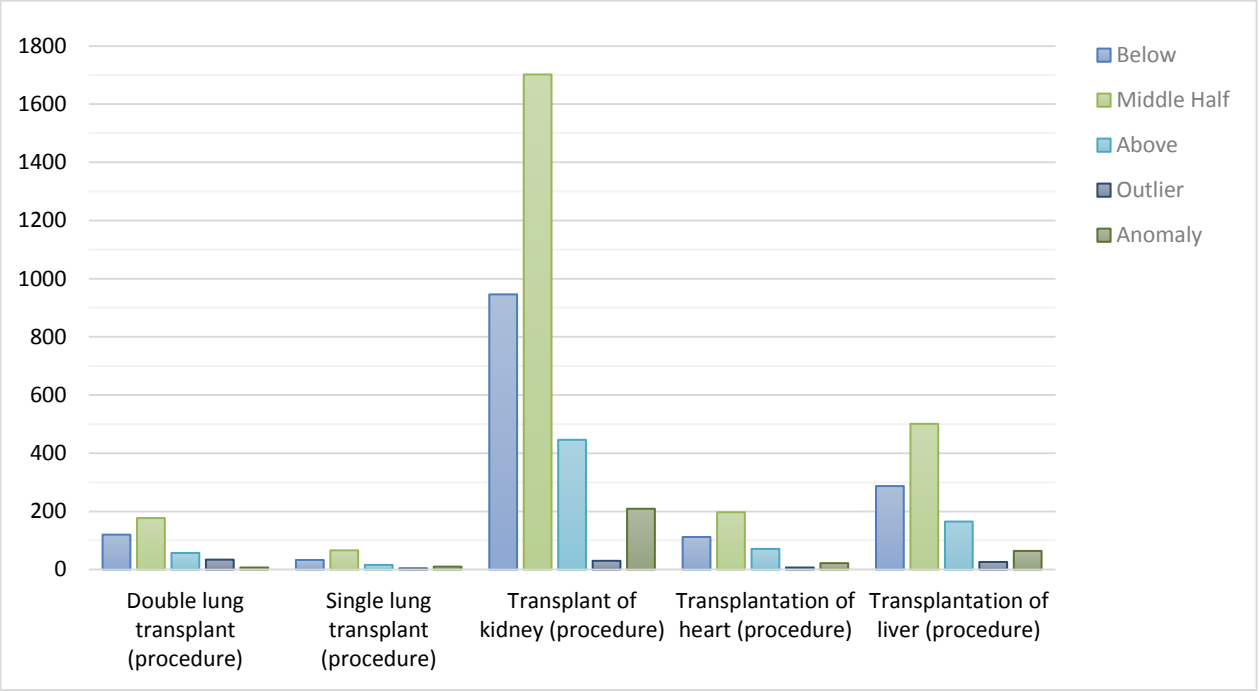

Figure 20: LOS category distribution per procedure type

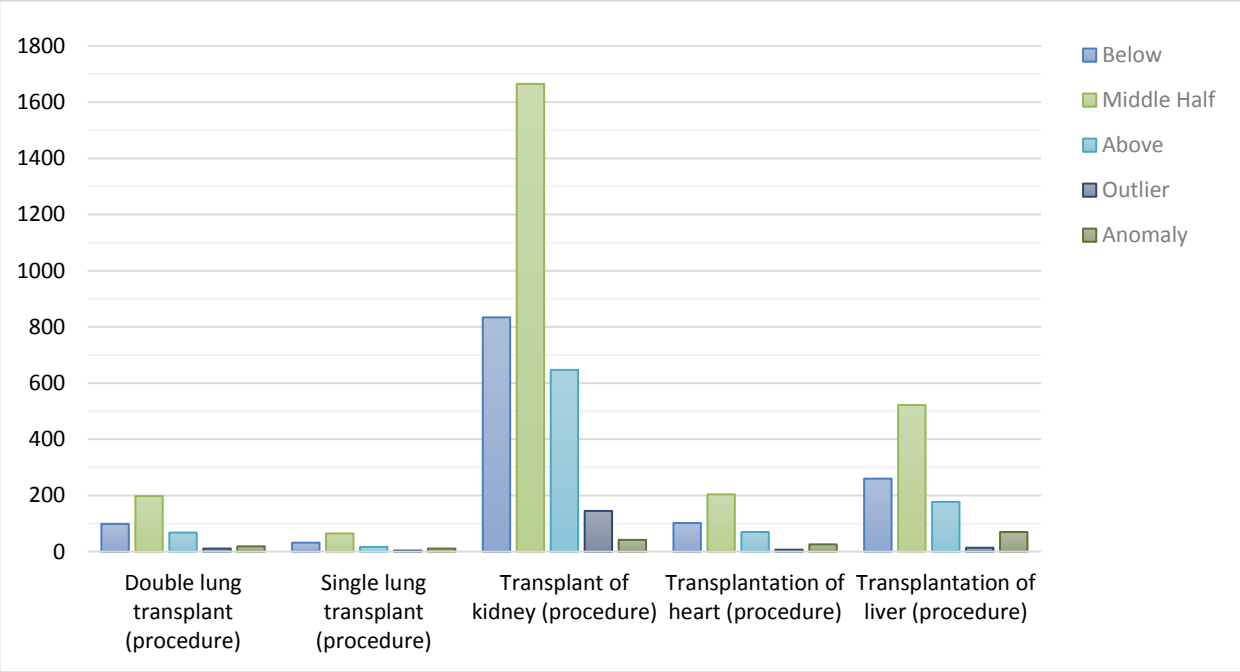

Figure 21: Cost category distribution per procedure type

The cleaned data was then used to construct the three domains as discussed next.

4.3.3 Domain construction

The data was grouped into three domains. A cost domain described the procedure a patient underwent, the cost of the procedure, as well as the length of stay of the patient in the hospital. A patient domain described the demographics of each observation including the patient's age, gender, the hospital they visited, the person responsible for the payment of their hospital bill, as well as the outcome of the procedure. Lastly, a clinical domain which described the clinical diagnosis of the patient, the severity of the diagnosis, the likelihood of the patient to die, as well as the co-morbidities present within the patient.

In order to construct the domains as graph databases, they were first modelled in GRAD, a graph database modelling language, introduced in Section 3.4.1. The graph database model of the cost domain is presented in Figure 22.

$\mathcal{C}$ = { $\mathcal{C}$, $\mathcal{L}_e$,, Λ_v, Λ_e,V_e, E_{as}, E_g}

where

$\mathcal{C}$ = {Observation, Procedure, Body Structure, Substance, Qualifier Value, Cost, LOS}

$\mathcal{L}_e$ = {lengthOfStay, cost, procedureSite, directSubstance, method, isA, underwent}

Λ_v = {PATIENTID, SNOMED CID, CATEGORY}

Λ_e = {LOS, TOTCHG}

V_e = {V_1, V_2, V_3, V_4, V_5, V_6, V_7} and

V_1 = <Observation, PATIENTID>

V_2 = <Procedure, SNOMED CID >

V_3 = <Body Structure, SNOMED CID>

V_4 = <Substance, SNOMED CID>

V_5 = <Qualifier Value, SNOMED CID>

V_6 = <Cost, CATEGORY>

V_7 = <LOS, CATEGORY>

$\mathcal{C}$ = Set of classes

$\mathcal{L}_e$ = Set of labels on the entity edges

Λ_v = Set of entity node identifiers

Λ_e = Set of edge attributes

V_e = Set of entity nodes described by the pair < $\mathcal{C}_i$,ID_i>,

where $\mathcal{C}_i$ is the class the entity belongs to and ID_i is the unique node identifier

E_{as} = Set of association edges

E_g = Set of generalisation edges

E_{as} = {E_1, E_2, E_3, E_4, E_5, E_7} and

E_g = {E_6} where

E_1 = <Observation, LOS, Association, lengthOfStay, {LOS}>

E_2 = <Observation, Cost, Association, cost, {TOTCHG}>

E_3 = <Procedure, Body Structure, Association, procedureSite, - >

E_4 = <Procedure, Substance, Association, directSubstance, - >

E_5 = <Procedure, Qualifier Value, Association, method, - >

E_6 = <Procedure, Procedure, Generalization, isA, - >

E_7 = <Observation, Procedure, Association, underwent, - >

Figure 22: Cost domain model

Next, the $\mathcal{G}_{cost}$ graph database model with the selected data was used to construct the cost domain graph database in Neo4J (Robinson *et al*., 2015). Neo4J (Robinson *et al*., 2015) is an open source graph database management system with a NOSQL data store as introduced in Section 3.4.3. A subgraph

of the constructed cost domain graph database illustrating the cost of heart transplant procedures is presented in Figure 23.

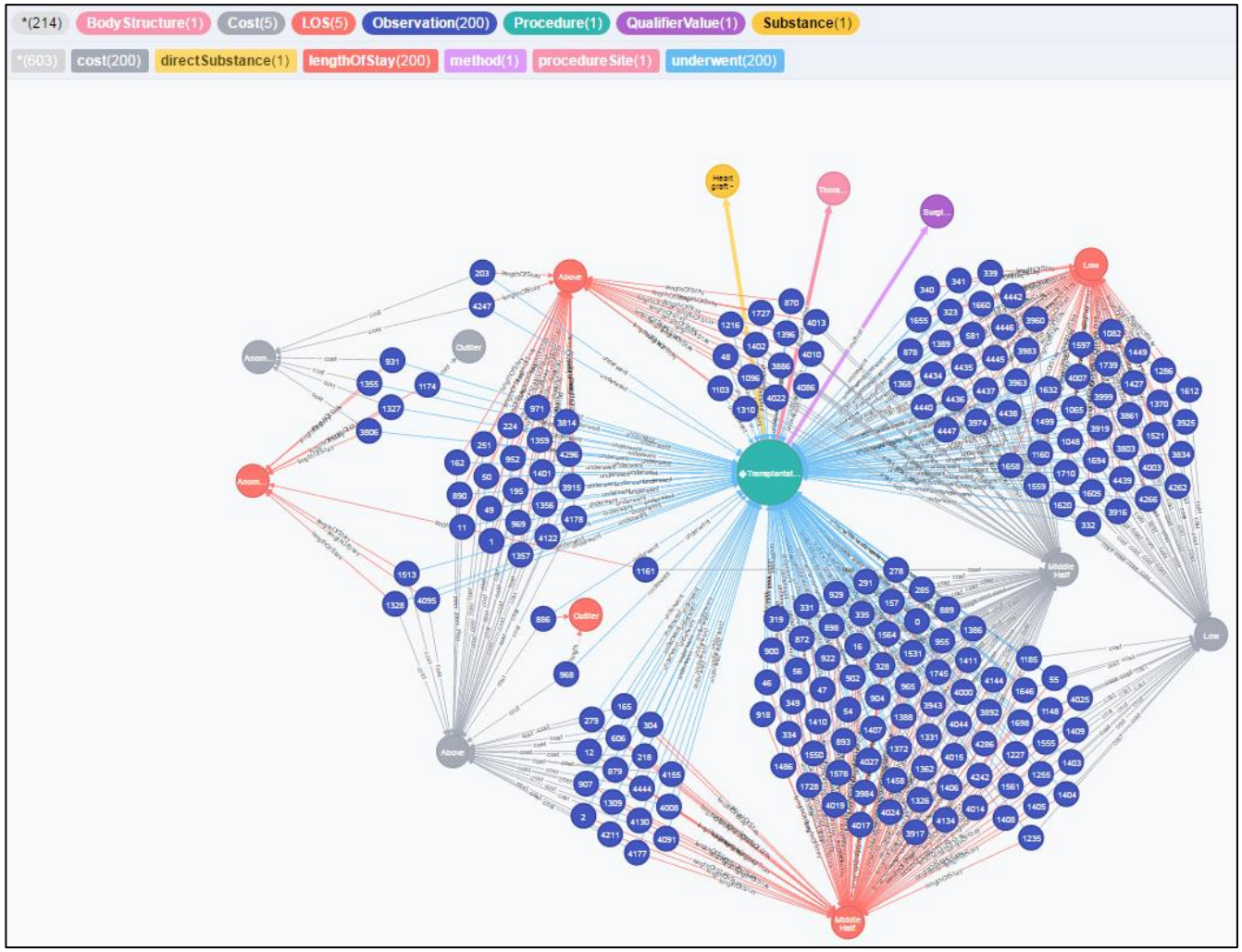

Figure 23: Subgraph of the cost domain: heart transplants (Neo4J)

The second domain to be constructed was the patient domain. The same steps were followed. First, the patient domain was modelled in GRAD. Then the patient domain was constructed, according to the $\mathcal{G}_{patient}$ model, in Neo4J (Robinson *et al*., 2015) with the patient domain data. The patient domain model is presented in Figure 24 and a subgraph of the patient domain, constructed in Neo4J, which includes observations from the Keck Hospital of USC, is illustrated in Figure 25.

$\mathcal{G}_{patient}$ = { $\mathcal{C}$, $\mathcal{L}_e$,, Λ_v,V_e, E_{as}, E_g}

where

$\mathcal{C}$ = {Observation, Gender, Age, Payer, Hospital, Mortality}

$\mathcal{L}_e$ = {ofGender, ofAge, payedBye, visits, died}

Λ_v = {PATIENTID, SNOMED CID, FEMALE, AGE, PAYID, HOSPID, DIED}

V_e = {V_1, V_2, V_3, V_4, V_5, V_6} and

V_1 = <Observation, PATIENTID>

V_2 = <Gender, FEMALE>

V_3 = <Age, AGE>

V_4 = <Payer, PAYID>

V_5 = <Hospital, HOSPID>

V_6 = <Mortality, DIED>

$\mathcal{C}$ = Set of classes

$\mathcal{L}_e$ = Set of labels on the entity edges

Λ_v = Set of entity node identifiers

Λ_e = Set of edge attributes

V_e = Set of entity nodes described by the pair < $\mathcal{C}_i$,ID_i>,

where $\mathcal{C}_i$ is the class the entity belongs to and ID_i is the unique node identifier

E_{as} = {E_1, E_2, E_3, E_4, E_5}

where

E_1 = <Observation, Gender, Association, ofGender,->

E_2 = <Observation, Age, Association, ofAge,- >

E_3 = <Observation, Payer, Association, payedBy,->

E_4 = <Observation, Hospital, Association, visits, - >

E_5 = <Observation, Mortality, Association, died,- >

Figure 24: Patient domain model

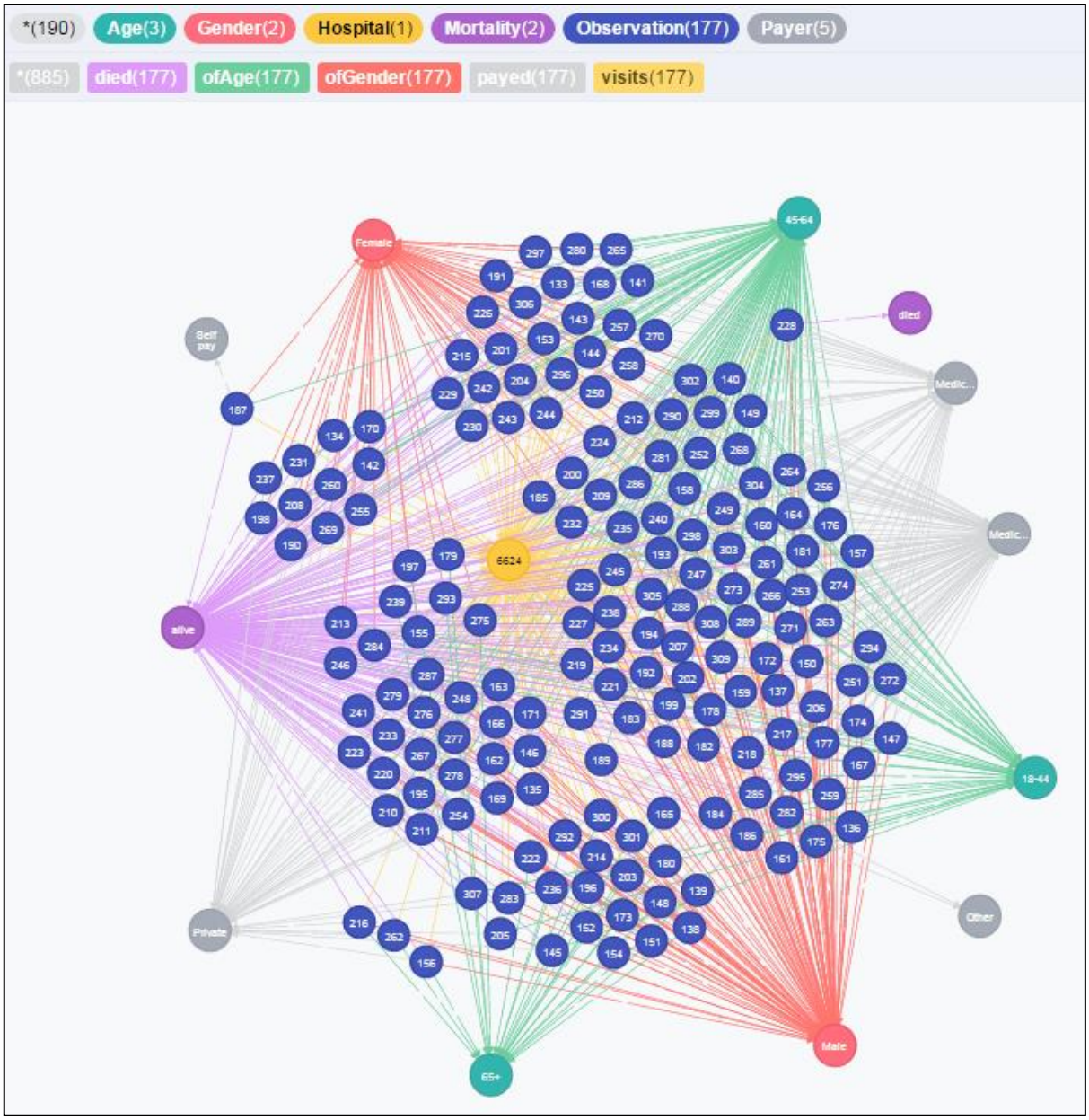

Figure 25: Subgraph of the patient domain: Keck Hospital of USC (Neo4J)

Lastly, the third domain, the clinical domain, was modelled and constructed in a similar way. The clinical domain model is presented in Figure 26 and constructed according to the $\mathcal{G}_{clinical}$ model in Neo4J. A subgraph of the clinical domain including heart and lung transplant observations is illustrated in Figure 27.

$\mathcal{G}_{clinical}$ = { $\mathcal{C}$, $\mathcal{L}_e$,, Λ_v,V_e, E_{as}, E_g} where

$\mathcal{C}$ = {Observation, NChronic, DiagGroup, Diagnosis, MortalRisk, Severity, Comorbidity}

$\mathcal{L}_e$ = {noCond, diagnosedWith, inGroup, riskDying, severeness, sufferFrom}

Λ_v = {PATIENTID, NCHRONIC, APRDRG, DX1, MORTALRISK, SEVERITY, COMORBIDITY}

V_e = {V_1, V_2, V_3, V_4, V_5, V_6} and

V_1 = <Observation, PATIENTID >

V_2 = < NChronic, NCHRONIC >

V_3 = < DiagGroup, APRDRG >

V_4 = < Diagnosis, DX1>

V_5 = < MortalRisk, MORTALRISK >

V_6 = < Severity, SEVERITY >

V_7 = < Comorbidity, COMORBIDITY >

$\mathcal{C}$ = Set of classes

$\mathcal{L}_e$ = Set of labels on the entity edges

Λ_v = Set of entity node identifiers

Λ_e = Set of edge attributes

V_e = Set of entity nodes described by the pair < $\mathcal{C}_i$,ID_i>,

where $\mathcal{C}_i$ is the class the entity belongs to and ID_i is the unique node identifier

E_{as} = {E_1, E_2, E_3, E_4, E_5, E_6} where

E_1 = <Observation, NChronic, Association, noCond,->

E_2 = <Observation, Diagnosis, Association, diagnosedWith,->

E_3 = <Observation, DiagGroup, Association, inGroup, - >

E_4 = <Observation, MortalRisk, Association, riskDying, - >

E_5 = <Observation, Severity, Association, severeness, - >

E_6 = < Observation, Comorbidity, Association, sufferFrom, - >

Figure 26: Clinical domain model

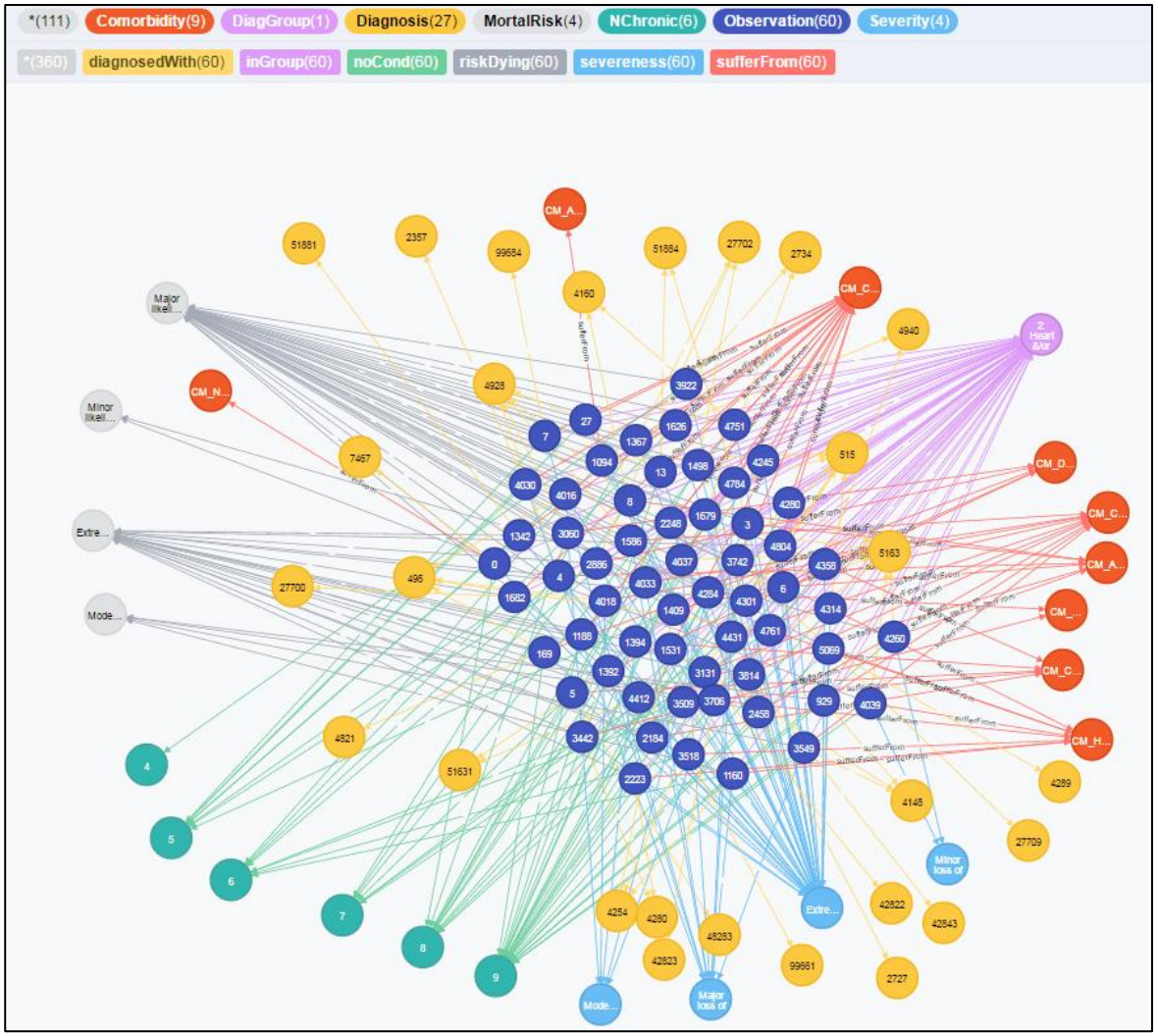

Figure 27: Subgraph of the clinical domain: heart and lung transplants (Neo4J)

On account of the complexity of the healthcare sector, it lends itself to be modelled as a semi-structured, heterogeneous information network, as stated in Section 2.3.2. Recall that the function of a network schema is to provide a meta level description of a network in a similar way that metadata provides structured information regarding a particular data source. Therefore, the next step was to extract a network schema from each of the three domain models. This entailed converting the graph data model into a heterogeneous information network, followed by the extraction of its schema as explained next.

The cost domain graph data model, $\mathcal{G}_{\text{cost}}$, was converted into the following information network:

$G_{cost} = (V, E)$ with an object type mapping function $\tau: V \rightarrow A$ and a relation type mapping function $\emptyset: E \rightarrow R$, where each information unit $v \in V$ belongs to one particular object type $\tau(v)\ \epsilon\ A$ and each relation $e\ \in E$ belongs to a particular relation type $\emptyset(e) \in R$, where

A = {Observation, Procedure, Body Structure, Substance, Qualifier Value, Cost, LOS} and

R = {lengthOfStay, cost, procedureSite, directSubstance, method, isA, underwent}

From this, the network schema denoted as $T_{Gcost} = (A,R)$, as illustrated in Figure 28, was derived and depicted in a k-partite graph structure presented in Figure 29.

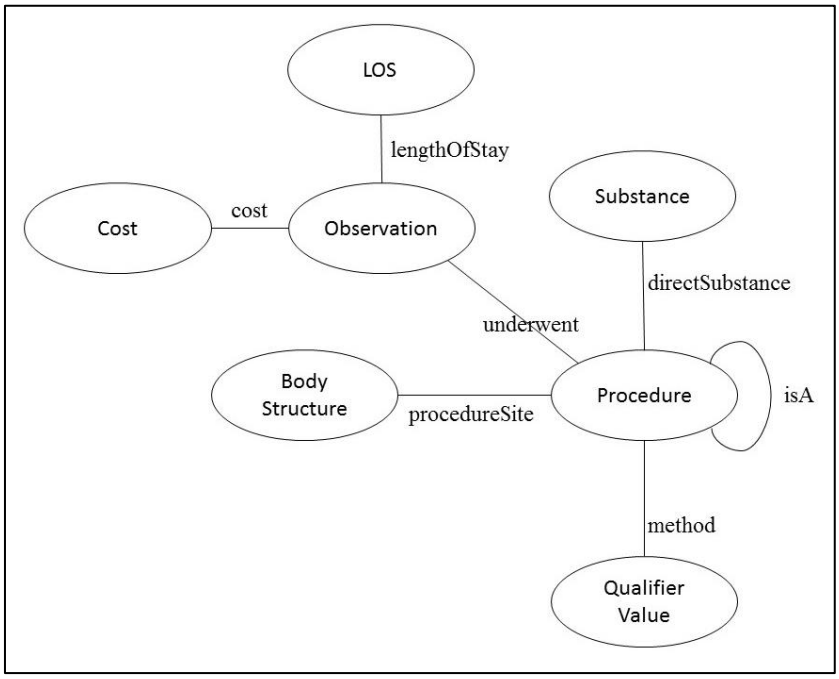

Figure 28: Cost network schema

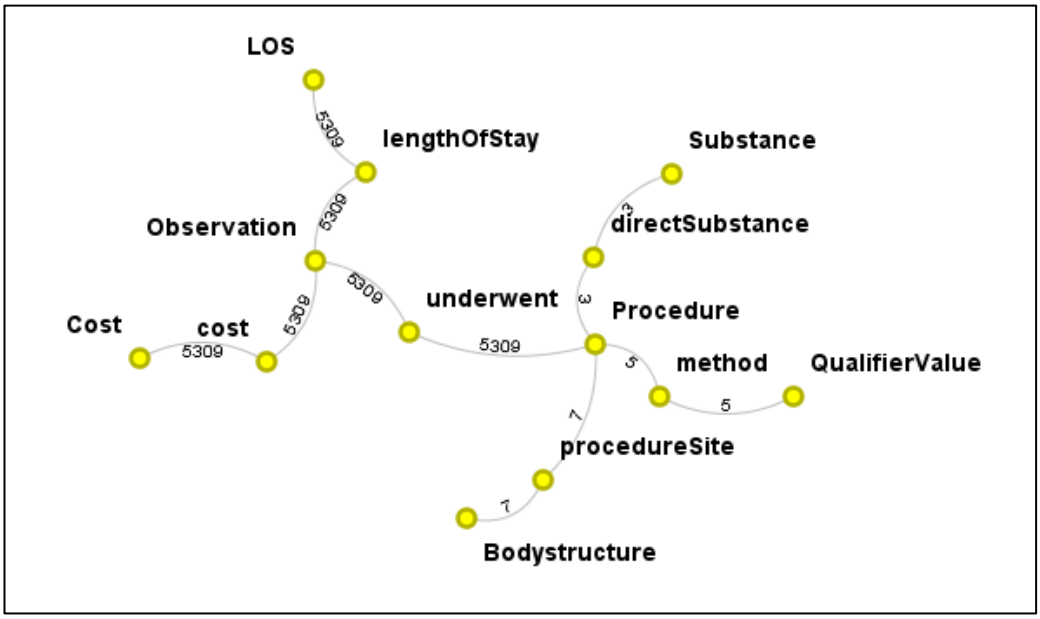

Figure 29: Cost k-partite graph structure

In a similar way, the patient and clinical domains were converted into the following two information networks:

$G_{clinical} = (V, E)$ with an object type mapping function $\tau: V \rightarrow A$ and a relation type mapping function $\emptyset: E \rightarrow R$, where each information unit $v \in V$ belongs to one particular object type $\tau(v)\ \epsilon\ A$ and each relation $e\ \in E$ belongs to a particular relation type $\emptyset(e) \in R$, where

A = {Observation, NChronic, DiagGroup, Diagnosis, MortalRisk, Severity, Comorbidity}

and

R = {noCond, diagnosedWith, inGroup, riskDying, severeness, sufferFrom}

$G_{patient} = (V, E)$ with an object type mapping function $\tau: V \rightarrow A$ and a relation type mapping function $\emptyset: E \rightarrow R$, where each information unit $v \in V$ belongs to one particular object type $\tau(v)\ \epsilon\ A$ and each relation $e\ \in E$ belongs to a particular relation type $\emptyset(e) \in R$, where

A= {Observation, Gender, Age, Payer, Hospital, Mortality}

and

R= {ofGender, ofAge, payedBye, visits, died}

From this, the network schema denoted as $T_{Gpatient}$ = (A,R) depicted in Figure 30 was derived and presented in the bi-partite graph structure in Figure 31.

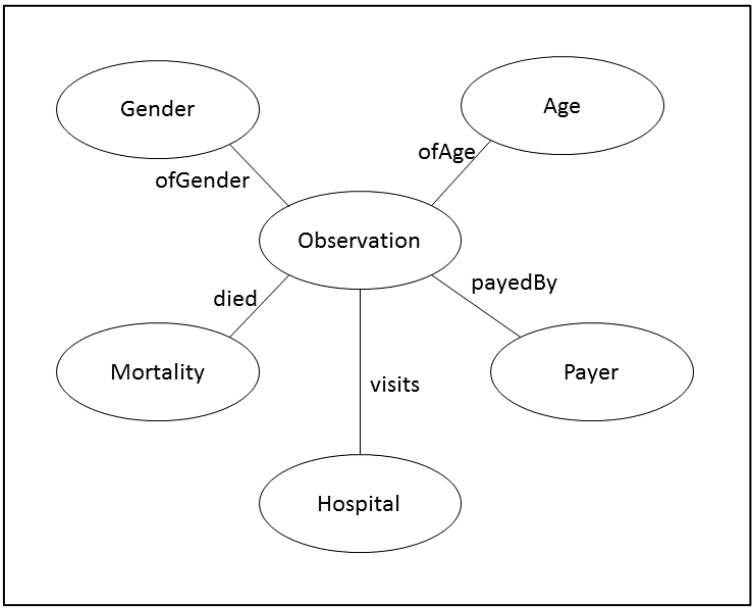

Figure 30: Patient network schema

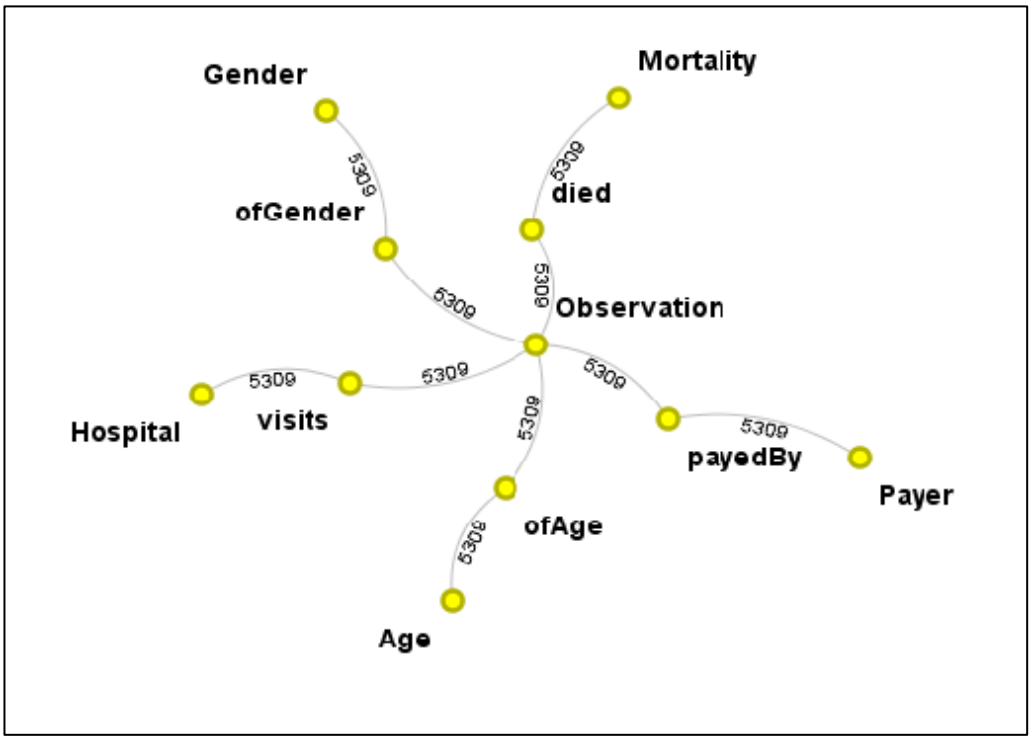

Figure 31 : Patient bi-partite graph structure

The network schema $T_{Gclinical} = (A,R)$ depicted in Figure 32 was also derived and presented in the bi-partite graph structure in Figure 33.

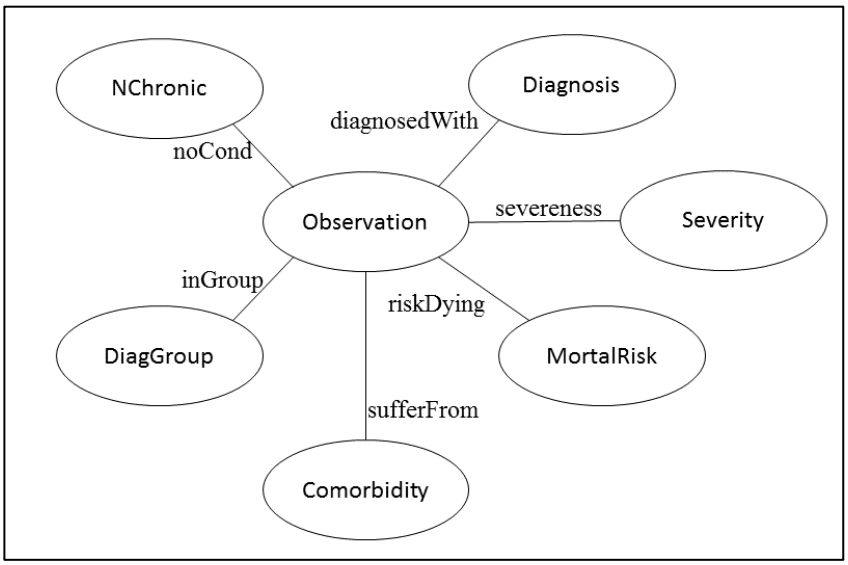

Figure 32: Clinical network schema

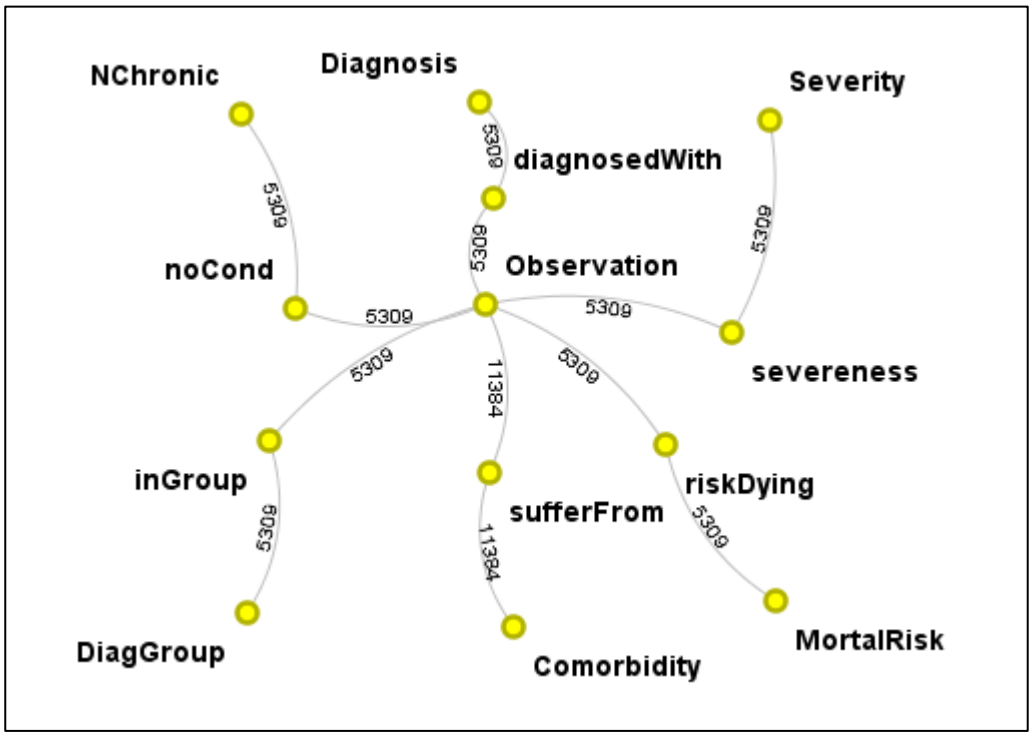

Figure 33 : Clinical bi-partite graph structure

The final step in the data preparation stage was the integration of the three domain networks into an integrated, heterogeneous information network (IHIN) as presented next.

4.3.4 Network integration

For the purpose of this study, KNIME (Berthold *et al.*, 2009), an open source data driven analytical platform, that was developed at the University of Konstanz, was used. KNIME, which stands for Konstanz Information Miner, was developed by a team of developers from a Silicon Valley software company and released for the first time in 2006. However, since its inception in 2006, its use and capabilities have grown exponentially and now supports a large user community, many of whom contribute extensions on a regular basis to the analytics platform. The KNIME workbench allows the user to create workflows. A workflow is a sequence of single processing units called nodes, linked together, that manipulate, analyse or visualise data.

During the domain construction step, which is part of the data preparation stage, a domain generation workflow was developed for each of the three domains. The three workflows were then combined into one network integration workflow presented in Figure 34. When executed, the network integration workflow constructs the integrated, heterogeneous information network (IHIN) of which the outcome is illustrated in Figure 35.

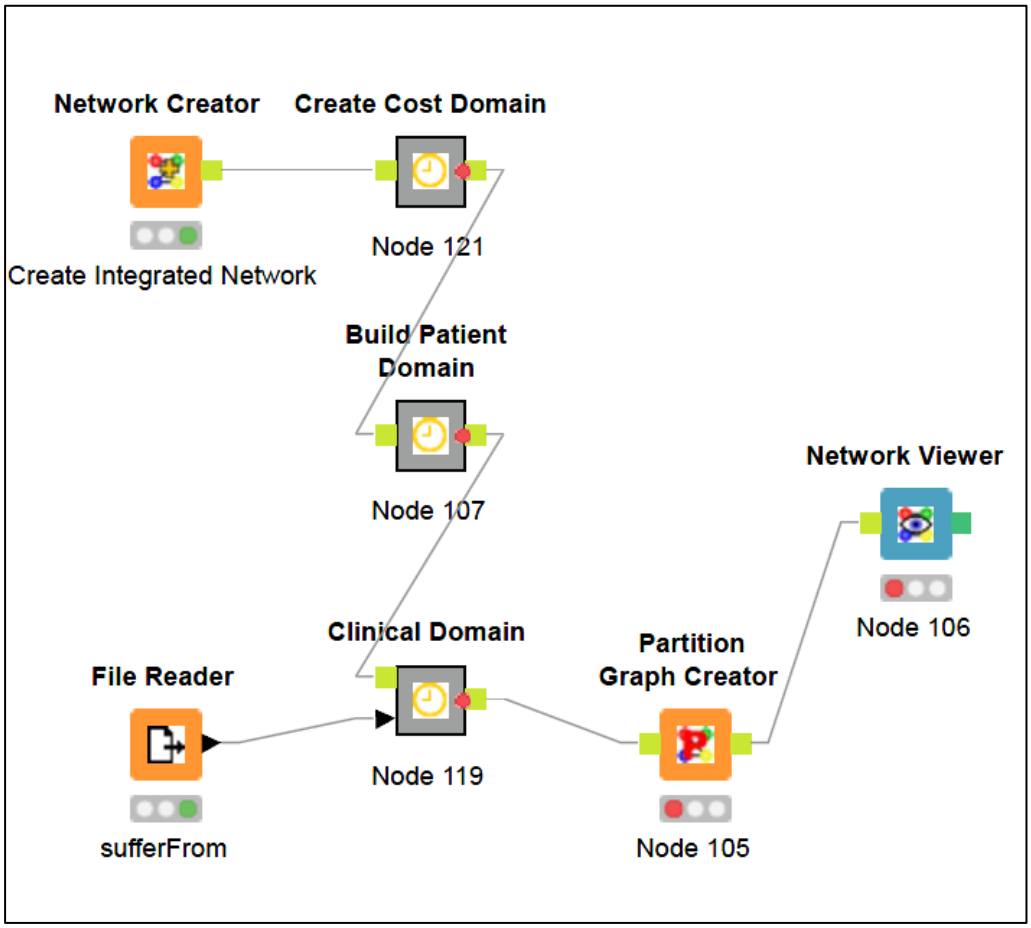

Figure 34: Network integration workflow (KNIME)

The IHIN can be described as follows:

B(V, E) is a BisoNet that contains all information with *V* representing the vertices and *E* the edges. The edge (u, v) represents an edge connecting the two vertices $u, v \in V$.

Where

V= {Observation, Gender, Age, Payer, Hospital, Mortality, Procedure, Body Structure, Substance, Qualifier Value, Cost, LOS, Nchronic, DiagGroup, Diagnosis, MortalRisk, Severity, Comorbidity}

and

E= {ofGender, ofAge, payedBye, visits, died, lengthOfStay, cost, procedureSite, directSubstance, method, isA, underwent, noCond, diagnosedWith, inGroup, riskDying, severeness, sufferFrom}

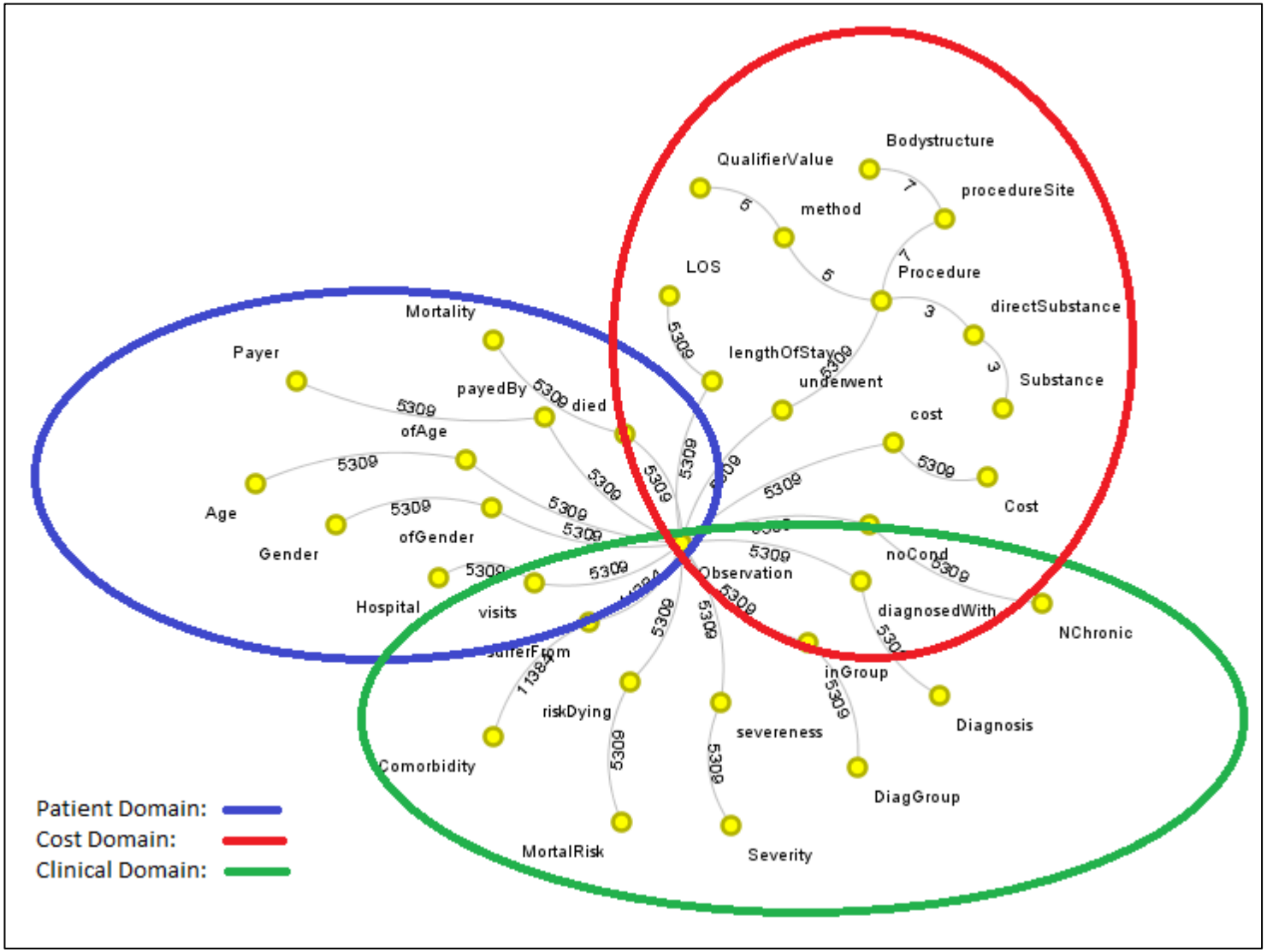

Figure 35: Integrated, Heterogeneous Information Network (IHIN)

The IHIN complies with the definition of a BisoNet as the information network was constructed by integrating three diverse domains into a unified framework. From Figure 35 it is noticeable that firstly, the vertices are loosely coupled; secondly, each vertex belongs to one and only one specific domain which makes these domains habitually incompatible. Lastly, the BisoNet does not store any other additional data, except for its source reference i.e. the domain it originates from.

It is important to note that the methodology proposed in this study strives towards computationally reproducible research. In Section 3.4 the importance of computationally reproducible research, especially within the field of data-intensive science, is discussed. The three guidelines to produce reproducible research that were introduced are

- the automation of the research method,
- provenance tracking, and
- the availability of data and software.

The automation aspect emphasises the importance of automating the individual steps that were executed for processing the data. This was accomplished by creating automated workflows in both SAS and KNIME. SAS process-flows, as illustrated in Figure 14, were created to automate the data selection and description steps that form part of the data understanding phase. KNIME workflows, as illustrated in Figure 34, were created to automate the data preparation stage combined with the frequent pattern mining step of the data mining phase.

Provenance tracking necessitates the documentation of the platforms used for the research. For the purpose of this study, the workbench consisted of SAS version 9.3, KNIME version 3.2.1 and Neo4J version 3.0.6 on a Windows 7 Professional operating system.

Lastly, the use of public available data and software is of high importance to make the research easily reproducible. The data used for the

purpose of this study was Nationwide Inpatient Sample (NIS) data, from the 2014 release, which forms part of the Healthcare Cost and Utilization Project (HCUP). Also used was the SNOMED CS ontology that is downloadable free of charge for academic use. Both KNIME and Neo4J are open source software which is downloadable via the internet, also free of charge.

The BisoNet constructed during this data preparation stage of the DMKD process model served as the input to the data mining stage as presented next.

4.4 Data mining

As discussed in Section 3.4.2 the method chosen for the extraction of information from the BisoNet was by means of concept graph detection. Because concept graphs are subgraphs of the BisoNet, the mining stage can be described as a clustering task of specifically a heterogeneous information network. The clustering was accomplished by means of frequent pattern mining, and the chosen machine learning algorithm was the Jaccard Itemset Mining (JIM) algorithm, introduced in Chapter 2.

4.4.1 Frequent pattern mining

The JIM algorithm is an extended form of the Eclat algorithm (Zaki *et al*., 1997) and requires a transaction database as input. The transaction database was generated by extracting an adjacency list for each of the subgraphs. An adjacency list converts an information network into a table format by listing each information unit with the set of its neighbouring information units. For the purpose of this study, each concept member was represented by an observation; as a result, a subset of the adjacency list which included only those subgraphs belonging to the observations was used, as illustrated in Table 13.

Table 13: Adjacency list of single lung transplant observations

Observation	Neighbours														
362011102120819	232657004	45-64	12	CM_HTN_C	F	CM_COAG	P3	N8	H36336	Function-Extreme	5163	Live	2	APR DRG-2	CM_DM	Dying-Major
402011100446346	232657004	CM_HTN_C	Function - Moderate	CM_DEPRESS	4928	CM_DM	P1	12	Dying-Major	Live	H40038	3	45-64	APR DRG-2	N8	F
362011102156524	232657004	N7	H36336	F	1	Function-Major	P1	12	Dying-Major	45-64	51883	CM_CHRNLUNG	CM_PULMCIRC	Live	APR DRG-2	CM_HTN_C

The adjacency lists of the observations belonging to each of the five selected transplant procedures served as the input item list for the JIM algorithm. For the first iteration only the anomalies, based on the cost of each observation, was included in the input item list. The goal of this frequent pattern mining algorithm was to detect observations, within the input item list, that share the same neighbours. The execution of the algorithm was controlled by the following four input variables:

- the minimum Jaccard index
- minimal support
- minimum item set size and
- target type

The minimum item set size value determines the minimum number of aspects contained in an itemset for an itemset to be deemed frequent. As a result, the minimum set size establishes the minimum size of the concept graph.

The minimum support value determines the minimum number of observations that must contain the aspects of the specific itemset to make the itemset a frequent item set.

The minimum Jaccard index value determines the similarity threshold of the produced frequent itemsets, which is the overlapping threshold of the observations. The value lies between zero and one. A value of one indicates no overlapping, implying that the observations share all their aspects exclusively. Relaxing the value allows the concept graphs to overlap, the closer the value gets to zero the more overlapping is allowed.

Lastly, the target type variable determines the frequent itemset representation and could be set to closed, maximal or frequent. When set to “closed” only frequent itemsets for which no superset with the same support exists are included. When set to “maximal” only frequent itemsets that have

no immediate supersets that are frequent, are included. The default value is "frequent" which includes all frequent itemsets in the representation.

Table 14 JIM algorithm execution values

Procedure	Min itemset size	Min support	Min Jaccard index value	Target type
Transplant of kidney (procedure)	9	10%	.01	Maximal
Transplant of liver (procedure)	10	10%	.01	Maximal
Transplant of heart (procedure)	10	10%	.01	Maximal
Double lung transplant (procedure)	10	20%	.02	Maximal
Single lung transplant (procedure)	10	20%	.02	Maximal

The aim of the data mining step was to extract concept graphs with the highest level of support and largest diameter. This was a balancing act, because as the diameter increases the support decreases. Therefore, the trade-off is either the concept graph is very specific with relatively little support, or more general with higher support. This implies that what is lost on specificity is gained in support. In addition, the aim was to create similar concept graphs in terms of diameter and support across the different procedure types. For this reason, the variable settings across the five procedure types had to be consistent.

The IHIN had a diameter of 15, of which one node was the observation ID. Three of the node types, SNOMED_CID, APRDRG and total-charge-category stayed constant within each procedure type. Therefore, only eleven of the 15 node types contributed to the specificity of the itemset. The minimum itemset size was set to 10, which included the three constant value node types, as well as seven of the remaining eleven node types.

For the purpose of this study, the variables were set to the values as depicted in Table 14. The minimum support was set to 10%, except for the two smallest sets, namely the single and double lung transplants, as 10% for them worked out to one observation each. As a result, their minimum support, to be of significance, was set to 20%. Note that the minimum itemset value for the kidney transplant procedure was set to nine as no frequent itemsets with a size of 10 could be generated by the JIM algorithm.

The extracted concept graphs for each of the five selected transplant procedure types, with cost anomalies, are summarised in Table 15.

Table 15: Extracted concept graphs

Liver Transp	APRDRG	TChg	LOS	Age	Risk_Mortality	APRDRG_Severity	Comorbidities	Outcome	Diagnosis		Various	ISSize	Support	JaccIdx
18027006	APRDRG-1	5	15		Dying-Extreme	Function-Extreme	CM_LIVER	Live	DX51881		M	10	9	13
18027006	APRDRG-1	5	15	45-64	Dying-Extreme	Function-Extreme	CM_LIVER	Live			M	10	8	11
18027006	APRDRG-1	5	15	45-64	Dying-Extreme	Function-Extreme	CM_LIVER	Live			P3	10	8	11
18027006	APRDRG-1	5	15	45-64	Dying-Extreme	Function-Extreme	CM_LIVER		DX51881		Live	10	7	10
Kidney Transp	**APRDRG**	**TChg**	**LOS**	**Payer**	**Outcome**	**APRDRG_Severity**	**Comorbidity**	**Comorbidity**	**Diagnosis**	**Gender**	**Various**	**ISSize**	**Support**	**JaccIdx**
70536003	APRDRG-440	5		P1	Live	Function-Major	CM_ANEMDEF		DX40391	M		9	6	14
70536003	APRDRG-440	5		P1	Live	Function-Major	CM_ANEMDEF	CM_LYTES		M		9	6	14
70536003	APRDRG-440	5	15	P1	Live		CM_ANEMDEF	CM_LYTES		M		9	5	12
70536003	APRDRG-440	5	15	P1	Live	Function-Major	CM_ANEMDEF			M		9	5	12
70536003	APRDRG-440	5	15	P1	Live	Function-Major			DX40391	M		9	5	12
70536003	APRDRG-440	5	15	P1	Live			CM_LYTES		M	45-64	9	5	12
70536003	APRDRG-440	5		P1	Live		CM_ANEMDEF		DX40391	M	Dying-Moderate	9	5	12
70536003	APRDRG-440	5		P1	Live	Function-Major	CM_ANEMDEF			M	45-64	9	5	12
70536003	APRDRG-440	5		P1	Live	Function-Major	CM_ANEMDEF			M	Dying-Moderate	9	5	12
70536003	APRDRG-440	5		P1	Live	Function-Major		CM_LYTES	DX40391	M		9	5	12
70536003	APRDRG-440	5		P1	Live	Function-Major			DX40391	M	Dying-Moderate	9	5	12
70536003	APRDRG-440	5		P1	Live	Function-Major		CM_LYTES		M	Dying-Moderate	9	5	12
70536003	APRDRG-440	5		P1	Live	Function-Major	CM_ANEMDEF	CM_LYTES	DX40391			9	5	12
70536003	APRDRG-440	5		P1	Live		CM_ANEMDEF	CM_LYTES	DX40391		Dying-Moderate	9	5	12
70536003	APRDRG-440	5		P1	Live	Function-Major	CM_ANEMDEF		DX40391	Age	Dying-Moderate	9	5	12
70536003	APRDRG-440	5		P1	Live		CM_ANEMDEF		DX40391	45-64	Dying-Moderate	9	5	12
70536003	APRDRG-440	5		P1	Live		CM_ANEMDEF		DX40391	45-64		9	5	12
70536003	APRDRG-440	5		P1	Live	Function-Major	CM_ANEMDEF	CM_LYTES		45-64	Dying-Moderate	9	5	12
70536003	APRDRG-440	5		P1	Live	Function-Major	CM_ANEMDEF	CM_LYTES			Dying-Moderate	9	5	12
70536003	APRDRG-440	5		P1	Live	Function-Major	CM_ANEMDEF	CM_LYTES			N6	9	5	12
70536003	APRDRG-440	5		P1	Live		CM_ANEMDEF		DX5856	45-64	Dying-Moderate	9	5	12
Heart Transp	**APRDRG**	**TChg**	**Outcome**	**Various**	**Risk_Mortality**	**APRDRG_Severity**	**Comorbidity**	**Age**	**Diagnosis**	**LOS**	**Various**	**ISSize**	**Support**	**JaccIdx**
32413006	APRDRG-2	5	Live	M	Dying-Extreme	Function-Extreme		45-64	DX51881	15	H51022	11	3	12
32413006	APRDRG-2	5	Live	M	Dying-Extreme	Function-Extreme	CM_LYTES	45-64		N5	CM_COAG	11	3	12
32413006	APRDRG-2	5	Live	M	Dying-Extreme	Function-Extreme	CM_LYTES	45-64		15		10	3	12
32413006	APRDRG-2	5	Live	M	Dying-Extreme	Function-Extreme			DX51881	15	P3	10	3	12
32413006	APRDRG-2	5	Live	M	Dying-Extreme	Function-Extreme	CM_LYTES			15	CM_COAG	10	3	12
32413006	APRDRG-2	5	Live	M	Dying-Extreme	Function-Extreme	CM_LYTES		DX51881	15		10	3	12
32413006	APRDRG-2	5	Live	M	Dying-Extreme	Function-Extreme	CM_LYTES			15	P1	10	3	12
32413006	APRDRG-2	5	Live	M	Dying-Extreme	Function-Extreme	CM_LYTES			15	H51022	10	3	12
32413006	APRDRG-2	5	Live	M	Dying-Extreme	Function-Extreme	CM_LYTES	45-64	DX51881			10	3	12
32413006	APRDRG-2	5	Live	M	Dying-Extreme	Function-Extreme	CM_LYTES	45-64			P1	10	3	12
32413006	APRDRG-2	5	Live	M	Dying-Extreme	Function-Extreme	CM_LYTES	45-64			H42323	10	3	12
32413006	APRDRG-2	5	Live	M	Dying-Extreme	Function-Extreme	CM_LYTES			P1	CM_HTN_C	10	3	12
Dlung Transp	**APRDRG**	**TChg**	**Outcome**	**Gender**	**Payer**	**APRDRG_Severity**	**Risk_Mortality**	**Hospital**	**Comorbidity**	**LOS**		**ISSize**	**Support**	**JaccIdx**
232658009	APRDRG-2	5	Live		P1	Function-Extreme	Dying-Extreme	H42323	CM_LYTES	14		10	4	21
Slung Transp	**APRDRG**	**TChg**	**Age**	**Gender**	**Payer**	**APRDRG_Severity**		**Hospital**	**Comorbidity**	**LOS**		**ISSize**	**Support**	**JaccIdx**
232657004	APRDRG-2	5	45-64	F	P3	Function-Extreme		H48057	CM_HTN_C	13		10	3	27

From the extracted concept graphs the following interesting facts could be derived.

- Firstly, for the liver anomaly transplant procedures, 100% of the concept graph descriptions included the combination of an anomaly total-charge value (5) and an anomaly length-of-stay value (15). Compared to only a third of the kidney anomaly transplant procedure's concept graph descriptions and half of the heart anomaly transplant procedure's concept graph descriptions which included the same combination. This supports the notion, identified during the data description step of the data understanding stage, that the liver transplant procedures had the strongest correlation between the total-charge and length-of-stay attribute values and the kidney transplant procedures the weakest.
- Secondly, the liver and heart anomaly transplant procedure concept graphs shared the same diagnosis code of DX51881 (acute respiratory failure).
- Lastly, 71% of all the observations categorised as anomalies had extreme mortal risk and severity values. However, none of the kidney transplant concept graph descriptions included these extreme values. When mortal risk and severity node types were included in the kidney transplant concept graph descriptions, their values were a mere moderate and major respectively.

The last step of the data mining stage was to construct the extracted concept graphs as described next.

KNIME Workflow

KNIME Workflow

4.4.2 Concept graph construction

For the purpose of this study, one concept graph from each procedure type was selected for construction. The selection was based on how interesting the graph was and diversity. Subsequently the five concept graphs highlighted in the concept graph table, Table 15, were constructed. These workflows were once again automated in KNIME. Figure 36 illustrates the liver anomaly transplant procedure concept graph extraction workflow — KNIME Workflow

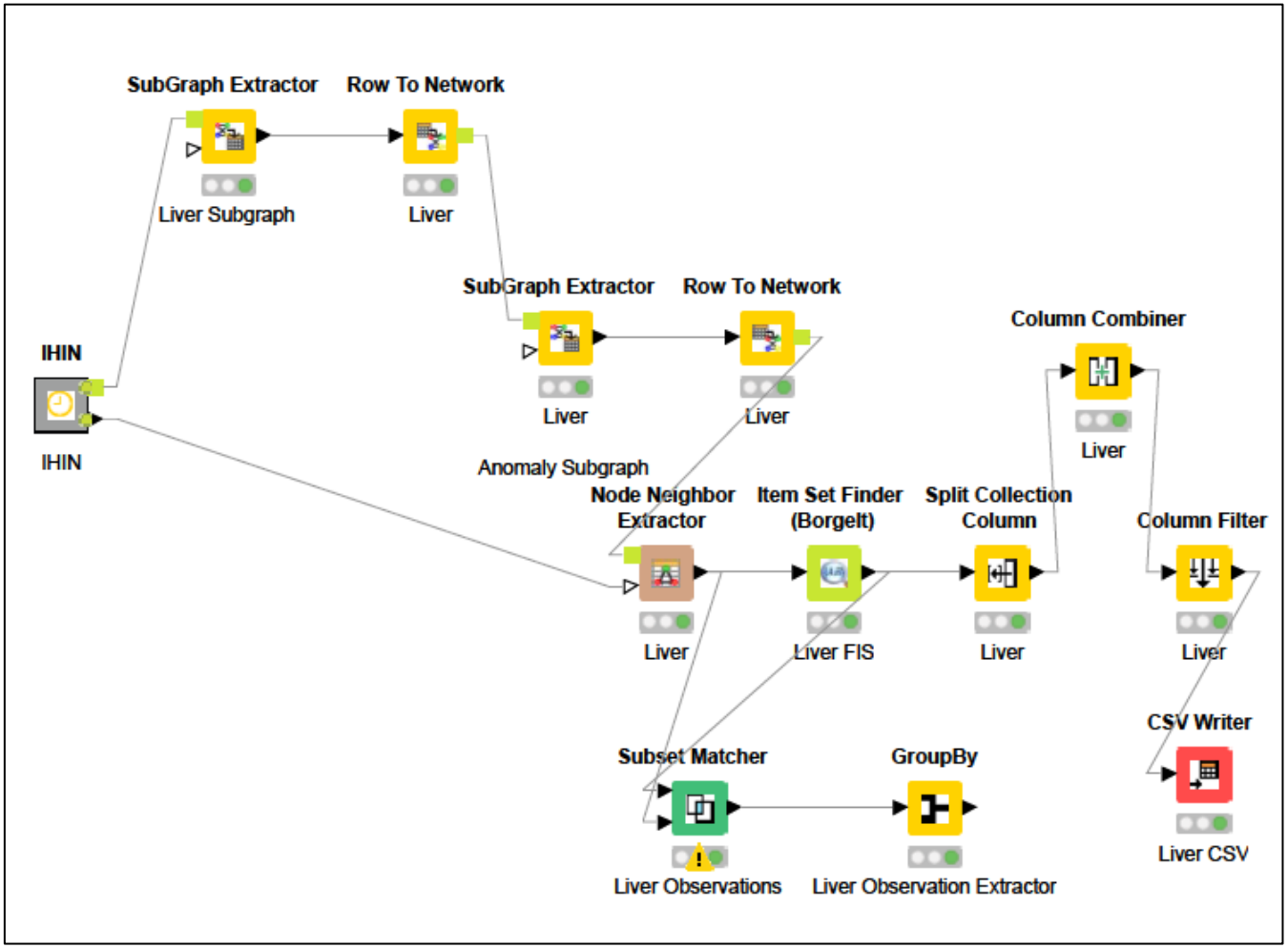

Figure 36: Liver anomaly transplant procedure concept graph extraction workflow (KNIME)

The output information network of the IHIN workflow was the input to the concept graph extraction workflow, as illustrated in Figure 36. This workflow generates two outputs. Firstly, the extraction of the frequent itemsets as described in Section 4.1.1 is written to a file in the "CSV Writer" step. Secondly, the observations that belong to the specific concept graph are extracted by the "Subset matcher" and "Groupby" steps. These observations are then used as part of the Cypher query as illustrated in Figure 37 to extract and visualise the concept graph as illustrated in Figure 38 in Neo4J.

```
MATCH (o:Observation)-[:underwent]-(p:Procedure),
      (o:Observation)-[:diagnosedWith]-(d:Diagnosis),
      (o:Observation)-[:lengthOfStay]-(l:LOS),
      (o:Observation)-[:inGroup]-(g:DiagGroup),
      (o:Observation)-[:cost]-(c:Cost),
      (o:Observation)-[:riskDying]-(m:MortalRisk),
      (o:Observation)-[:severeness]-(s:Severity),
      (o:Observation)-[:ofGender]-(q:Gender),
      (o:Observation)-[:ofAge]-(a:Age) ,
      (o:Observation)-[:payed]-(r:Payer),
      (o:Observation)-[:died]-(t:Mortality),
      (o:Observation)-[:visits]-(h:Hospital),
      (o:Observation)-[:sufferFrom]-(z:Comorbidity),
      (o:Observation)-[:noCond]-(n:NChronic)
where o.PATIENTID = 182011100011204  or
o.PATIENTID = 362011102160735 or o.PATIENTID = 402011100443099 or
o.PATIENTID = 422011100862917 or o.PATIENTID = 422011100862980 or
o.PATIENTID = 422011101004588 or o.PATIENTID = 422011100881339 or
o.PATIENTID = 482011100444509 or o.PATIENTID = 482011100457296
RETURN o,p,d,l,g,c,m,s,q,a,r,t,h,z,n
```

Figure 37: Cypher query (Liver anomaly transplant procedure concept graph)

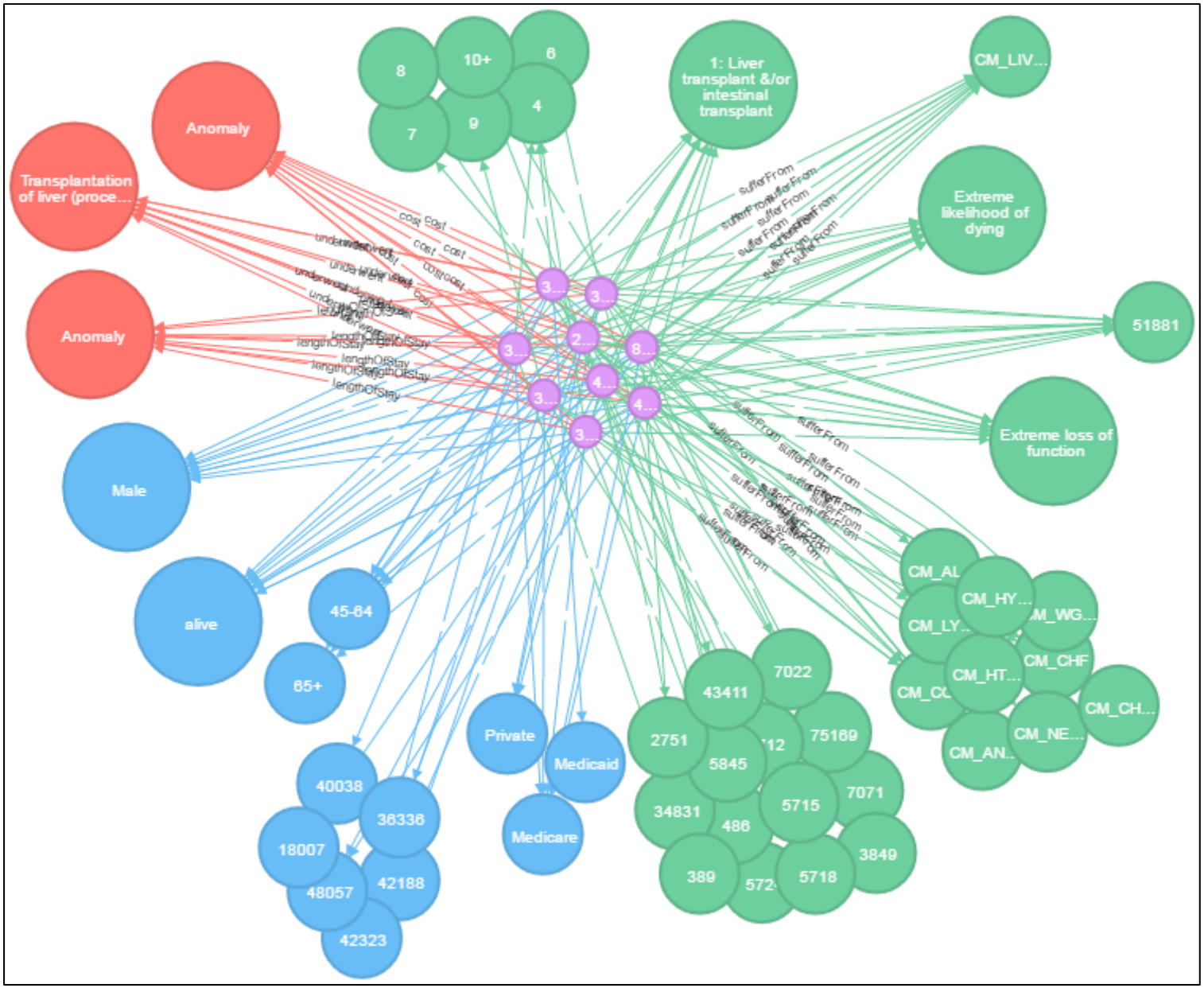

Figure 38: Liver anomaly concept graph

The following four concept graphs were constructed in a similar way

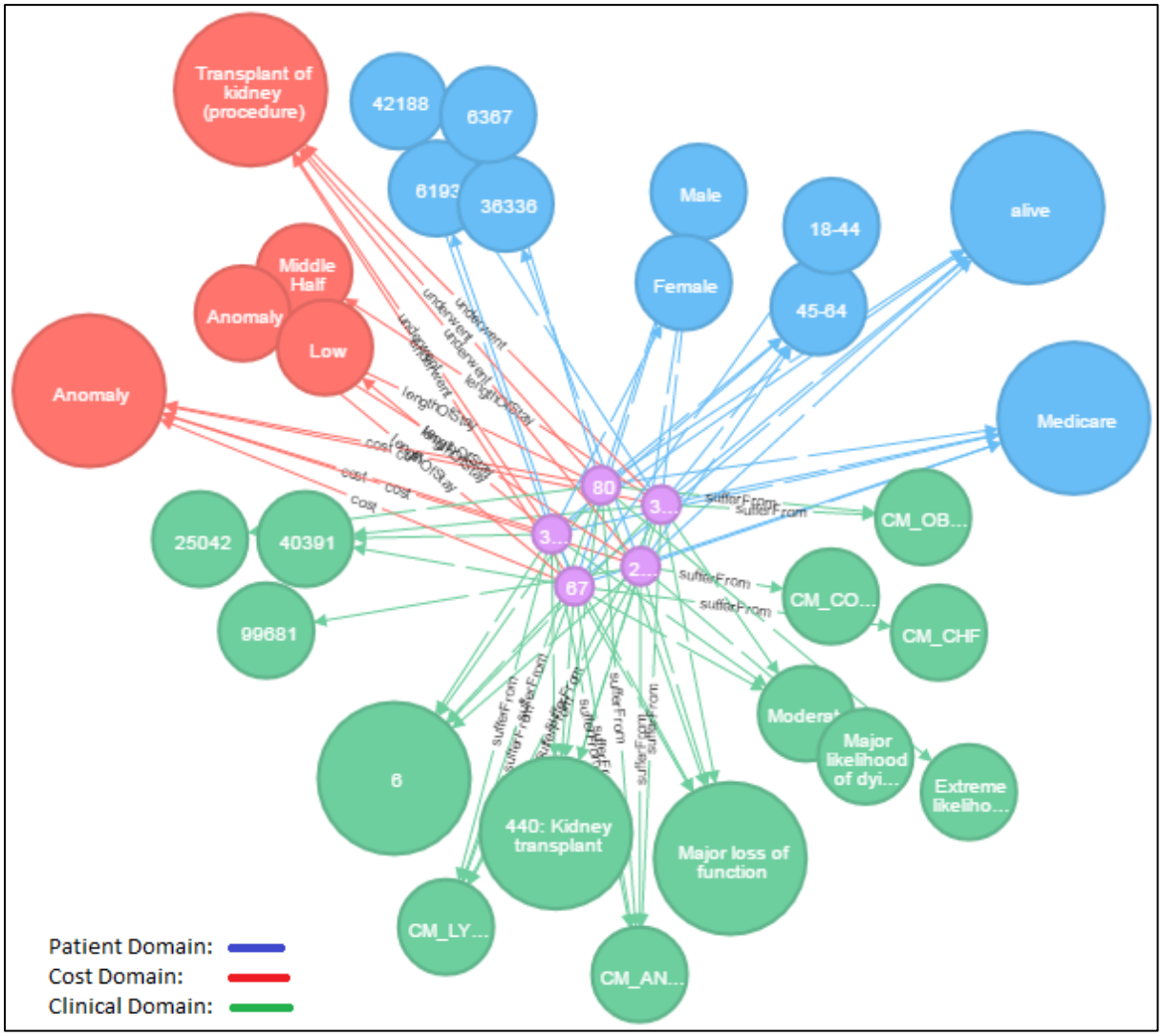

Figure 39: Kidney anomaly concept graph

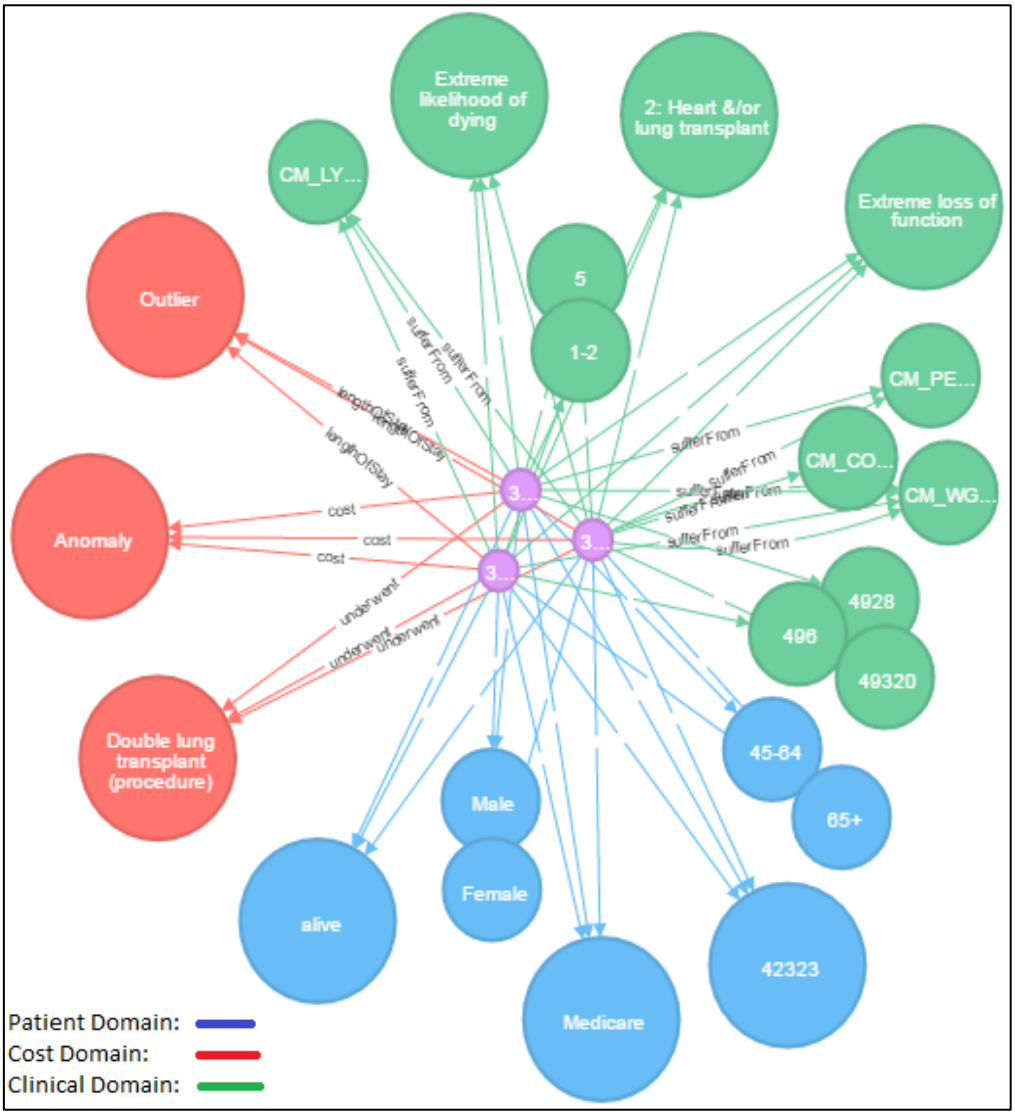

Figure 40: Double lung anomaly concept graph

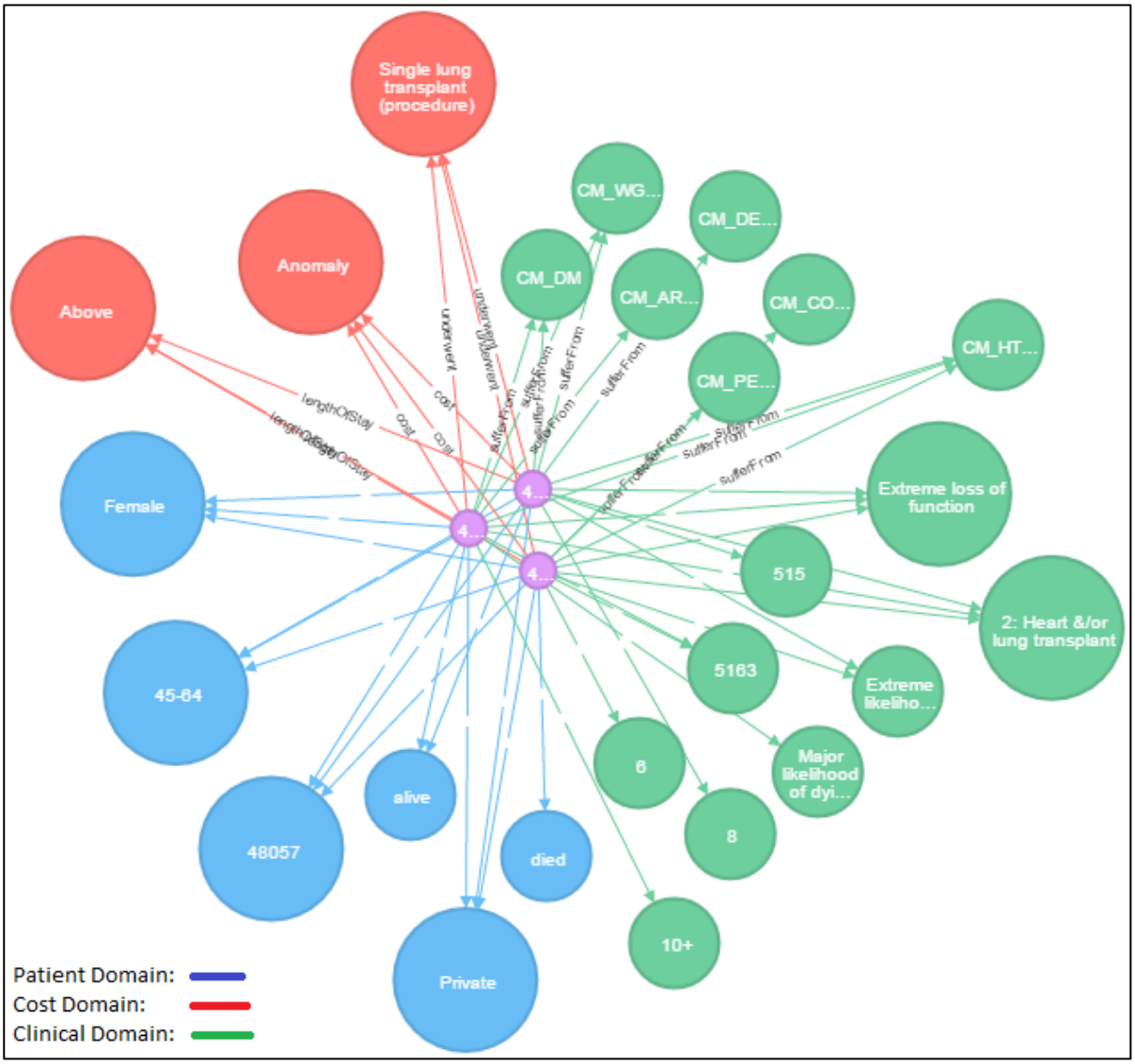

Figure 41: Single lung anomaly concept graph

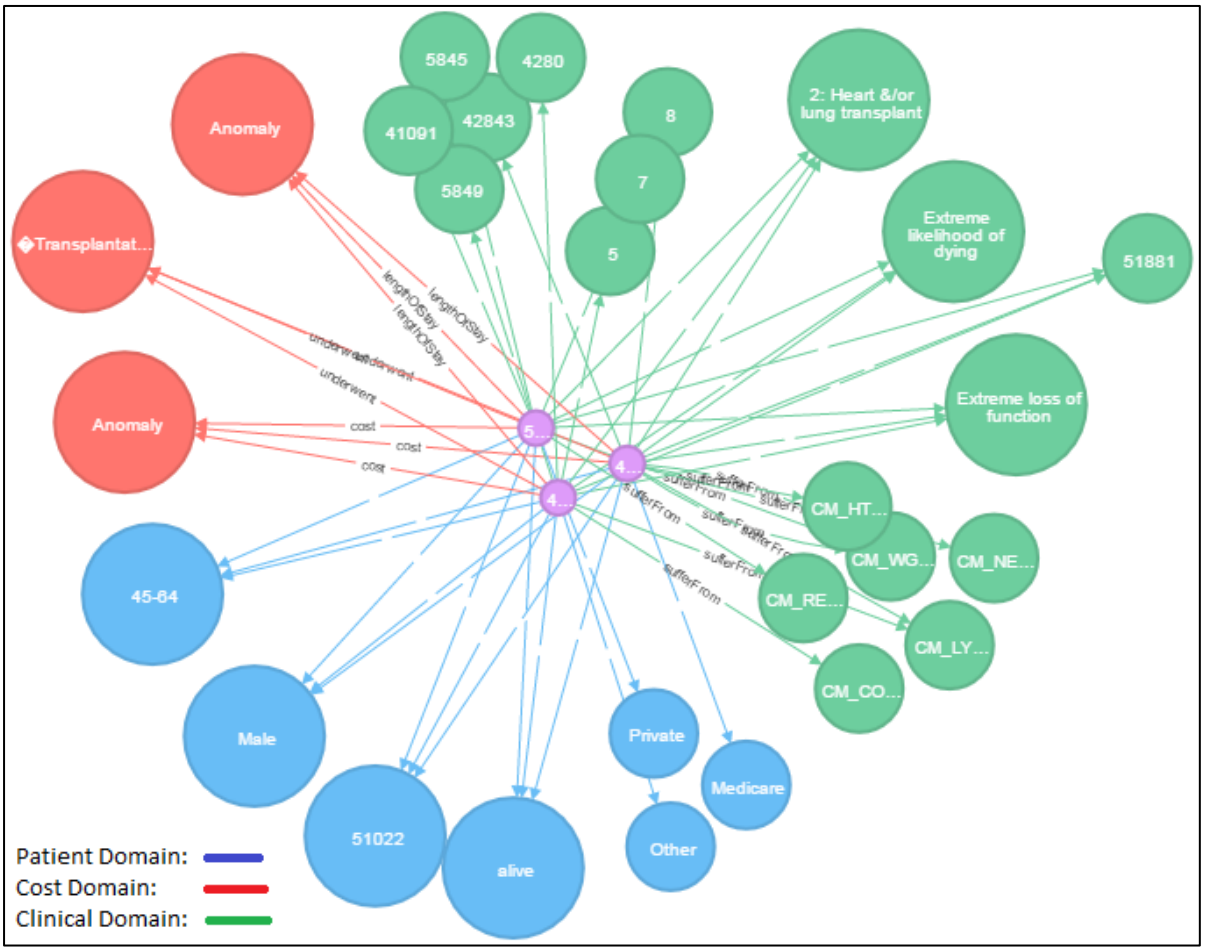

Figure 42: Heart anomaly concept graph

Each of the five concept graphs represented a concept with its member and aspect sets and could be defined as follows:

$C_{liver\ anomaly}(V_A, V_M, E_C) \subseteq B$ defines the concept graph $C_{liver\ anomaly}$ in the BisoNet B.

Where $V_A \subseteq V$ represent the aspect set and $V_M \subseteq V$ represent the member set and $V_A \cap V_M = \emptyset$. $E_C \subseteq E$ is the subset of all edges that connect vertices within the concept graph and $E_C = \{\{u, v\} \in E : u, v \in V_A \cup V_M\}$.

Where

V_A = {18027006, APRDRG-1, 5, 15, Dying-Extreme, Function-Extreme, CM_LIVER, Live, DX51881, M} and

V_M = {182011100011204, 362011102160735, 402011100443099, 422011100862917, 422011100862980, 422011101004588, 422011100881339, 482011100444509, 482011100457296}

A similar process was followed to define the four concept graphs illustrated in Figure 38 to 41.

In the graphs, the members (observations) were represented by purple nodes and the aspects by large blue, red and green nodes. The colour of the nodes corresponds to the colour of the domain the node belongs to, blue to the patient domain, red to the cost domain and green to the clinical domain. The smaller blue, red and green nodes represent the non-frequent items. From these concept graph visualisations, it can be clearly seen that the graphs contain aspects from different domains, therefore they were interpreted as ***domain bridging*** concept graphs. Secondly, from these visualisations it was clear to see that graphs also shared aspect members and therefore could be interpreted as ***overlapping*** concept graphs connecting concepts from heterogeneous domains.

The novelty and interestingness of the knowledge discovered is discussed next.

4.5 Knowledge evaluation

The previous section illustrated the ability of frequent pattern mining algorithms to detect frequent subgraphs in integrated heterogeneous information networks by making use of concept graph detection. This section demonstrates the use of concept graphs to identify domain bridging concepts, as well as the use of overlapping concept graphs to identify domain bridging graphs. In both instances, it was for the purpose of bisociative knowledge discovery.

4.5.1 Domain bridging concept discovery

Most frequently, domain bridging concepts are ambiguous and ill-defined, not contributing towards novel knowledge discovery but rather towards incorrect conclusions. On the contrary, domain bridging concepts can be precise, well defined metaphors and have the potential to contribute towards bisociative knowledge discovery by connecting apparently unrelated information units (Kötter, Thiel, & Berthold, 2010).

The ‘Cost anomaly liver transplant’ concept graph demonstrated the ability of graphs to detect novel domain bridging concepts. The members of this concept graph’s aspect set, as illustrated in Figure 43: Cost anomaly liver transplant concept 432, were from all three domains namely: patient, cost and clinical.

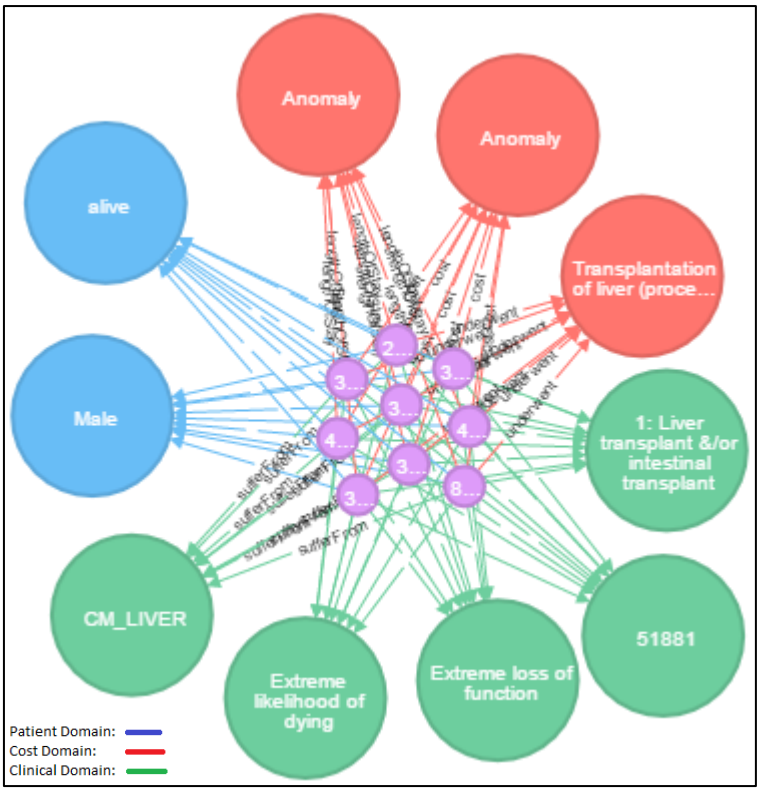

Figure 43: Cost anomaly liver transplant concept

Three aspect members were from the cost domain namely:

- anomaly of type cost,
- anomaly of type LOS and
- transplant of liver of type procedure.

Two aspect members were from the patient domain namely:

- male of type gender and
- alive of type outcome.

Lastly, four aspect members were from the clinical domain namely:

- liver disease co-morbidity,
- extreme likelihood of dying as type risk of mortality,
- extreme loss of function as type severity of illness,
- 51881 (acute respiratory failure) diagnosis and
- liver transplants as APR-DRG class.

In the context of this study, it could be said that the alive aspect member was ambiguous as the majority of the observations in the original dataset had an outcome type of alive. However, the combination of the domain bridging concepts:

- transplantation of liver and
- cost anomaly

could be interpreted as a metaphor for the combination of domain bridging concepts:

- length of stay anomaly,
- liver disease co-morbidity,
- extreme likelihood of dying,
- extreme loss of function,
- acute respiratory failure and
- liver transplant APR-DRG class.

Consequently, a patient who was diagnosed with acute respiratory failure, with a predicted extreme likelihood of dying, an extreme loss of function, classified with an APR-DRG class of liver transplant and an abnormal length of stay in the hospital was the type of liver transplant patient that would have incurred abnormal high costs.

Similarly, from the ‘Cost anomaly heart transplant’ concept graph, a cost anomaly heart transplant patient could be described as a patient between

the age of 45 and 64, who visited hospital 51022, was diagnosed with acute respiratory failure, with a predicted extreme likelihood of dying, extreme loss of function, was classified in the APR-DRG class of heart and/or lung transplant and had an abnormal length of stay in hospital.

A 'Cost anomaly kidney transplant' patient could be described as an anaemic, Medicare patient, with six chronic diseases, fluid and electrolyte disorders, a major loss of function and classified in the APR-DRG class of kidney transplant.

Interesting to note was that unlike the other transplant procedures selected for this study, the incurrence of abnormal cost of a kidney transplant does not have a relation to the length of stay of the patient, a predicted extreme likelihood of dying or an extreme loss of function. Also, novel to notice was that the only cost anomaly concept graph that included a female gender type was one belonging to the single lung transplant patients.

The discovered knowledge should be further investigated and subsequently be applied to refine the IHIN.

4.5.2 Domain bridging subgraph discovery

Domain bridging graphs are subgraphs of overlapping concept graphs with the potential that these subgraphs might lead to novel insights (Kötter *et al*., 2010). The 'Heart transplant cost anomaly' and 'Liver transplant cost anomaly' concept graphs demonstrated this ability of overlapping concept graphs to identify domain bridging subgraphs. Concept members from both concept graphs were connected by their shared aspect members namely: male of type gender, alive of type outcome, anomaly of type cost category, anomaly of type length of stay category, a predicted extreme likelihood of dying, an extreme illness severity level and a diagnosis of acute respiratory failure, as illustrated in Figure 44.

Figure 44: Domain bridging subgraph

The hypothesis that could be generated from this domain bridging subgraph is that male, heart and liver transplant patients that suffer from acute respiratory failure are most likely to incur a higher cost for their transplant procedure. This hypothesis should be investigated further.

4.6 Summary

In this case study:

- **The application of bisociative knowledge discovery revealed unexpected, interesting relationships within the data.**

  In this section, the use of concept graphs to identify domain bridging concepts was demonstrated. For example, the 'cost anomaly liver transplant' concept graph demonstrated the ability of this approach to detect novel domain bridging concepts. In addition, the use of overlapping concept graphs, to identify domain bridging subgraphs, was demonstrated.

- **The data of complex organisations, modelled as integrated, heterogeneous information networks, provided visual insights into the structure of the data that would not otherwise have been visible using traditional approaches**.

  The network schema presented in Figure 35 served as a high level abstraction of the BisoNet which allows the user a visual insight into the structure of the three incompatible domains and their interconnectedness. By means of frequent subgraph mining, concept graphs were extracted from this BisoNet. The visualisation of the domain bridging concept graphs, as illustrated in Figure 43, made the detection of aspects from different domains visible. Similarly did the detection of domain bridging subgraphs, by identifying their shared aspect members, as illustrated in Figure 44.

- **A methodology was developed to unpack complex organisations into layers of abstraction and develop functions that automate the integration of these layers.**

  Figure 45 summarises the computational model for biomedical informatics that was developed by the application of

computational thinking principals, as introduced in Section 3.2.1, during the implementation of this case study.

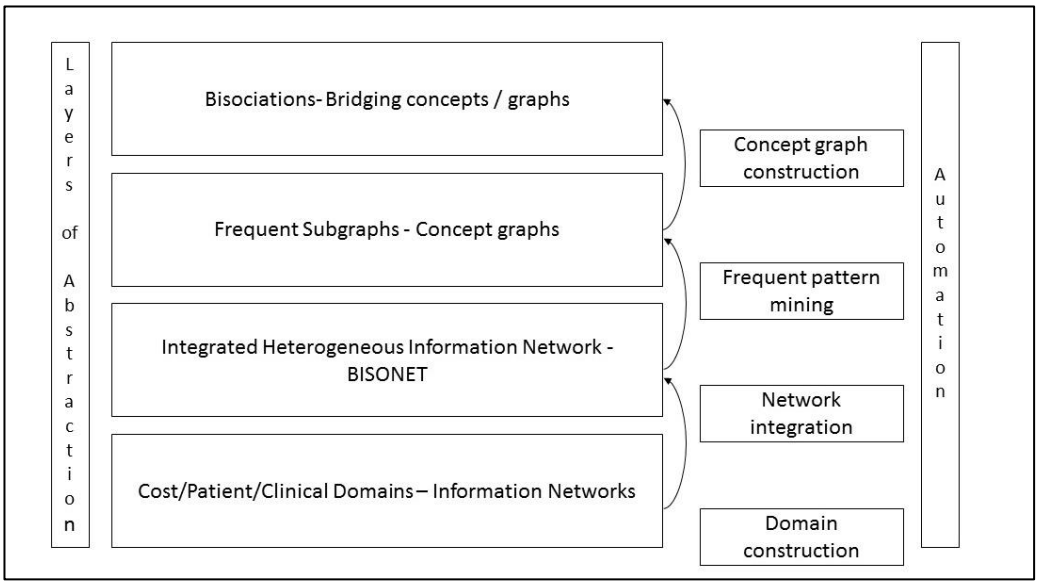

Figure 45: Computational model for biomedical informatics

Firstly, the complex healthcare sector data was unpacked into three habitually, incompatible domains namely; cost, clinical and patient. Each domain was then modelled as an information network. The following layer of abstraction was the BisoNet modelled as an integrated, heterogeneous information network. These two layers interact with each other by means of an abstraction function known as the network integration function. This function was automated using KNIME.

This layer was followed by concept graphs modelled as frequent subgraphs. The BisoNet layer and concept graph layer interact by means of frequent pattern mining algorithms which were also automated using KNIME.

The highest layer of abstraction consisted of the bisociations modelled as bridging concepts and graphs. The interaction

between the concept graphs and bisociation layer was by means of a human-machine combination.

Chapter 5

Summary and conclusion

5.1 Summary

Recall from Section 1.2 that the objective of this study was to construct and motivate a decision support process model dedicated to the field of biomedical informatics. This was achieved by the application of an exploratory approach to data-intensive science where previously hidden patterns emerged from the data. This was accomplished by the application of existing methods in a new and novel way compared to the traditional confirmatory approach, where only hypothesised patterns expected from data are examined.

In Chapter 3, a new data mining and knowledge discovery process model for bisociative knowledge discovery was introduced. This process model makes use of graph structures to integrate heterogeneous data into different information networks. These integrated networks then serve as the starting point for exploratory knowledge discovery. The proposed process model utilizes graph mining techniques, to analyse complex patterns in the information networks. During the knowledge evaluation phase the discovered knowledge is either used to refine the developed decision support model, or used to develop a concrete hypothesis for further investigation.

The new process model was then applied to a real-world case study as discussed in Chapter 4. Healthcare sector data was used to investigate the dispersed cost structures of certain medical procedures. Through the

application of computational thinking during the execution of the knowledge discovery process model, a computational model for biomedical informatics was constructed.

In this section the conclusions drawn from this work are presented as part of the summary of the entire research study.

5.2 Conclusions

From this work the following conclusions were drawn:

A. The proposed explorative data mining method using bisociative knowledge discovery revealed unexpected, potentially interesting relationships within the constructed information network.

Classical data mining tasks make use of one of two typical approaches to discover knowledge from data. The one approach, supervised learning, is to fit a model to a given dataset with the aim being for it to predict the behaviour of some underlying system within a certain level of accuracy. The other approach, unsupervised learning, describes part of the dataset in terms of clusters or frequent item sets, which then gives some insight to what led to these patterns. Both approaches have the same assumption in common which is that a hypothesis can be formulated for the data and that the subsequent processing is then determined by a concrete question, based on the hypothesis. This obviously requires some prior knowledge or decision regarding the data involved. Hence, for both of these approaches, the problem to be solved is simplified either by narrowing down the problem through prior knowledge i.e. a training set, or by specifying the form of outcome i.e. cluster or frequent item set.

Explorative data mining overcomes the simplification of a problem phenomenon, for the purpose of knowledge discovery, by creating a more abstract view of the entire dataset. This enables the search for arbitrary interesting patterns on a structural level which are detached from the semantics of the presented information. However, this leads to the discovery of too many patterns and the challenge then becomes how to identify the interesting scarce details i.e. the anomalies in the case of this study. One way to accomplish this is by finding unexpected and potentially interesting relationships that will trigger a user's interest from the many connections presented. This unexpectedness is based on the connection of habitually unrelated domains by means of exclusive domain crossings i.e. bisociations.

Domain crossings occur in different ways; in this study, two types were investigated namely: domain crossings by means of domain bridging concepts, as discussed in Section 4.5.1, and domain crossings by means of domain bridging graphs, as discussed in Section 4.5.2.

B. This study showed that modelling data from the healthcare sector as an information network allowed visual insights into the structure of the data, which supported the detection of novel insights that otherwise would not have been revealed.

Most organisations' data stores that underlie their decision support systems, frequently known as data warehouses, are structured in a relational way. This is accomplished by either making use of relational data models or dimensional data models based on relational theory. For the purpose of this study, healthcare data structured as a relational data warehouse was remodelled into a graph structure i.e. an integrated, heterogeneous information network. Graph databases became famous in the well-known Human Genome Project where it

was and still is used for genome mapping purposes. However, one mostly comes across this preferred, semi-structured or unstructured way of data modelling in the social media sector, such as modelling social networks i.e. Facebook and LinkedIn, with the emphasis on the interconnectedness of individuals. It is also used in the library science sector: modelling data from bibliographic databases i.e. Scopus and Schools-Wikipedia, with the emphasis on the interconnectedness of authors, authoring papers as well as researchers citing the research papers.

For the purpose of this study, data of the complex healthcare sector, originally structured as a relational data model, was remodelled in a novel way, as an integrated, heterogeneous information network. The important difference between these two approaches is that the focus in the latter is on the structure of the data and not the semantics as it is in the former. In a network, semantics could be provided as additional attributes attached to nodes. But, this only serves to provide a necessary link to the semantic layer, for example to determine where a node originated from. The main advantage of a graph modelling approach is that a structure is not predefined but it emerges from the data, compared to a relational model where data are forced into a predefined structure.

Data of the healthcare sector originated from three habitually incompatible domains and hence could be modelled as a bisociative information network (BisoNet). The three domains were the cost, patient and clinical domains. A network schema was extracted from the data. This schema served as a higher level of abstraction of the data and allowed the user a visual insight into the structure of the data and an ability to drill down into the actual relationships of the individual information units. It is from these relationships that potentially new knowledge was discovered.

Mapping the data as a BisoNet allowed for the detection of domain-bridging associations between otherwise weakly connected domains. These associations provided novel insights into the data, as discussed in Section 4.5, which otherwise would not have been detected.

C. In this study, the success was demonstrated by unpacking a complex organisation into rich layers of abstraction and automating the interpretation thereof through computing.

Computational thinking, as introduced in Section 3.2.1, is a type of analytical thinking with rich abstractions of notions and the automation thereof at its core (Wing, 2008)(Wing, 2008). Crucial to the analysis is the definition of these abstractions, which entails having the knowledge to decide which details should be included in the abstractions and which should not. Further, also important is the automation of the abstractions which entails the simultaneous interpretation of these abstractions by a computer, possible through a clear understanding of the relationships between these layers. These computers can be machines, humans or a combination thereof.

Hence, by means of computational thinking, large complex systems that model complex organisations can be constructed by unpacking the complexity in rich layers of abstraction and automating the interpretation thereof through computing. This was illustrated by the computational model for biomedical informatics developed during this study, as discussed in the summary of Chapter 4.

By creating automated workflows for the creation and simultaneous interpretation of the abstraction layers, as illustrated by the developed computational model, this research adheres to the guidelines of computationally reproducible research as discussed in Section 4.3.4. This makes this data-centric research clear, transparent and organised.

5.3 Recommendations and future work

The challenge in the field of informatics is to automate the process to extract meaning with tools that were primarily designed for data manipulation. In terms of information systems, many disruptive technologies have been introduced in the past five to ten years. These included distributive storage systems as in Apache Hadoop, which uses the MapReduce model, storing data on distributed servers. Also, in-memory database systems like SAP-HANA, IBM-DB2 BLU and Apache-IGNITE were developed which use the main memory as its data store. These technologies allowed industry to model and analyse large volumes of diverse data in real–time.

However, although the underlying technologies of existing knowledge discovery models have changed, the theories and methods that are included in these models have not. This study demonstrated the need to include new data modelling and knowledge discovery methods in an automated process to enable the field of informatics to move forward and make a more significant contribution to support the decision making processes within complex organisations.

To support decision making in complex organisations, industry must move away from using only traditional data manipulation tools, such as relational databases and transactional machine learning algorithms. New methods using information networks, which have the ability to store the interconnectedness of entities within organisations, as input data should be considered. Furthermore, machine learning algorithms that have the ability to extract knowledge hidden within these structures should be applied.

In Section 1.1, Bellazzi *et al*'s (2008) argument regarding the successful application of knowledge discovery process models within the field of biomedical informatics concludes that the healthcare sector has several distinguishing features that must be taken into consideration when considering such models.

These features include:

- the necessity of the integration of data from multiple, unrelated, heterogeneous sources
- the safety critical context of the healthcare sector which requires such models to be interpretable and inspectable
- the need of an explorative interface for the domain experts, for the discovery and formalisation of new hypotheses

The data mining and knowledge discovery process model for biomedical informatics developed during this study addresses these features by firstly, integrating data from heterogeneous sources in a heterogeneous, integrated information network. Secondly, the graph structure of the modelled data and the knowledge discovered allowed visual insights into the structure of the data which makes such models interpretable. Lastly, the explorative data mining approach that was followed allowed for the discovery of new hypotheses by creating a more abstract view of the entire dataset.

As healthcare organisations are increasingly more driven by value-based payment models, it is of critical importance to them to deliver quality care at lower cost. Hence, the importance of discovering the interconnectedness of these, cost vs quality of care, and many more habitually incompatible domains. From the perspective of a private hospital group, it is important to be able to predict quality care which is measured using mortality, infection control, and antimicrobial data, to name a few, and its relationship to cost and patient outcome data. From the perspective of a health insurance organisation, it is important to predict the over servicing of members using co-morbidity, chronic illness and patient data, and its relationship to treatment data.

This study, as in any academic research, was subject to certain limitations. The knowledge discovery process model developed during this study was applied only to the healthcare sector by means of a case study and

the research should be expanded to address the specific needs of the different type of healthcare organisations, such as private hospital groups, health insurance organisations and pharmaceutical companies, amongst others.

References

Ackoff, R. . (1989). From data to wisdom. *Journal of Applied Systems Analysis*, *16*, 3–9.

Adriaans, P., & van Benthem, J. (2008). Introduction: Information is what information does. In P. Adriaans (Ed.), *Philosophy of Information* (Vol. 8, pp. 3–26). Elsevier.

Adriaans, P., & Zantinge, D. (1996). *Data mining*. Harlowi: England: Addison-Wesley.

Aggarwal, C. C. (2013). *Outlier Analysis*. New York: Springer.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceeding VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases* (Vol. 1215, pp. 487–499). http://doi.org/10.1.1.40.6757

AMIA. (n.d.). Informatics Areas: Clinical Informatics. Retrieved June 5, 2014, from http://www.amia.org/applications-informatics/clinical-informatics

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*, 1–3. Retrieved from http://www.wired.com/2008/06/the-end-of-theo/

Anderson, J. G. (1997). Clearing the way for physicians' use of clinical information systems. *Communications of the ACM*, *40*(8), 83–90.

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the Data Deluge. *Science*, *323*, 12971298.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. *International Journal of Medical Informatics*, *77*(2), 81–97. http://doi.org/10.1016/j.ijmedinf.2006.11.006

Bendoly, E. (2003). Theory and support for process frameworks of knowledge discovery and data mining from ERP systems, *40*, 639–647.

Bernstam, E. V, Smith, J. W., & Johnson, T. R. (2010). What is biomedical informatics? *Journal of Biomedical Informatics*, *43*(1), 104–10. http://doi.org/10.1016/j.jbi.2009.08.006

Berthold, M. R., Borgelt, C., Hoppner, F., & Klawon, F. (2010). *Guide to Intelligent Data Analysis*. London: Springer-Verlag.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kotter, T., Meinl, T., … Wiswedel, B. (2009). KNIME - The Konstanz Information Miner, *11*, 26–31.

Berthold, M. R., & Hand, D. (1999). *Intelligent Data Analysis: An Introduction* (1st ed.). New York: Springer-Verlag.

Biomedical Information research Network. (n.d.). Retrieved August 11, 2015, from http://www.birncommunity.org/

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bodenreider, O. (2008). Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of Medical Informatics*, 67–79.

Brachman, R. (1996). The process of knowledge discovery in databases. In *Advances in knowledge discovery and data mining* (p. 611). Menlo Park: American Association for Artificial Intelligence.

Buchan, I., Winn, J., & Bishop, C. (2009). A Unified Modeling Approach to Data-Intensive Healthcare. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The*

Fourth Paradigm (pp. 91–98). Microsoft Research.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A Survey. *ACM Computing Surveys*, *41*(3), 1–58. http://doi.org/10.1145/1541880.1541882

Chiasson, M. W., & Davidson, E. (2004). Pushing the contextual envelope: developing and diffusing IS theory for health information systems research. *Information and Organization*, *14*(3), 155–188. http://doi.org/10.1016/j.infoandorg.2004.02.001

Cios, K. J., & Kurgan, L. a. (2005). Trends in Data Mining and Knowledge Discovery. In *Advanced Techniques in Knowledge Discovery and Data Mining* (pp. 1–26). London: Springer-Verlag. http://doi.org/10.1007/1-84628-183-0

Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, *26*(1–2), 1–24.

Cios, K. J., Teresinska, A., Koniezna, S., Potocka, J., & Sharma, S. (2000). A Knowledge Discovery Approach to diagnosing myocardial Perfusion. *IEEE Engineering in Medicicne and Biology*, *19*(4), 16–25.

Collen, M. F. (1986). Medical Inkonnatics Origins of Medical Informatics.

Dubitzky, W., Kotter, T., Schmidt, O., & Berthold, M. R. (2012). Towards Creative Information Exploration Based on Koestler's Concept of Bisociation. In M. R. Berthold (Ed.), *Bisociative Knowledge Discovery* (pp. 11–32). Springer Online.

Dunn, J. M. (2008). Information in Computer Science. In P. Adriaans & J. van Benthem (Eds.), *Philosophy of Information* (Vol. 8, pp. 581–608). Elsevier. http://doi.org/10.1016/B978-0-444-51726-5.50019-4

Elliot, T. (1934). The Rock. In T. Elliot (Ed.), *Book of Word*. London.

Fayyad, U., & Uthurusamy, R. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Commun. ACM*, *39*(11), 27–34. http://doi.org/10.1145/240455.240463

Fensel, D. (2001). *Ontologies: A silver bullet for knowledge management and electronic commerce*. New York: Springer-Verlag. http://doi.org/10.3359/oz0234121

Floridi, L. (2008). Trends in the philosophy of information. In P. Adriaans & J. van Benthem (Eds.), *Philosophy of Information* (Vol. 8, pp. 113–131). Elsevier. http://doi.org/10.1016/B978-0-444-51726-5.50009-1

Floridi, L. (2014). Semantic Conceptions of Information. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. Retrieved from http://plato.stanford.edu/archives/spr2014/entries/information-semantic/

Futschik, M. ., Sullivan, M., & Kasabov, N. (2003). Prediction of clinical behaviour and treatment for cancer. *Applied Bioinformatics*, *2*, S53–S58.

Gardner, R. M., Overhage, J. M., Steen, E. B., Munger, B. S., Holmes, J. H., Williamson, J. J., & Detmer, D. E. (2009). Core content for the subspecialty of clinical informatics. *Journal of the American Medical Informatics Association : JAMIA*, *16*(2), 153–7. http://doi.org/10.1197/jamia.M3045

Ghrab, A., Vaisman, A., Zimányi, E., Romero, O., & Skhiti, S. (2016). GRAD : On Graph Database Modeling. *arXiv:1602.00503, [cs.DB]*.

Gossen, T., Nitsche, M., Haun, S., & Andreas, N. (2012). Data Exploration for Bisociative Knowledge Discovery: A Brief Overview of Tools and Evaluation Methods . In M. R. Berthold (Ed.), *Bisociative Knowledge Discovery* (pp. 287–300). Springer Online.

Gray, J. (2007). NRC-CSTB_eScience.

Han, J. (2009). Mining Heterogeneous Information Networks by Exploring the

Power of Links. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 13–30). Berlin: Springer-Verlag.

Han, J., Cheng, H., Xin, D., & Yan, X. (2007). Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery*, *15*(1), 55–86. http://doi.org/10.1007/s10618-006-0059-1

Haux, R. (2006). Health information systems - past, present, future. *International Journal of Medical Informatics*, *75*(3–4), 268–81. http://doi.org/10.1016/j.ijmedinf.2005.08.002

Haux, R. (2010). Medical informatics: past, present, future. *International Journal of Medical Informatics*, *79*(9), 599–610. http://doi.org/10.1016/j.ijmedinf.2010.06.003

HCUP Central Distributor. (2015). *Introduction to the HCUP inpatient sample (NIS) 2011* (Vol. 4287). HCUP Central Distributor.

Health expenditure, total (% of GDP). (2015).

Hey, T., Tansley, S., & Tolle, K. (2009). Jim Gray on eScience: A Transformed Scientific Method. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The Fourth Paradigm* (p. xvii). Microsoft Research. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Ji, M. (2012). Classification of Heterogeneous Information Networks. In *Mining Heterogeneous InformationNetworks*. Morgan & Claypool Publishers.

Kaplan, B. (1994). Reducing barriers to physician data entry for computer-based patient records. *Topics in Health Information Management*, *15*(1), 24–34.

Kitzes, J., Turek, D., & Deniz, F. (2017). *The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences*. (J. Kitzes, D. Turek, & F. Deniz, Eds.). Oakland, CA: University of California

Press.

Koestler, A. (1964). *The Act of Creation*. New York: The MacMillan Company.

Kötter, T., & Berthold, M. R. (2012). From Information Networks to Bisociative Information Networks. In M. R. Berthold (Ed.), *Bisociative Knowledge Discovery* (pp. 33–50). Springer Online.

Kötter, T., Thiel, K., & Berthold, M. R. (2010). Domain Bridging Associations Support Creativity. In *Proceedings of the International Conference on Computational Creativity* (pp. 200–204).

Kulikowski, C. a, Shortliffe, E. H., Currie, L. M., Elkin, P. L., Hunter, L. E., Johnson, T. R., … Williamson, J. J. (2012). AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. *Journal of the American Medical Informatics Association : JAMIA*, *19*(6), 931–8. http://doi.org/10.1136/amiajnl-2012-001053

Lorenzi, N. M., & Riley, R. T. (2004). *Managing Technical Change. Organizational Aspects of Health Informatics*. New York: Spriger Science + Business Media Inc.

Markou, M., & Singh, S. (2003). Novelty detection: A review - Part 1: Statistical approaches. *Signal Processing*, *83*(12), 2481–2497. http://doi.org/10.1016/j.sigpro.2003.07.018

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in Ecology and Evolution*, *27*(2), 88–93. http://doi.org/10.1016/j.tree.2011.11.016

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill. http://doi.org/10.1007/978-3-642-21004-4_10

Morris, M. (2016). *2016 Global health care outlook: Battling costs while*

improving care. *Deloitte*.

Musen, M. A., Noy, N. F., Shah, N. H., Whetzel, P. L., Chute, C. G., Story, M.-A., … Team, N. (2012). The National Center for Biomedical Ontologies. *Journal of the American Medical Informatics Association : JAMIA*, *19*, 190–195.

Nagel, U., Thiel, K., & Tobias, K. (2012). Towards Discovery of Subgraph Bisociations. In *Bisociative Knowledge Discovery* (pp. 263–283). Springer.

Nevins, J. ., Huang, E. ., Dressman, H., Pittman, J., Huang, A. ., & West, M. (2003). Towards integrated clinico-genomic models for personalised medicine, comining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Human Molecular Genetics*, *12*(Spec Issue 2), 153–157.

Newman, H. B., Ellisman, M. H., & Orcutt John A. (2003). November 2006/Vol. 46, No. 11 COMMUNICATIONS OF THE ACM. *Communications of the ACM*, *46*(11), 68–77.

Pete, C., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S., & Wirth, R. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, 76.

Piatetsky, G. (2014). KDNuggets. Retrieved August 15, 2015, from http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Pigliucci, M. (2009). The end of theory in science? *EMBO Reports*, *10*(6), 534. http://doi.org/10.1038/embor.2009.111

Pomeroy, S. ., Tamayo, P., Gaasenbeek, M., Sturla, L. ., Angelo, M., McLaughlin, M. ., … Goumnerova, L. . (2002). Prediction of cnetral nervous system embryonal tumour outcome based on gene expression. *Nature*, *415*, 436–442.

Robinson, I., Webber, J., & Eifrem, E. (2015). *Graph databases* (2 nd). USA: O'Rielly Media, Inc.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, *33*(2), 163–180. http://doi.org/10.1177/0165551506070706

Russell, S., & Norvig, P. (1995). *Artificial Intelligence: A modern approach*. New Jersey: USA: Prentice Hall.

Sacha, J., Cios, K. J., & Goodenday, L. (2000). Issues in Automating Cardiac SPECT diagmnosis. *IEEE Engineering in Medicicne and Biology*, *19*(4), 78–88.

Sarkar, I. N. (2010). Biomedical informatics and translational medicine. *Journal of Translational Medicine*, *8*, 22. http://doi.org/10.1186/1479-5876-8-22

Schmidt, O., Kranjc, J., Mozetič, I., Thompson, P., & Dubitzky, W. (2012). Bisociative Exploration of Biological and Financial Literature Using Clustering. In *Bisociative Knowledge Discovery* (pp. 438–451). Springer.

Scott, W. R., Ruef, M., Mendel, P. J., & Caronna, C. A. (2000). *Institutional Change and Healthcare Organizations*. Chicago: The University of Chicago Press.

Segond, M., & Borgelt, C. (2011). Item Set Mining Based on Cover Similarity. In *Proc. 15th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*. Springer-Verlag.

Shortliffe, E. H., & Blois, M. S. (2006). The computer meets medicine and biology: emergence of a discipline. In *Biomedical Informatics* (3 rd, pp. 3–45). New York: Springer.

Sun, Y., & Han, J. (2012). *Mining Heterogeneous Information Networks : Principles and Methodologies*. Morgan & Claypool Publishers.

Sun, Y., & Han, J. (2013). Mining Heterogeneous Information Networks : A Structural Analysis Approach. *SIGKDD Explorations Newsletter*, *14*(2), 20–28.

Sveiby, K. (1996). Transfer of Knowledge and the Information Processing Professions. *European Management Journal*, *14*(4), 379–388.

Tobias, K., & Berthold, M. R. (2012). (Missing) Concept Discovery in Heterogeneous Information Networks. In *Bisociative Knowledge Discovery* (pp. 230–245). Springer.

Torio, C. M., & Andrews, R. M. (2013). *STATISTICAL BRIEF # 160 National Inpatient Hospital Costs : The Most* (Vol. 31).

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

United Nations. (2015). *World population Ageing*. *United Nations, Department of Economic and Social Affairs, Population Division Department of Economic and Social Affairs, Population Division*.

vant' Veer, L. ., Dai, H., van de Vijver, M. ., He, Y. ., Hart, A. ., Mao, M., … van der Kooy, K. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*, 530–536.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33. http://doi.org/10.1145/1118178.1118215

Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *366*(1881), 3717–3725. http://doi.org/10.1098/rsta.2008.0118

World Health Organization. (2015). *World Report on Ageing and Health*. Luxembourg.

Yin, R. (2009). *Case study research : Design and methods (4th ed., Applied*

social research methods series ; v. 5) (4 th). Los Angeles: Sage Publications.

Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997). New Algorithms for Fast Discovery of Association Rules. *3rd Intl Conf on Knowledge Discovery and Data Mining*, *20*(651), 283–286. http://doi.org/10.1.1.42.5143

Appendix A

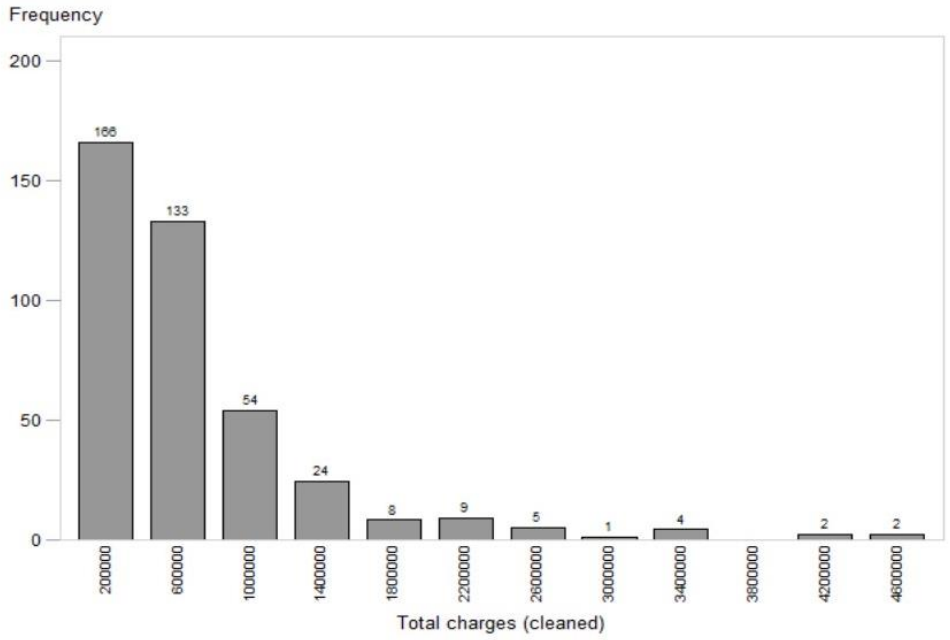

Figure A.1: Heart transplant procedure

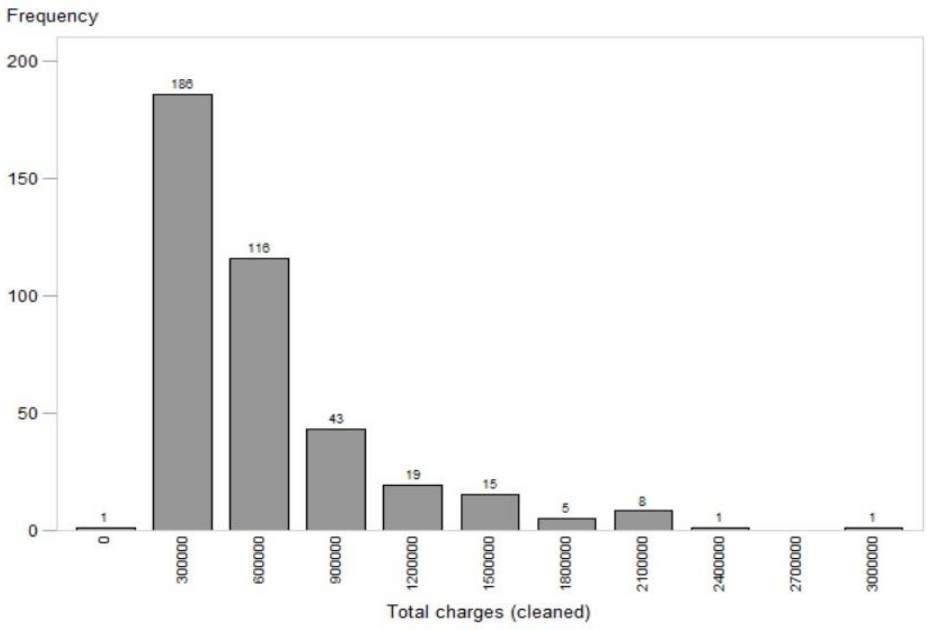

Figure A.2: Double lung transplant procedure

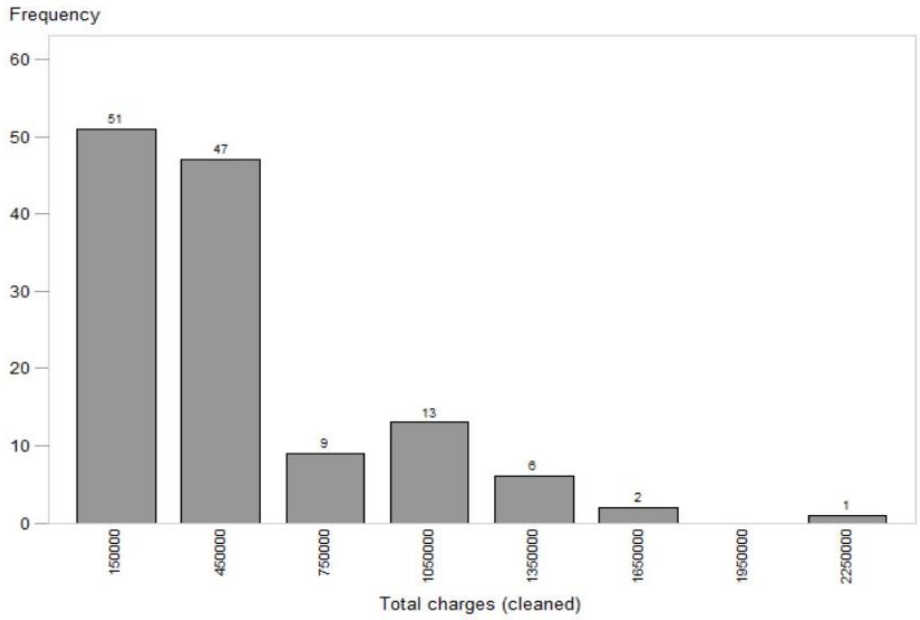

Figure A.3: Single lung transplant procedure

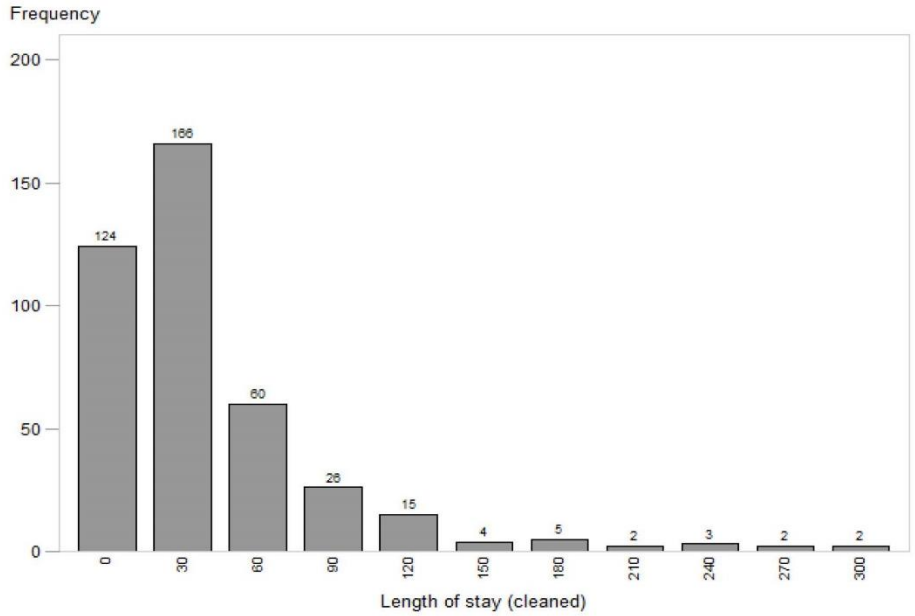

Figure A.4: Heart transplant procedure

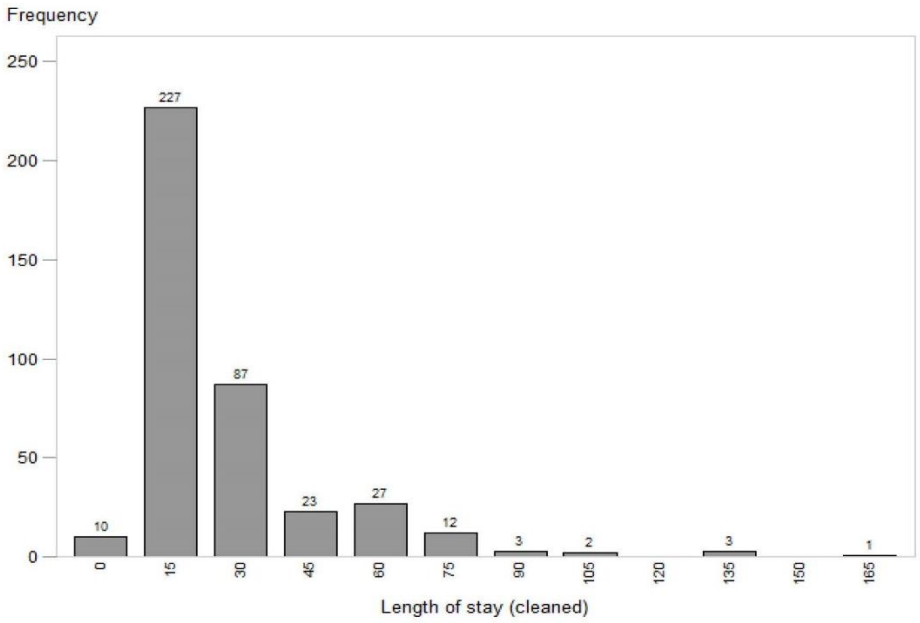

Figure A.5: Double lung transplant procedure

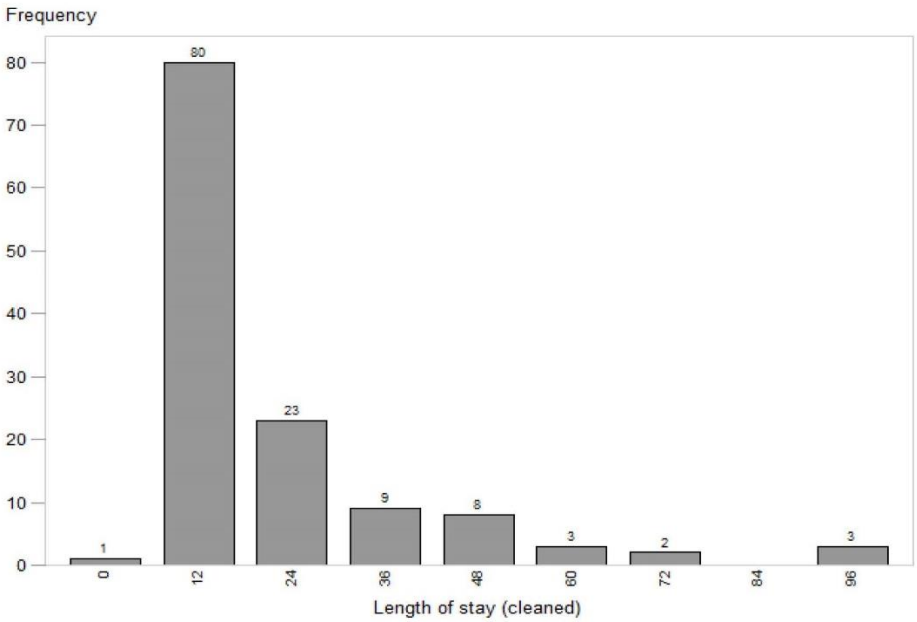

Figure A.6: Single lung transplant procedure

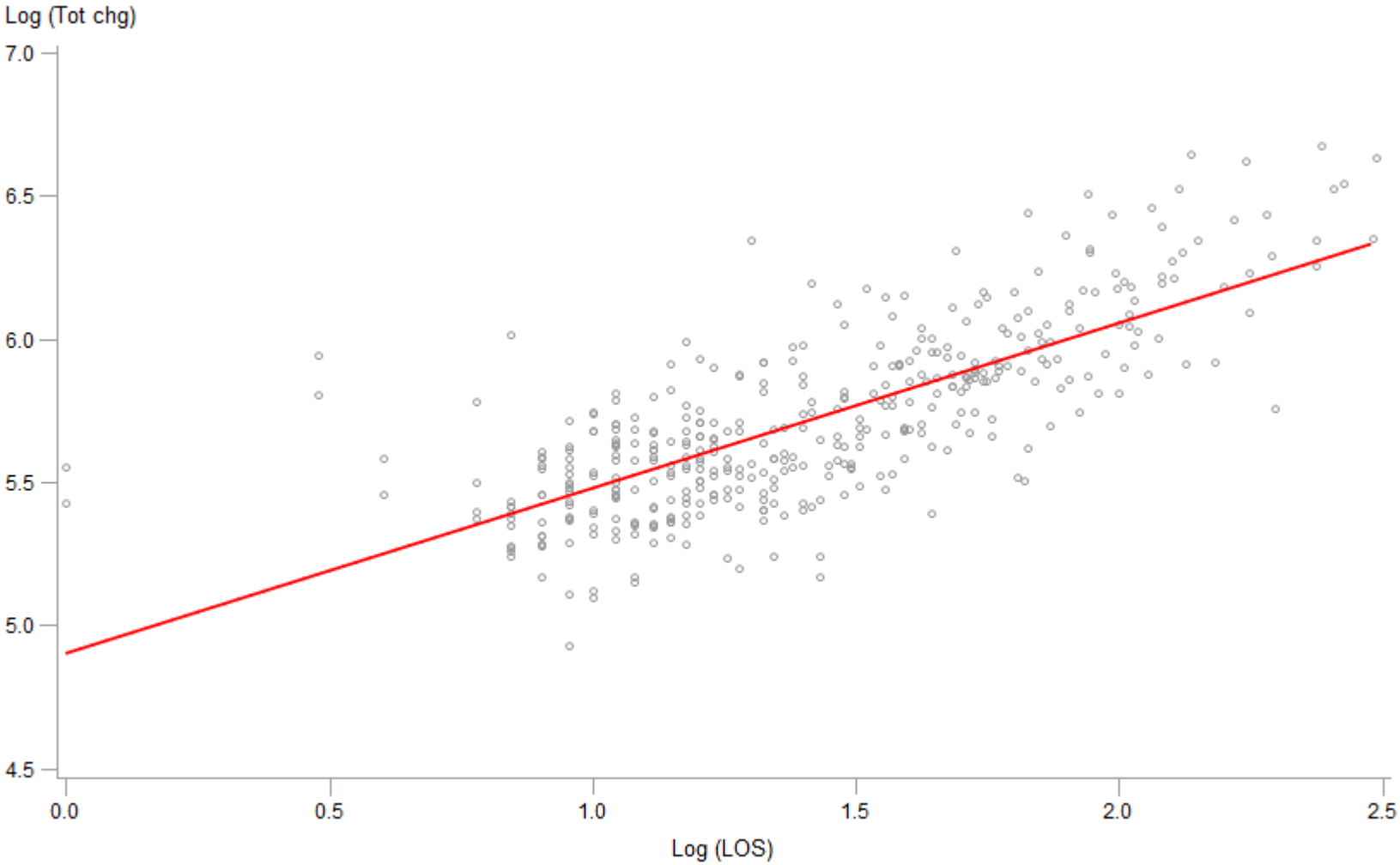

Figure A.7: Transplant of heart (procedure)

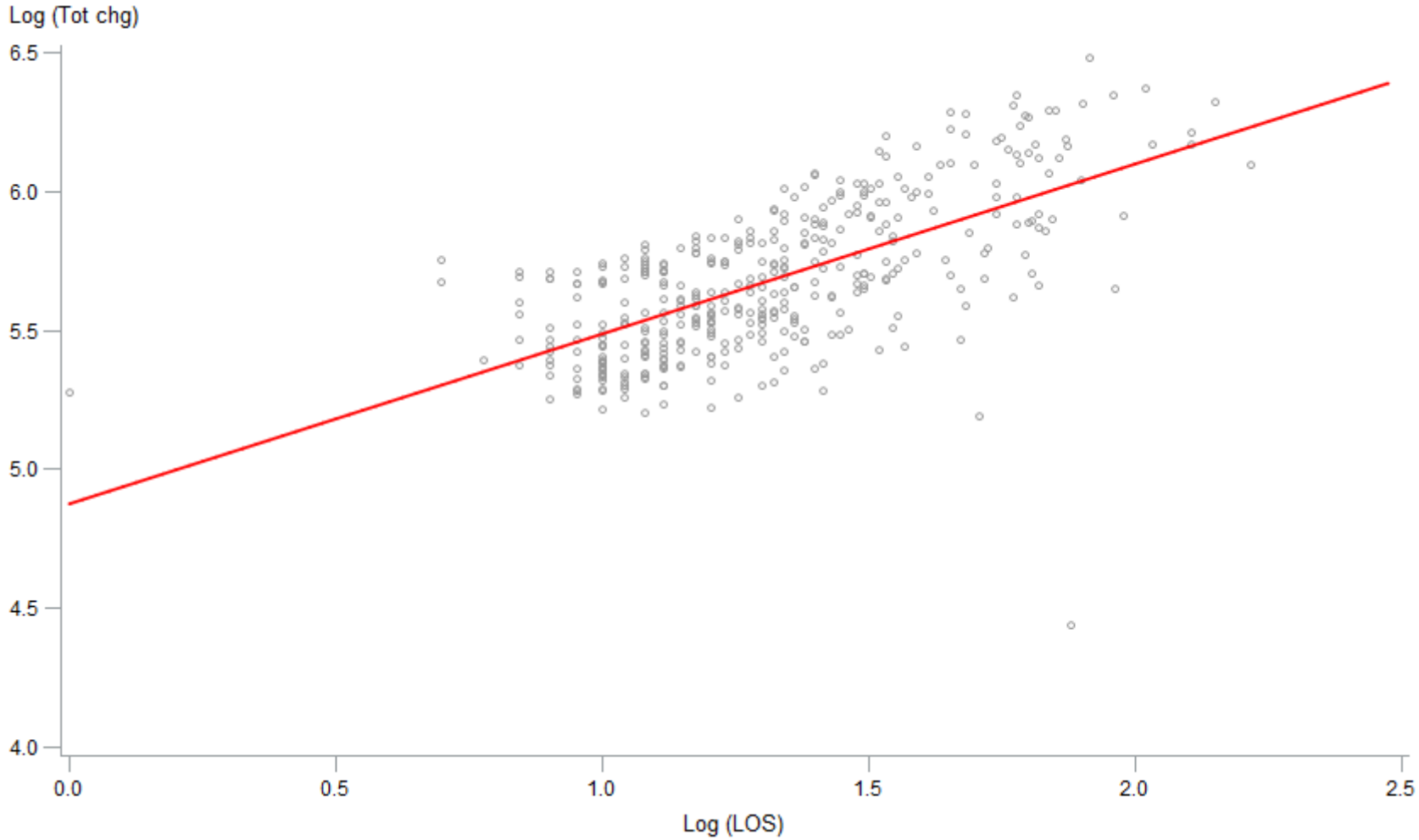

Figure A.8: Double lung transplant (procedure)

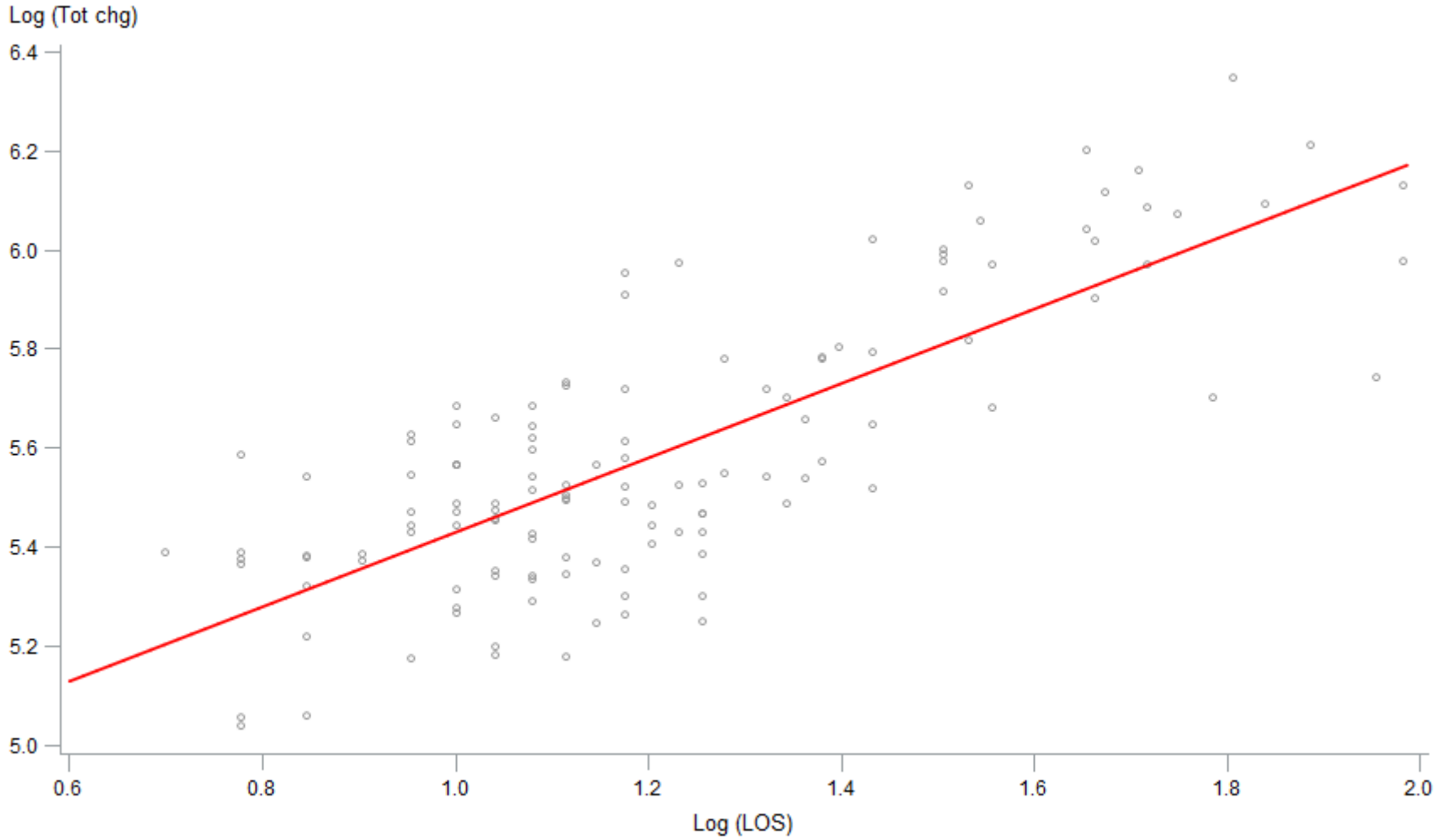

Figure A.9: Single lung transplant (procedure)

Appendix B

SAS Script

KNIME Workflow