



Universiteit Utrecht

MASTER'S THESIS

Patient Careflow Discovery

K.P.A. van Wanrooij
3020037

August 15, 2012
Version 1

Supervisor Deloitte:
Martijn Ludwig

Deloitte.

Supervisor UU:
Ad Feelders

Institute of **Information &**
ICS **Computing**
Sciences



Name: Kenneth van Wanrooij
Student number: 3020037
Title thesis: Patient Careflow Discovery
Date: August 15, 2012

University: Utrecht University
Faculty: Information & Computing Sciences
Company: Deloitte

Supervisor UU: Ad Feelders
Arno Siebes
Supervisor Deloitte: Martijn Ludwig



Abstract

In this study, we explored how different mining techniques can be used to gain insight into the healthcare domain. More specifically, we developed a methodology that takes a set of activity sequences from a Hospital Information System to analyze patient careflow.

We developed a data-based methodology able that provides insight into patient careflow, based on a standardized data structure from the Dutch DBC Information System and other external resources. This approach provides a set of techniques able to analyze any type of care profile, for any specialism, within any hospital and combinations of either. After an initial data collection, an event log is prepared containing high-level activities describing the logistic carepath of patients. Secondly, different types of care profiles are identified by clustering using a Partitioning Around Medoids algorithm based on the Tanimoto distance between paths. The third step is to apply classification in order to identify the main characteristics of each type of profile. As a fourth and final step, each cluster is analyzed using the Trace Alignment plugin in ProM, which allows the identification of both a cluster's main process pattern and specific deviations from this process for individual carepaths.

A variety of insightful visualization techniques allows medical specialist to interpret the results of this methodology without specific knowledge on the Data and Process Mining techniques. The insights gained from this methodology support the improvement of patient careflow in three different ways: by treating patients according to the cheapest path with the highest quality of care, by improving standardization of carepaths and by developing a robust, optimal operating schedule using predictive modeling based on the patient types defined by this analysis.

Keywords: Healthcare, Patient Careflow, Advanced Analytics, Visualization, Data Mining, Clustering, Classification, R, Process Mining, Trace Alignment, Event Logs, ProM

Acknowledgement

This research is the result of my graduation project at the department of Computing Science at Utrecht University (COSC/ACS). The study was executed during an internship at the Business Intelligence and Analytics service line of Deloitte Consulting.

Both Deloitte and Utrecht University have provided me with valuable support, therefore I would like to express my utmost gratitude to Ad Feelders (UU) and Martijn Ludwig (Deloitte) for their supervision, support and ideas on this research.

Also, I am grateful for the support of all my colleagues within Deloitte (BIA, Innovation and S&O). Special thanks goes out to Mark Boersma, Machiel Westerdijk, Ernst Blaauw, Rogier Prince, Willem-Jan Swiebel, Frank Korf, Peter van Stijn, Kim Verkooij, Joshua Ratha, Nadiéh Bremer and Bertien Dumas for hearing – and answering – my many questions, and Stefan van Duin for providing me with the time and resources within Deloitte.

I want to thank Ronny Mans and J.C.B. Rantham Prabhakara (TU/e) for their input on Process Mining and ProM in particular, and Han Hoogeveen and Marjan van den Akker (UU) for helping me get here in the first place.

Last but not least, I would like to show my gratitude to my friends and family who gave me the energy and motivation to finish this project successfully, with in particular Stephanie Blom and Suzette Obbink who fed me delicious coffee and cheesecake during the final stages, and my brother for sharing his knowledge and experience.

Many thanks to all of you,

Kenny



Contents

1	Introduction	1
1.1	Assignment background	1
1.2	Research statement	3
1.3	Research design	4
1.3.1	Business Understanding	4
1.3.2	Data Understanding	5
1.3.3	Data Preparation	5
1.3.4	Modeling	5
1.3.5	Evaluation (Case Study)	5
1.4	Outline	6
2	Theoretical background	7
2.1	Healthcare domain	7
2.1.1	Defining patient careflow	9
2.1.1.1	DBC	10
2.1.1.2	DOT	11
2.1.1.3	Careflow definition	12
2.1.2	Hospital Information System	14
2.2	Data tasks	15
2.2.1	Data Mining	15
2.2.2	Process Mining	16
2.2.3	Data quality	18
2.3	Related work	19
2.3.1	Data Mining	19
2.3.2	Process Mining	20
2.4	Conclusion	21
3	Technical background	23
3.1	Clustering algorithms	23
3.1.1	Clustering using Vector Distances	26
3.1.2	Compression Clustering	27
3.1.2.1	Compression algorithms	28
3.1.2.2	Clustering compressed activity strings	29
3.2	Classification	29

3.3	Process Mining	31
3.3.1	Heuristics Miner	31
3.3.2	Trace Alignment	31
3.4	Conclusion	32
4	Methodology design	33
4.1	Business Understanding	33
4.1.1	Which insights do we require to assess patient careflow?	33
4.1.2	How can we compare, evaluate and advise different carepaths?	34
4.2	Data Understanding	36
4.2.1	Collection	36
4.2.2	Understanding	37
4.2.3	Audit	39
4.2.4	Visualization	40
4.3	Data Preparation	42
4.4	Modeling	44
4.4.1	Create Clusters	44
4.4.1.1	Using Vector representations	44
4.4.1.2	Using String representations	47
4.4.2	Classification of resulting clusters	48
4.4.3	Identification of frequent patterns	49
4.5	Conclusion	49
5	Evaluation (Case Studies)	53
5.1	Arthrosis (hip) - surgical/clinical with joint prosthesis	53
5.1.1	Step 1: Data Preparation	53
5.1.2	Step 2: Analysis	53
5.1.3	Step 3: Results	56
5.2	Arthrosis (knee) - surgical/clinical with joint prosthesis	59
5.2.1	Step 1: Data Preparation	59
5.2.2	Step 2: Analysis	59
5.2.3	Step 3: Results	65
5.3	Malignant breast neoplasm - surgical/clinical	67
5.3.1	Step 1: Data Preparation	67
5.3.2	Step 2: Analysis	67
5.3.3	Step 3: Results	71
5.4	Summary	73
6	Discussion and Conclusions	77
6.1	Business Understanding	77
6.2	Data Understanding and Preparation	78
6.3	Modeling	79
6.3.1	Data Mining	79
6.3.1.1	Clustering	79



6.3.1.2	Classification	80
6.3.1.3	Software package	81
6.3.2	Process Mining	81
6.4	Analysis results	82
6.5	Future Work	82
6.6	Conclusion	84
A	Database structure	III
A.1	DIS_Import	III
A.2	DIS_Stage1	V
A.3	DIS_Stage2	VI
A.4	DIS_Data	VI
A.5	R	VII
A.6	ProM	VIII
A.6.1	Using MS Access to prepare the output datasets	VIII
A.6.2	ProM Import	VIII
B	Modeling images	IX
B.1	Vector clustering ZPK 1 to 8	IX
B.2	Comparing hclust with pam for different distance measures	X
C	CP-Tables from rpart	XIII
D	Arthrosis (hip) - surgical/clinical with joint prosthesis	XIV
D.1	Activity frequency histograms	XIV
D.2	Trace Alignments	XVII
E	Arthrosis (knee) - surgical/clinical with joint prosthesis	XIX
E.1	Activity frequency histograms	XIX
E.2	Trace Alignments	XXV
F	Malignant breast neoplasm - surgical/clinical	XXVIII
F.1	Activity frequency histograms	XXVIII
F.2	Trace Alignments	XXXVI
G	ZPK-code overview	XL



Chapter 1

Introduction

In this study we provide insight into patient careflow by analyzing a hospital derived dataset using various data-based methods, which can be used to improve quality of care whilst decreasing operational costs. To do so, we used a variety of techniques, which have previously shown success in the analysis of healthcare processes. This includes data mining and process mining [28, 36, 43, 55, 59, 61, 68, 78], but also business objectives, mining objectives and external data analysis that have proved useful [4]. This results in a new methodology, that implements the techniques offering the best performance, and at the same time standardizes the business- and mining objectives. The advantage of this methodology is that it is applied to a standard dataset, and thus can be applied to any hospital, any department and for any type of patient careflow.

We start with a short description of the problem in Section 1.1. Section 1.2 will go further into defining our research goal. Section 1.3 describes the approach for the actual research, with a description of the applied case study. An outline of the remainder of this thesis is given in Section 1.4.

1.1 Assignment background

Patients follow a certain *carepath* consisting of a variety of *medical activities* during treatment. These carepaths vary for similar patients not only in *cost*, but also in the *ordering* and *medical content* of activities. The *Diagnosis Treatment Combination (DBC)*¹ system, introduced in 2005, encouraged and improved standardization for similar carepaths of specific diagnosis/treatment combinations [74]. Such a standard is called a protocol. In practice, however, there is often a significant difference between the activities described in a protocol, and the activities actually performed [45]. Hospital board members therefore lack insight and knowledge about the practical implementation of these processes and the system is sensitive to fraud [29, 35]. A new reimbursement model, *DBC Towards Transparency (DOT)*², was introduced in 2012 and aims to provide more transparency in costs and care activities. Together, these factors put a lot of pressure on hospitals to “*do more with less*” [4, 5, 15].

The core elements of the reimbursement systems are DBC-codes. These codes represent a sequence of medical activities for the treatment of patient, where each code corresponds to a specific problem with a specific treatment for a specific medical discipline [74]. The standardization of patient

¹In Dutch: Diagnose Behandeling Combinatie

²In Dutch: DBC Op weg naar Transparantie

careflow has been shown to improve quality of care [6]. Postoperative stay of patients undergoing standardized treatment is shortened and hospital costs decrease [65]. It is however not an easy task to accomplish: manual analysis of these patient careflow processes by interviews is time-consuming and often sub-optimal [55, 68]. Besides, manually defined processes are not applicable for standardization across different departments and hospitals [32]. Previous work has shown that improvement of the treatment process receives significant attention in healthcare, but the use of readily available data is still limited [59].

Instead, data-driven techniques (data mining and process mining) allow us to perform thorough analysis and model patient careflow. Although these techniques are similar, we consider them separate as process mining also includes process modeling in addition to data mining (see Sections 2.2.1 and 2.2.2 for further elaboration). Both mining techniques help to determine the current process and find homogeneous flows and sequences. Homogeneity for a group of carepaths can be defined based on, amongst others, the medical content, the logistics process, total costs or a combination of these elements. As a result, we obtain objective clusters, statistics and other insights that can be used for further standardization of DBC's.

Dutch hospitals use a variety of Hospital Information Systems (HIS) and other embedded systems (e.g. an X-ray device) that track the activities performed in a carepath. Each of these systems records a huge amount of data, for example every diagnosis and treatment activity performed. Information Technology continuously aims to support and improve the healthcare sector, which results in an ever-increasing large quantity of electronic data from operational systems [32, 37, 43, 73].

It is stated that "Data in and of itself has no value! The only value data/information has to offer is in the context of the business processes, decisions, customer experiences, and competitive differentiators it can enable." [33] Today's buzzwords such as Business Activity monitoring (BAM), Business Operations management (BOM) and Business Process Intelligence (BPI) describe the need to more fine-grained techniques that can help improve healthcare processes with the many data available [52, 56, 73]. The only requirement for these systems – referred to as Business Process Management Systems or Workflow Management Systems (WfMS) – is that they record activities in so-called *event logs* [45]. The next step is to use these data to identify possible improvements of processes and efficiency, to help organizations cope with cutbacks and the increasing demand of care.

For this study we aim to use a subset of the many available data as described in the *DBC Information System (DIS)*³, provided to us by six different hospitals. The DIS collects both delivered and billed healthcare products from all healthcare providers (further described in Section 2.1.2) [20]. These event logs record non-trivial careflow processes, and contain huge numbers of distinct activities in countless different combinations. This makes analysis non-trivial and often results in spaghetti-like models.

New and powerful data mining and process mining tools and techniques are continuously being developed and implemented by researchers and software vendors. Most of the techniques make assumptions that do not hold in practical situations, and few of the more advanced techniques have been tested on real-life processes [56]. The goal is to describe some of these techniques and evaluate their applicability on standard DBC registration data. Continuing, we describe a methodology suitable for analyzing specific patient careflow and to provide insights in the medical activities performed for individual carepaths.

³National system that collects and maintains all the information on DBC's (Section 2.1.1.1)



1.2 Research statement

The general research objective for this thesis is defined as follows:

To explore advanced Process- and Data mining techniques, and to define a methodology that provides insight into patient careflow for specific DBC's in a hospital environment.

In the previous section we already stated that few of the available techniques have been applied to real-life cases. In Section 2.3, we dive deeper into the available literature on earlier exploratory research. As we aim to develop a general methodology that provides insight into the patient careflow based on empirical data⁴, our first step is to find answers to the following three questions:

1. Can data- and process mining techniques be applied to gain insight in patient careflow?
 - 1.1 Which data mining techniques are applicable?
 - 1.2 Which process mining techniques are applicable?

Basic statistics and simple analysis are insufficient to gain insights that medical professionals are able to work with. Not only is expert input an absolute requirement, we also need to quantify the results in a way that is understandable for both the analyst and medical expert. We need to establish the required parameters and statistics, on which we can develop our model. Also, we need to define a transparent way to present the resulting insights to the medical professionals.

Our second research question aims to identify parameters and variables, required for our model.

2. Which insights do we require to assess patient careflow?
 - 2.1 Which criteria (logistic/medical/cost) are used for the assessment?
 - 2.2 Which elements of a specific patient careflow can we use?
 - 2.3 Which parameters/techniques do we need to calculate quality for a specific carepath?
 - 2.4 Which elements of a carepath do we have available as input?

data mining is a subjective and iterative process, which requires both *statistical* and *domain specific* knowledge. The former extracts numbers and statistics from data, whilst the latter adds “meaning” to the results. Mining does not provide specific optimizations and objective results. In any case, it is a combination of interaction between the *analyst*, available *techniques* and *expert domain knowledge*.

The third question targets means to compare and evaluate our results.

3. How can we compare, evaluate and advise different carepaths?
 - 3.1 How do we visualize patient careflow?
 - 3.2 How do we define patient careflow quality? (What defines a *good* carepath or cluster?)
 - 3.3 How do we compare different cluster outcomes?

The goal of this thesis is not to define a standard protocol for healthcare processes, but rather to evaluate the applicability and value of mining techniques for classifying standard behavior. We aim to suggest a methodology that helps to indicate the deviation of existing protocols or the (statistical) possibility to introduce new standardized patient careflow.

⁴I.e. the recorded set of activities performed for a specific Diagnosis/Treatment combination

1.3 Research design

The previous section described the research statement for this project, which is a typical data mining project. This type of projects are thoroughly documented and supported by the *Cross Industry Standard Process for Data Mining (CRISP-DM)* as shown in Figure 1.1 [75]. CRISP-DM has been at the basis for various other healthcare projects, both in data mining and process mining [4,61]. Similarly, we apply this methodology to our project, as described in the following section. The last phase, Deployment, is not covered in this thesis, as software implementation and deployment is not part of this study.

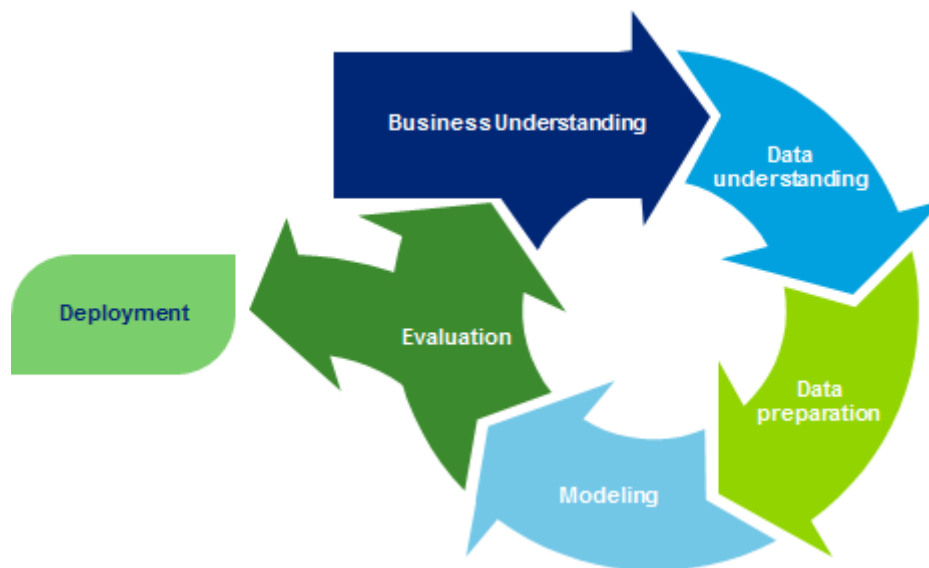


Figure 1.1: (Simplified) CRISP-DM model [75].

The understanding gained and decisions made in the different phases of CRISP-DM are supported by *experts*: a number of colleagues within Deloitte, who have years of consulting experience within hospitals and healthcare. Each of them has been actively involved in various projects regarding improving healthcare processes, and worked with both the DBC and the DOT systems. Together with a medical degree, a PhD in health economics, and a PhD in machine learning, this team offers a valuable and suitable source of expertise for this study.

1.3.1 Business Understanding

The first phase is to define the *business objectives* and a means to measure or quantify the analysis results. In Section 1.1 and 1.2, we already introduced our primary objectives which include the search for a quality measure. This section describes the initial plan to reach these objectives.

In Chapter 2, we provide a more thorough overview of the healthcare domain with its dynamics and complexity. A short literature study summarizes the results of previous data mining and process mining projects in healthcare, which provides a basis for the selection of a few specific techniques.

With the prerequisites in mind, we continue to the second phase: getting to understand the data.



1.3.2 Data Understanding

The second phase is concerned with the initial data collection. Before we try to analyze the data, it is important to understand the acquired data: which columns do we have for what scales, what do they mean, etc.

Quality assessment (or *Data Audit*) is also a big part of this phase: is the dataset complete, how do we cope with missing values, who entered the data (i.e. are typo's common?). Some relatively simple visualizations can help add insight to these data.

For this project, we tried to use a sample set from the official DIS as described in Section 2.1.2. Unfortunately, these data were not publicly available at the collection phase of this project. Instead we used similar datasets that are at our disposal, also containing DBC data as described in Section 4.2. This section also describes a thorough data audit.

Once we collected and defined the required data, we prepare the sets for analysis in the next phase.

1.3.3 Data Preparation

The third phase is preparing the raw data into a sample set that can be used for modeling. This contains every step necessary to prepare the final analysis set from the raw initial data. Example steps are: selecting the right tables and records, merging multiple sets, cleaning missing data and transforming the overall structure, as the required structure may depend on the type of analysis.

Together with Data Understanding, this phase often takes up the largest part of any data mining project. Since we are working with sample sets instead of an extract from the official system, we have to go through a number of cleaning and validation steps. Also, we need multiple transformation steps to prepare the final sets since we are working with two different types of analysis that require different data structures. The entire preparation phase is described in Section 4.3.

Once we have the sets available, we start our actual modeling and analysis.

1.3.4 Modeling

The fourth phase describes the actual modeling of the data. In the modeling phase we try a range of techniques to develop models that provide insight into the patient careflow. The different model results are assessed, compared, and sometimes even combined.

Section 4.4 describes the application of the techniques explained in Chapter 3. We present the steps required to build the models and apply the techniques to a number of sample care products.

The final step of this phase is an addition to the steps described by CRISP-DM: we summarize the steps performed in the previous phases in a methodology, describing only the most applicable techniques. This methodology can be used to gain insight into any type of patient careflow, and is evaluated in the second-last step of the CRISP-DM circle.

1.3.5 Evaluation (Case Study)

The fifth phase before *deployment* is to validate and evaluate the methodology and models obtained in the previous phase, which requires expert domain knowledge. At this phase, merely statistics are

insufficient and human input is required for a thorough model assessment. It is the final check before Deployment, and consists of two main themes:

1. Evaluation/validation from a *Technical* perspective.
2. Evaluation/validation from a *Business* perspective, it is crucial to validate whether the model contributes to accomplishing the Business Objectives stated in the first phase.

In Chapter 5, we assess the final methodology by performing a number of Case Studies. We show that we have developed a methodology that provides valuable insights, that is (widely) applicable to identify improvements in patient careflow.

1.4 Outline

After this introduction, we first give more detail on the healthcare domain and mining in general (Chapter 2). This chapter also elaborates on previous work performed in this area. Specifics on the data mining techniques used in this project are explained in Chapter 3. Chapter 4 describes the development of our methodology according to the CRISP-DM model. The methodology is then tested on a number of Case Studies in Chapter 5. The project is evaluated in Chapter 6 before we conclude in Chapter 6.6.



Chapter 2

Theoretical background

This chapter offers background information on the research area of this project, which is required for a deep understanding of the business objectives and techniques applied in this study. We start with a general overview of the healthcare domain in Section 2.1. A thorough description of the dynamics and complexity of this domain points out the need for advanced analysis techniques. The applicability of these techniques is supported by huge amounts of data readily available from many different Information Systems. For the purpose of this project, we restrict the data set to the standard administration structure of DBC's (and in the future DOT) which is built to describe patient careflow. Next, we provide a general introduction to data mining and process mining in Section 2.2, which explains how these techniques can be applied to provide useful insights. Section 2.3 provides a short overview of related work on mining in the healthcare domain. Based on these earlier results, we select a number of algorithms and techniques as a basis for our methodology design.

2.1 Healthcare domain

Our main goal is to gain insight into patient careflow, which is a far from trivial task in a healthcare environment [5]. As shown in Figure 2.1, there are a number of main types of *healthcare processes* [43, 47]. The *medical treatment processes* describe the diagnostic-therapeutic cycle – patient observation,

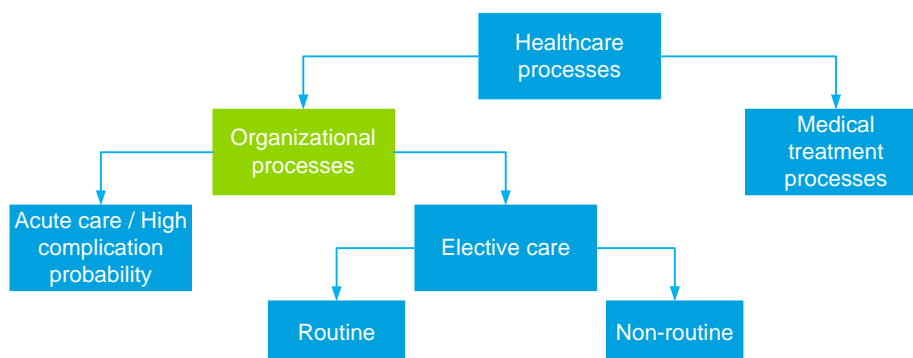


Figure 2.1: Overview of the main types of healthcare processes.

medical reasoning and decision making. We put our focus towards the *organizational processes*, which include inter-operating healthcare professionals, disciplines and departments. In contrast to the medical processes, these processes are not about the medical content of individual paths, but rather about the logistics of work processes. The difficulty is that these type of processes are intertwined with each other and directly linked to the individual patient's care and requirements [5, 38, 43, 47].

One would argue that each individual patient is unique, in contrast to the highly standardized products in e.g. car manufacturing. It is important to mention that healthcare processes are *case-driven*, as each execution of a step can be attributed to exactly one specific case. As a consequence, each instance of the process varies in the way it is executed [43]. Often used terms to characterize healthcare processes are *dynamic* and *flexible* [28, 46]. There are a number of characteristics that distinguish healthcare from other domains. This section gives an overview of the main characteristics.

Dynamic, flexible and complex Medical knowledge is continuously evolving: not only the development of new treatments and diagnostic procedures, but also the discovery of new diseases result in constant changes in healthcare processes. Other causes of change are e.g. the implementation of new (information) systems – both administrative and operational – or the discovery of new drugs. Because each patient is unique, each treatment path can be complex because each individual might respond differently to certain types of treatment. Even a relatively simple treatment for one diagnosis can become complex when treatment is required for another diagnosis due to complications. Therefore, strong flexibility within these processes is required [5, 28, 36].

Ad hoc actions and process changes Similar to the above, ad hoc changes are sometimes required to deal with a patient's unique complications. Following standardized carepaths comes second to saving a human life, physicians act according to their knowledge and experience. Therefore not every principle of general operations management is applicable, sometimes physicians are required to deviate from the standardized path [5, 36].

Multi-disciplinary and cross-functional The treatment of a single patient involves procedures from many disciplines (e.g. management, IT, etc.) and specialized departments (e.g. radiology, cardiology, etc.) within a healthcare organization. It is also directly linked to financial tasks [5, 53].

Automation issues In medical decision making, experts and physicians perform trade-offs, decisions and actions based on their knowledge and experience. It is not (yet) possible to automate these steps in systems such as a Decision Support System. Instead, the degree of collaboration among humans and automated systems plays a crucial role in delivering high quality services to patients [28].

Data management issues Every department often has its own Information System or application, which records data about individual patients. The data from these various applications is often redundant and not linked to other processes and applications. This results in redundant, inaccurate, uninformative and even confusing data storage [28].

Classification issues To analyze healthcare processes, it is important to classify patients and processes in order to define what constitutes such a process. Several techniques have been suggested:



1. Four levels of process classification: *type of care* (acute versus elective), *complication probability* (high versus low), *complexity of care* (high versus low) and *whether or not the diagnosis is known* [44].
2. A type of ISO-process grouping, which provides homogeneous groups in terms of costs and the process described by the carepaths [70].
3. The Dutch reimbursement systems DBC and DOT, which have defined patient careflow for similar DBC's [74]. These systems contain a lot of detail on the medical process: they describe specific medical activities for each care product. This is possibly a big obstacle for process optimization; despite the distinction of medical content for activities, the logistic processes they describe can be similar [36].

Healthcare processes are less structured because of these characteristics and issues. Therefore “it is not known what happens in a healthcare process for a group of patients with the same diagnosis” [46]. Hospitals are searching for reliable techniques that provide valuable insights of these processes and the variety of executions. The techniques must offer easily obtainable results in an interactive way, such that useful insights can be concluded from them [55].

Before we continue to make any type of analysis, we need to define what exactly we mean by patient careflow. The DBC and DOT systems are at the basis of reimbursement for provided patient treatment in healthcare organizations. Every hospital stores data according to a predefined data structure, in order to be able to make declarations. These datasets are easily combined in order to analyze multiple hospitals simultaneously – on a national level, this is done by the DIS (Section 2.1.2). For this reason, we base our analysis on these standardized and readily available datasets, using a definition of patient careflow based on the existing DBC and DOT systems.

In the next section, we give a thorough overview of how the reimbursement systems DBC and DOT are designed. With this in mind, we propose a definition of patient careflow used in our analysis.

2.1.1 Defining patient careflow

The old reimbursement model used in the Netherlands distinguished four budget components: *housing*, *availability*, *capacity* and *production*. This resulted in a fixed budget for the various components. The model was unable to cope with the dynamic and complex nature of the healthcare domain as described earlier, which led hospitals to frequently exceed their provided budget [29]. In order to establish a transparent financing system for healthcare organizations, a new reimbursement model, the DBC system, was introduced. Although the system offered many improvements, not everything turned out as anticipated [25]. The recently introduced system DOT aims to cope with these limitations [29, 74].

The DBC and DOT systems aim to provide both medical and financial homogeneity in careflow groups or clusters. To reach this goal, a number of generalized groups of Diagnosis/Treatment combinations are specified, describing patient careflow [74]. The first part of this section focuses on the developed methodologies and techniques behind these two systems. We use these methodologies and techniques as a starting point for our analysis. In the last part of this section, we propose our definition of patient careflow which is based on assumptions and definitions as described in DBC and DOT.

2.1.1.1 DBC

In 2005, the Dutch government decided to introduce the DBC system to provide more transparency by forging a stronger link between funding and performance. This system is an example of a case-based funding system¹, consisting of predefined average care profiles that describe carepaths for specific diagnoses. These profiles have established prices that are used to calculate the fees hospitals receive for their delivered care services [19, 25, 29, 74].

During the development of the DBC system, administrative differences (i.e. the manner in which they allocate cost to activities) between hospitals were ignored. The goal was to find groups of DBC's with homogeneous care profiles to provide a manageable product structure [74]. The level of cost homogeneity also played an important role, as cost inhomogeneity leads to financial risks for hospitals [8]. The first clustering of DBC's is based on statistical data analyses, which is later refined by scientific committees based on their medical judgment [74].

The DBC system stimulates hospitals to increase efficiency for individual careflows, because reimbursement is based on a fixed price per DBC. A part of these prices – the *A-segment* – is regulated by the Dutch government. However, prices for about 70 percent of all DBC's – the *B-segment* – are freely negotiable between hospitals and insurers, which allows for managed competition between hospitals. This stimulates hospitals to provide the best quality care for the lowest cost [29, 55, 59, 74].

A DBC-code can be viewed as an abstraction of what is inside such a predefined care profile. Each code describes the sequence of medical activities for the entire path a patient goes through from the diagnosis of a problem, to the treatment of the problem, to the final discharge [29, 74]. A DBC consists of four attributes as shown in Table 2.1.

Table 2.1: DBC-code structure, each code consists of four attributes. This is an example for 0305.11.1701.0223: arthrosis (hip) - surgical/clinical with joint prosthesis.

DBC-code	Definition	Description	Example
0305-..-....-....	specialism	department or discipline	orthopedics
....-11-....-....	care type	acute, regular or follow-up	regular
....-..-1701-....	diagnosis		arthrosis: pelvic/hip/thigh
....-..-....-0223	treatment	conservative or surgical, ambulatory or clinical	surgical, clinical episode with joint prosthesis

Not only the DBC's, also the activities they describe are directly linked to a specific specialism. Each DBC or care product describes a number of specific activities. An example is given in Table 2.2a for three outpatient department visits, five nursingdays, surgery, hip imaging and six labtests. The ordering is not given in the care products The Dutch Healthcare Authority, previously known as the "College Tarieven Gezondheidszorg", is responsible for determining the tariff for each individual treatment activity. These activities are represented by a CTG-code, and classified by an Activity Class (ZPK). ZPK-codes describe the type of activity (e.g. diagnostic or surgical) and are equal for all specialisms. This allows the comparison between two carepaths to include similarity measures between two distinct activities (e.g. two diagnostic activities are more similar than a diagnostic and a surgical activity) [74]. A subset

¹Note that the Dutch DBC is similar to DRG. The variations between the two systems mainly lay in starting point and intention, build-up structure, and in financial incentive [29].



of ZPK-codes is listed in Table 2.2b.

Table 2.2: Examples for CTG-codes and ZPK-codes.

(a) Overview of the activities described by a DBC-code (this is an example, not an official extract). (b) Overview ZPK-codes 1...8 (see Table G.1 for the complete overview).

CTG	ZPK	Description	#	ZPK	Definition
190011	1	First outpatient visit	1	1	Outpatient department / ER
190013	1	Outpatient department	2	2	Daycare
190204	3	Nursingday	5	3	Clinic
38567	5	Replacement of the hip	1	4	Diagnostic activities
89202	7	Imaging for hip	1	5	Surgical activities
70611	8	Antibodies	1	6	Other therapeutic activities
70702	8	Hemoglobin	2	7	Medical imaging
79991	8	Laboratory research	3	8	Chemistry / hematology (labtest)

Although the introduction of the DBC system did offer more insight into pricing, content and quality of care, there are still areas of improvement. With around 30.000 different DBC's², the final product set is quite large, which reduces the overall transparency. Because this system has defined its set of unique DBC's per specialism, it is possible that one specific care profile is defined more than once in different specialisms. Furthermore, the large size of the product set and reduced transparency allows organizations to manipulate the resulting reimbursement, which makes the system more fraud-sensitive [29, 31]. The newly introduced DOT system transcends individual specialisms and decreased the total set of care-products [74].

2.1.1.2 DOT

The basic idea behind the DOT system is similar to that of DBC: to provide both medical and financial homogeneity in patient careflow. Although DOT is largely based on DBC, there are three main points of improvement:

Based on ICD-10 DBC diagnosis classification systems are unique systems developed by separate medical specialist communities, and therefore incomparable amongst specialisms. Instead, care products in DOT follow the ICD10 (International Statistical Classification of Diseases and Related Health Problems) system. This allows the comparison of performance data on both a *local* and *international* level, since every specialism – both local and international – uses the same coding system [29].

Reduction of products An important and effective difference is a serious reduction of over 30.000 DBC's into only 4400 care products [19, 74]. As opposed to the previous system, products are now defined in a *specialism transcending* fashion³, where medically equivalent products from different specialties are combined into one single product. This reduction offers a large increase in uniformity amongst different specialties and departments, which supports negotiations between

²Theoretically, this number can be much higher.

³In Dutch: specialisme overstijgend.

care providers and health insurers. Instructions for all specialties will be uniform and – similar to DBC and DRG systems – the compensation per product will be case-mix based [29, 74]. Note that some specialties are yet to be included in this new system.

Grouper Another problem with the previous system is that the validation process does not offer a strict or uniform system due to its high *administrative complexity*. This causes risk of *upcoding* and *overdeclaration*. In DOT, care products are deduced from registered activities by a web-based *grouper*, as opposed to being selected and validated by medical specialists. This system is intended to prevent upcoding, whilst increasing registration quality by alleviating the administrative burden [29, 54, 68].

The grouper is a central, online system that automatically deduces care products and add-ons from registered diagnosis and treatment activities. The specification of such a product is done according to a binary tree-like structure: stepwise choices or decision rules determine which specific care product is provided. These rules are based on registered activities only, which makes it important to register every performed activity [19].

Some argue that although an automated system is now responsible to select the appropriate price for DOT, the resulting care products offer less transparency than the original DBC's [29]. The DBC system offers a detailed report of diagnosis and treatment details, whereas the DOT system is a simplification of the myriad of products [19]. An important recommendation is that the system needs permanent adjustment in order to avoid future financial risk for hospitals [74].

The advantage of the DBC system is that it offers a greater level of detail, where DOT has improved homogeneity overall and allows comparison between different specialties. The lack of available DOT data however prevents us from statistic comparison of the different systems. In general, however, the patients and their carepaths are the same, and the methodology developed in this study is applicable to any type of reimbursement system. The next section gives an overview of our definition of patient careflow which is based on these systems, and lies at the basis of our analysis.

2.1.1.3 Careflow definition

Previously, we stated that patient careflow – or a clinical pathway – describes a group of similar care profiles, i.e. a number of medical treatments and activities. The large number of process variations in a healthcare environment make it difficult to analyze and compare performance between patients, departments and hospitals, without having a rigid patient careflow defined [8, 35, 74]. This system is widely adopted by large hospitals around the world, with the goal to decrease costs whilst maintaining or improving quality. They are designed to organize care activities, reduce use of resources and variations in practice, minimize treatment delays and reduce the length of hospital stay for individual patients. Collaboration between all kinds of hospital staff members is required for the design of patient careflow, whereas an individual patient is able to cause variation in the execution of the clinical pathway. In order to improve clinical pathway performance and homogeneity, a more dynamic and adaptive process is required [58]. The level of homogeneity in patient careflow can be described on three main levels [74]:

Logistics This level describes patient careflow on a higher level, focusing on the logistics and organizational process a patient undergoes during treatment. Individual activities are described by their ZPK, and detailed medical content is ignored (e.g. all surgical activities can be covered



by a single ZPK, whilst the individual activities show significant differences in cost, duration and intensity).

Homogeneity is measured based on throughput time, number of (similar) activities and the order in which they are performed. This allows the comparison of different DBC's with relatively distinct types of surgery, but overall similar care profiles.

Medical At this level, medical content such as the detailed description of individual activities performed is important. E.g. surgical activities have a big effect on determining the similarity between two carepaths.

For any type of analysis on this level, thorough medical knowledge is required. Statistic results do not provide sufficient insight in the similarity between medical content, which offers limited quality in clustering.

Financial Creating financially homogeneous groups of DBC's is a trivial task: the myriad of DBC's can be divided into *cheap* and *expensive* groups without any advanced analysis simply by looking at the total averaged cost. When a cluster of DBC's shows neither medical nor logistics homogeneity, this often implicates a large spread in cost. More interesting is using cost as either a fine-tuning mechanism, or as a verification of the clustering quality [74].

Earlier, we described the difference between *medical* and *organizational* processes (Figure 2.1). As is obvious, the medical content of a carepath defines the medical process, whereas the logistics describe a part of the organizational process. The financial label of a carepath is linked to both the medical and logistics process. We put our focus on the organizational process (logistics) within patient careflow, because our medical knowledge is limited and logistic process offers support for statistical analysis. In future work, where analytical knowledge is combined with more extensive medical domain knowledge, these results also offer a good starting point to include the analysis of the medical content.

Now we know what type of process described by patient careflow we are targeting, we continue to look at the available elements. Using the DBC system, a uniform set of codes (CTG) describe all types of treatments and activities. These individual treatments are already classified by a ZPK-code. As stated before, the DBC system contains a lot of detail, but on a ZPK-level there are only 24 distinct activities. This simplifies the number of possible descriptions for carepaths, which has proven to be a successful approach [43,46]. Experts have pointed out that for most DBC's only ZPK-classes 1 through 8 are significant, as the remaining classes do not describe activities affecting the logistic process or total costs.⁴ This gives us the possibility to use a restricted set of distinct activities and further simplify the representation of a carepath.

Another important aspect is whether to include the order in which activities are performed. A good measure to group different carepaths is comparing the activity content, i.e. the number of activities performed. When we include the ordering of activities, this can result in a different grouping of carepaths, because the similarity of subsequences can outweigh the difference in number of activities. One of our goals is to find out which representation is the most qualified, or whether a combination of representations is required for an optimal result. Therefore we propose three different representations for patient careflow as input for our analysis (examples are given in Table 4.2):

⁴ZPK 13 describes the use of prosthetic implants, which are in fact expensive and do have an effect the total costs for a specific path. However, for the purpose of this study, we assume that the costs are equal for each DBC case.

1. A string representation of the ordered sequences of activities on a ZPK-level.

Limited to a subset of ZPK-codes, this provides a high-level though extensive description of the sequence of activities performed.

2. Counts of the performed activities on a ZPK-level.

Limited to a subset of ZPK-codes, this is a total count of the number of performed activities per code and gives a high-level overview of the content of a carepath.

3. Counts of all performed activities on a CTG-level.

Since this level contains too much detail, this is purely for visualization purposes (see Figure 4.1b).

In the next step we review the available data and discuss how we can build these representations.

2.1.2 Hospital Information System

In Dutch hospitals a variety of HIS and other embedded systems are found, each of which record huge amounts of data that track e.g. device usage or performed treatment activities. Due to the dynamic and ever evolving nature of the healthcare environment, existing HIS must be periodically adapted to the current situation. The systems are often large and complex, and the amount of resources hospitals invest in the development of these systems is significant [53].

In order to track patient careflow, we need data that describe the activities performed in each execution. This type of data is generally collected by a WfMS. Such a system can be defined as follows: “A system that defines, creates and manages the execution of workflows through the use of software, running on one or more workflow engines, which is able to interpret the process definition, interact with workflow participants and, where required, invoke the use of IT tools and applications” [77]. Unfortunately, the data models in HIS and other types of WfMS often differ between hospitals and even departments, which makes it difficult to define one standard tool that can cope with any system of any hospital or specialism. Besides, in general these systems lack maturity and interoperability with other systems [5].

In the Netherlands a national system is available, the DIS, which is responsible for gathering all DBC data. It contains information from the HIS of all types of healthcare organizations, describing the provided and declared care. DIS is responsible for the data-exchange with a number of legal institutions (NZa, CBS, CVZ, Ministry of Health and DBC Onderhoud), and is – per request – allowed to share data with third-parties for e.g. research purposes [19]. This relieves the administrative burden of care providers and provides one single dataset where all DBC's are gathered. A summary of the available data is shown in Table 2.3 [18].

The level of detail registered by DIS is limited. For example, DIS only records the start-date for a performed activity. Specific timestamps for both start and end of an activity are not included. This limits the possibilities for our analysis, as the ordering of activities is limited to a day-to-day level and the duration of performed activities is unknown. As a consequence, we are forced to ignore the duration and specific ordering of activities on a single day, and we combine similar activities on a single day into a single activity.

This further simplifies the resulting representation, since numerous activities are represented by a single activity. On a logistics level this makes sense, because most activities do not require extra



Table 2.3: Four tables from the DIS – only a limited number of columns are provided (e.g. ZPK-codes are derived from different source). Each table contains a unique primary key (PK), such that they can be linked to each other by a foreign key (FK).

Patients		Carepath		Subcarepath		Activity	
PK	PatientID	PK	PathID	PK	SubpathID	PK	ActivityID
	Name		PathStart	FK	PathID	FK	SubpathID
	Birthyear		PathEnd		AGB		CTG-code
	Sex		Specialism		Diagnosis		Timestamp
		FK	PatientID		Treatment		Amount

preparations or actions to repeat the activity. E.g. one lab test is similar to doing numerous tests: one needs to take blood and send it to the lab. The same goes for surgery, assuming that the operations are combined in one OR-session. Other ZPK-levels such as “Clinic” representing a single nursingday are already limited to one single activity per day by design. For more problematic cases, we might lose important detail, e.g. when complications arise after surgery and a patient has to undergo surgery for a second time. However, this does not outweigh the advantage of the increase in clarity the new representation offers to the overall logistics process.

Now that we have our definition of patient careflow and know where to get the data, we look into what we can do with the data. First, we give an overview of the fields of both data mining and process mining, which gives a general impression of the techniques available. However, before we perform any type of analytical analysis, we need to prepare the dataset and validate the quality of the acquired data.

2.2 Data tasks

There is a variety of techniques available in data-driven analysis, capable of uncovering patterns and paths, and providing many other valuable insights. These techniques can be subdivided in *data mining* and *process mining*, of which many have been applied to various types of HIS before [61]. The next section gives an overview of these techniques in general. Data in itself has no value, but since we now have many means to record it, we can use it for the analysis of e.g. healthcare processes and types of patients. However, data-driven analysis only works if the input data is of sufficient quality; therefore we also dedicate a part of this section to *data quality and preparation*.

2.2.1 Data Mining



Figure 2.2: The process of Knowledge Discovery in Databases.

Data mining is the analysis step of the “Knowledge Discovery in Databases” (KDD) process (Figure 2.2); the non-trivial process of identifying valid, novel, potentially useful, and ultimately understand-

able patterns in data [23, 66]. It commonly refers to the computer-based methodology for uncovering patterns, and can be characterized as the extraction of implicit and potentially valuable information from datasets in a programmatically automated manner [76]. CRISP-DM describes similar steps to KDD in the phases on data understanding, preprocessing and modeling, but is more elaborate as it also has a large focus on other area's such as identifying business objectives.

Conventionally, data is analyzed manually, but an analyst may fail to find many hidden and potentially useful relationships and patterns [40]. Data miners often work with readily available data, because it is often impossible or too expensive (in both costs and time) to gather specific data. An important difference between data mining and the area of *experimental* or *statistical* analysis is that data mining focuses on the applicability of improvements and insights, where experimental or statistical analysis requires accurate proof and confidence intervals. In other words, data mining is relatively less concerned with specific relations between variables, and focuses on producing a solution that can generate valuable predictions or insights [62]. It is a topic that involves *learning* in a practical, non-theoretical sense. The two main types of learning commonly described are *supervised* and *unsupervised* learning.

Supervised learning is used when the resulting values are known a priori, and the dataset is used to build a model that can help predict values for new data points or observations [66]; a form of predictive modeling.

A well-known example of this is *Classification*: the attributes of data points in the original set are matched against predefined classes. The resulting model is then able to predict the class for new observations based on their attributes. The goal here is to understand the basis for the classification [61, 66].

Unsupervised learning is used when classes or values are unknown a priori, and the goal is to discover classes or patterns hidden in the data [66]; a form of descriptive modeling.

A well-known example of this is *Clustering*, where data points are clustered in groups of points with similar or related attributes. For the assignment of a data point to a cluster we look at the distance between individual data points, and the distance between an individual data point and a cluster of points. There are many different ways of calculating the distance between two points available, e.g. the *Euclidian distance*, each of which has its own characteristics. Also, clusters can be represented by different values, e.g. an mean value or a *medoid*. These variations have an effect on the resulting clustering [66].

Both learning types are useful tools for this project: we aim to identify standard patient careflow, which we hope to achieve by clustering a large collection of different carepaths. Each cluster describes a specific set of characteristics for the carepaths it contains. Once the clusters are generated, classification helps identifying the key characteristics that make up each cluster. The implementation and execution of the different algorithms is performed using R, an open source tool for statistical computing [26].

2.2.2 Process Mining

Process mining is a combination of data mining and *process modeling* (see Figure 2.3), with the goal to *discover*, *monitor* and *improve* real processes by extracting knowledge from case-based event logs [21, 52].

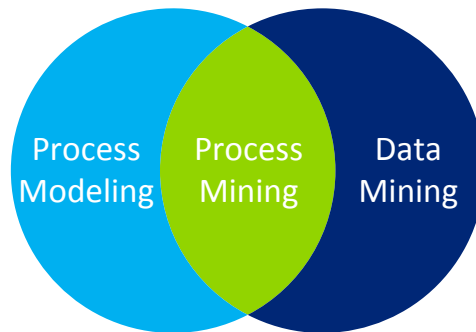


Figure 2.3: Process mining is a combination of process modeling and data mining.

One way to compare and improve patient careflow is to put medical specialists together in workshops to discuss similarities and possible improvements. This is a time-consuming and often suboptimal solution [55, 68, 74]. Instead, process mining offers a range of techniques to automate this process [21]:

Process discovery extracts or discovers process models from an event log, by inferring the ordering relations between the various recorded tasks. It offers a tool to find out how people and/or procedures really work.

Delta analysis or *conformance checking* verifies whether the real process matches an a-priori process model. Process mining offers a tool to monitor deviations by comparing the actual process as recorded in the event log with the intended process.

Process mining is not just about process discovery and improvement, and has an extensive variety of algorithms implemented in the ProM framework, an open source process mining tool [1, 21, 46, 52, 69].

The only strict requirement for process mining is the availability of a suitable *event log*, which lists all events executed on a certain case. These transaction-like logs are often collected by WfMS and other information systems (e.g. a HIS – Section 2.1.2). The definition and an example of a suitable event log is described by Table 2.4 [21].

We distinguish three *perspectives* on the event log, each of which focuses on other elements in the patient careflow. These perspectives represent the “*How?*” (process or control-flow), “*Who?*” (organizational) and “*What?*” (case or data) questions [2, 48, 56, 72].

Process or Control-flow focuses on the ordering of activities. The key elements are *process instances* (in this case: individual carepaths). The goal is to identify a good characterization of all possible paths.

Organizational perspective focuses on the performers or originators involved, and how they are related. Either these performers are classified in terms of roles and organizational units, or the relation between individual performers is shown.

Case or Data perspective focuses on properties of cases, i.e. the path of a case in the process, the originators working on a case, or values of the corresponding data elements. This perspective requires more detailed information such as extra data attributes of a case.

Table 2.4: Overview of a suitable event log, (2.4a) shows the events recorded for two cases, with activities ordered by date. (2.4a) shows the properties a log should comply to.

(a) Example log					(b) List of properties
Case	Event	Date	Activity	Cost	
1	23	25-03-2010	Triage	50	1. Each event refers to an <i>activity</i> .
1	24	25-03-2010	X-Ray	80	2. Each event refers to a <i>case</i> .
1	25	26-03-2011	Treatment 1	500	3. Each event can have a <i>performer</i> or <i>originator</i> .
4	28	26-03-2011	Triage	50	4. Events have a <i>timestamp</i> and are totally ordered.
1	29	26-03-2011	Treatment 2	550	5. Some logs contain more information on the case itself (e.g. age, sex and diagnosis).
4	31	26-03-2011	Treatment 1	500	
1	33	26-03-2011	Clinic	300	
4	37	26-03-2011	Clinic	300	
4	41	27-03-2011	Treatment 3	450	
4	42	27-03-2011	Clinic	300	

2.2.3 Data quality

Due to the increase of data-based methods, data quality is an increasingly important topic as bad data quality will “deteriorate the process of discovering patterns, relationships and structures when applying data mining” [7]. “If you don’t have the data, decisions can’t be made (by definition), and if decisions can’t be made, the organization cannot create value. So there is also an ‘opportunity cost’ associated with non-existent or bad data” [51].

In short, this is described by the *garbage in – garbage out* principle: good inputs generally result in good outputs, whilst bad inputs generally result in bad outputs [41]. Input problems come in three flavors:

Type problems These problems occur when the incorrect type of input is fed into a system, such as entering the age of a patient into the sex field. Although often easy to detect, this represents the maximum form of garbage if undetected.

Quality problems When the system is fed the correct type with wrong details, such as entering the wrong activity treatment number in the correct field, this is considered a quality problem. A minor typo in the activity number may merely have small financial consequences, whereas it may also have huge medical consequences.

Missing values A common problem in datasets is missing data, and can occur due to a number of reasons: the data may be unavailable, the person entering the data could have skipped a field by accident or the field may not be applicable. It depends on the context whether e.g. the rows containing missing values should be ignored or mean values should be used instead.

The DIS collects data from all hospitals and other healthcare organizations in the Netherlands [20]. The first risk lies with the data entry, as any system that requires human input is prone to human error [17]. These *type problems* are usually prevented by restricted drop-down boxes in the different HIS, and the DIS offers an extensive validation on data import. Another risk is that different hospitals



register activities in a different way than others. Note that this may not always be a conscious choice, since each hospital has a variety of HIS and other expert systems – each of which uses their own event log format (see Section 2.1.2). Often, the information required for data mining and other analyses is hidden in a data structure which is designed for correct and efficient storage. Extracting the right information from this data is a time-consuming task and requires domain knowledge, as the choices made during extraction influence the analysis results [11]. Section 4.3 tries to cope with these Quality problems.

For data mining it is important to define useful data types for the available attributes (e.g. regression is unable to cope with nominal target values, similarly it is impossible to classify continuous variables without binning them) [66]. Process mining treats all activities as equal, therefore it is important to have a uniform and consistent level of detail of events [11]. Practical applications require a number of preprocessing steps [24]. In Section 4.3 we apply a number of visualization techniques to gain insight into the data quality, and apply the necessary steps to remove unwanted garbage.

2.3 Related work

This section describes related work performed in the healthcare domain. The described literature helps to find an answer to our first research question: *Can data- and process mining techniques be applied to gain insight in patient careflow?* Based on the literature described below, we select the best possible algorithms applicable for this study. This section is split in two parts: data mining and process mining.

2.3.1 Data Mining

The *patient classification problem* is the grouping of patients with similar characteristics (e.g. medical history, diagnosis or carepath). These groups can be used to increase the predictability of carepath patterns and length of stay for patients [32], which supports carepaths standardization and allows for improved resource utilization. The next section provides an overview of previous studies that have tried to tackle the patient classification problem using a variety of techniques including clustering, classification and association rules.

Clustering Two well-known clustering algorithms applied to healthcare processes are *K-means* [32] and *hierarchical clustering* [74], which are applied to group patients with similar carepaths. Alternatively, Self Organizing Maps offer an intuitive and visual clustering technique [15, 43, 46], whereas Sequence Clustering offers an interesting technique for the clustering of similar sub-sequences [5].

The developers of the original DBC reimbursement system explored and applied a number of standard clustering techniques and defined an initial grouping of DBC's providing a statistically sound product structure. The resulting algorithm consists of two steps to identify the clinical pathways: first, clusters of DBC's with similar care profiles *within a given activity class* are identified using Hierarchical clustering on the Tanimoto distance. Secondly, clinical pathways are identified by analyzing the similarities in these sequences. The resulting care profiles were refined by a scientific committee, based on their medical judgment [74].

Compression techniques have proved useful in many domains like DNA and texture clustering. The application of compression benefits many unsupervised learning techniques, as compression techniques try to describe objects such as carepaths by the smallest representation possible. Given an initial carepath, we can then calculate the minimum number of changes we need to make to transform it into a second path [13,34,49]. The smaller the number of changes, the more two carepaths are alike. In this context, well-known compression algorithms can lead to elegant, parameter-free solutions to clustering and distance-function design [22].

Classification A variety of classification techniques such as Random Forest, CART Tree, Adaboost.M1, Naive Bayes and K-Nearest Neighbor have been evaluated by [68]. Especially the CART Tree appears to be a popular technique, recurring in the classification of Emergency Department patients [15] and for general Business Process Management [16,61].

Association rules With the potential to provide knowledge in form of recurring event patterns in patient careflow, this technique offers an easy to interpret set of rules describing the event log [28, 40,68]. Alternatively, association rules have been applied to provide more insight into the relation between carepaths and the diagnosis [58]. However, the unstructured properties of healthcare processes limit the quality of results in recent studies and test cases [5,28,40,58,68].

2.3.2 Process Mining

The first process mining application on WfMS is found in [3]. Ever since, a lot of work has been performed in search of a successful process mining technique for the healthcare domain [4]. Some say that process mining is able to provide insights where data mining does not [16]. This section gives a short overview of the different techniques available, focusing on the *process perspective*.

Popular example techniques for process discovery are e.g. the Heuristics Miner [72], the Fuzzy Miner [27], and the Genetic Miner [48]. The Heuristics Miner is able to separate the main behavior from noise in event logs and has proven to be an insightful tool in the healthcare domain [28,55,59,61,72]. However, the resulting process is often still spaghetti-like and difficult to understand [43,45,46]. The Fuzzy Miner has many options that help to aggregate and simplify the event log, but this can hide important detail within a healthcare process [43,46,55,59]. For non-free-choice activities results show that the Genetic Miner is best in dealing with parallelism and invisible tasks [35,55,59].

A variety of other techniques is also available. For the discovery of patterns within carepaths, time dependency is found to be one of the major factors using the Time Dependency plug-in [58]. Alternatively, a process mining implementation for association rule mining is built into the Association Rule Miner in ProM. This algorithm has proven to be useful to present behavioral patterns in the form of statements rather than models, but still has a number of limitations [5,28].

A technique complementary to existing data mining techniques is offered in the Trace Alignment plugin [9,10,39]. Although, to our knowledge, this specific technique has not been applied in the healthcare domain, we clearly recognize its potential in our search for groups of similar carepaths within a specific DBC.

The research described in [43,46] also covers work on the *organizational perspective* in healthcare processes. For the *performance perspective*, tools such as Visual Analytics [55,59] and the Dotted Chart Analysis [43,46,61] have proven to offer valuable insights [5].



2.4 Conclusion

Healthcare processes are dynamic and complex, as these processes are described by the activities required for the treatment of individual patients. Unique complications may require ad-hoc actions, which are often difficult to organize in the cross-functional setting of a hospital. The Dutch reimbursement systems DBC and DOT both applied retrospective analysis on the sequence of medical activities for the entire path patients to define care products (DBC's). Although these systems encourage standardization of patient careflow, in reality it is not known what happens for a group of patients with the same diagnosis. In order to gain insight into the myriad of different carepaths, hospitals are searching for reliable techniques that provide easily obtained insights in an interactive way. For the purpose of this study, we focus on the logistic process of individual paths. This allows for a higher level of abstraction during the analysis, whilst providing solid grounds for process improvement.

The research mentioned in the previous sections has pointed out the high potential for numerous data mining and process mining techniques. However, due to the complexity of healthcare processes there are still a number of limitations to these techniques. Based on the wide variety of clustering techniques and measures described in the literature, we conclude that traditional data mining techniques still offer the highest level of flexibility for the clustering of patients, but additional techniques are required to gain the required insights.

Clustering The success of the original DBC system as described in [74] supports our decision to explore techniques such as hierarchical clustering and partitioning around medoids, using different distance functions. On the other hand, Self Organizing Maps appear to offer less insight and flexibility of the clustering characteristics or attributes, therefore it will not be covered by this study.

Classification Since we are trying to gain insight into patient careflow and the clusters we create to describe different care profiles, we also explore the CART-algorithm, which has proven to be a popular tool for this exact purpose [15]. It allows us to identify the specific characteristics of each cluster.

Process mining The process mining techniques described in the literature are useful to identify and visualize the actual process, e.g. by applying the Heuristics Miner [43, 46]. As our goal is to use define different profiles within a specific care product, we are also interested in the visualization capabilities of the Trace Clustering algorithm. This algorithm has shown to offer great insight into both small and large process deviations, and can offer a lot of insight into specific differences and similarities (i.e. a standard careflow) within multiple carepaths.

For the development of our models, we use R and ProM. R offers a high level of flexibility and an extensive number of analytical algorithms and visualization techniques. A variety of packages is available in standard libraries in R, including numerous traditional mining techniques. Combined with successful previous experiences and extensive available documentation, R offers a suitable environment for the development of our clustering and classification techniques. The literature describing process mining was mainly based on ProM. We follow the literature by using ProM, as this software package offers plugins for both the Heuristics Miner and the Trace Alignment algorithm. These techniques are described in more detail in the next chapter.



Chapter 3

Technical background

In the previous section we have seen that we have a wide variety of both data mining and process mining techniques available, each of which has its own specific pros and cons. For the purpose of this study, we selected a number of potentially useful algorithms and measures: a number of clustering and classification techniques implemented in R, and some process mining techniques implemented in ProM. The next section offers a technical description of these techniques, and is required only when the goal is to imitate the results provided by this research.

3.1 Clustering algorithms

The first set of techniques aims to group the different carepaths into clusters of similar paths. The similarity of carepaths can be based on their activity frequencies as described by Representation 2, or also take into account the ordering of activities as described by Representation 1 (see Section 2.1.1.3 for more detail). In the previous chapter we stated that we have two main types of clustering: hierarchical and partitional. The next section gives a detailed description of the implementation of these techniques.

Hierarchical clustering This technique comes in two variants: agglomerative (bottom-up) and divisive (top-down) clustering, where in each step clusters are merged or split based on the value of a certain objective function.

In this study we use the `hclust` algorithm: an agglomerative variant of hierarchical clustering implemented in R. This algorithm starts at the bottom where each individual carepath represents a cluster. The algorithm takes a dissimilarity matrix as input; this matrix describes the distances between all pairs of carepaths, based on a certain distance function between two paths, as described in detail in Sections 3.1.1 and 3.1.2. Various methods can be applied to evaluate which two clusters are closest together. We use *Ward's minimum variance* method, which aims at finding compact, spherical clusters by taking the minimum value for the sum of squares objective function:

$$d_{ij} = d(X_i, X_j) = \|X_i - X_j\|^2 \quad (3.1)$$

Other methods are not covered in this study, as a short analysis showed Ward's method offers the best performance. Pseudocode for the algorithm is given in Algorithm 1.

An advantage of this technique is that the user can select the number of clusters *after* the algorithm is executed, but it is also computationally heavy. Clusters are never split after they are merged, which can lead to a suboptimal result for specific clusterings.

PAM This is an example of a K-medoids algorithm: a clustering algorithm related to the *K-means* and *medoidshift* algorithms. Both the K-means and K-medoids algorithms are partitional (breaking the dataset up into groups) and attempt to minimize the distance between points assigned to a cluster and a point designated as the center of that cluster. K-medoids chooses observed data points as centers (medoids) as opposed to the calculated average value of data points used in K-means [67]. Although neither algorithms can guarantee a global optimum, K-medoids is more robust to noise and outliers as compared to K-means, because a medoid is less influenced by outliers or other extreme values than a mean [67]. An example of this difference is given in Figure 3.2 [50]: a typical convergence to a local minimum for K-means is shown in (1a-1f), whereas (2a-2h) represent the obvious clustering by applying K-medoids to the same initial medoids.

K-means Taking user-input K , it tries to break up the dataset into K groups, attempting to minimize the distance between points in a single group and the center of that group.

K-medoids Similar to K-means, but using a single data point in a group for center (medoid).

Pseudocode is given in Algorithm 2. A disadvantage for these algorithms is that we need to specify the number of clusters K a-priori, which implies that the algorithm has to be executed for each value K . Numerous executions may be required to get satisfactory results, since there is no universal rule to determine the preferred number of clusters for any clustering technique [30, 32], and these algorithms cannot guarantee a global optimum.

An advantage is that the clustering for K clusters does not depend on the clustering of $K - 1$, and it may lead to better results (as shown in case study 3 – Section 5.3). Since the `pam` algorithm is computationally fast, this often outweighs the disadvantage of selecting K beforehand.

<hr/> <p>Algorithm 1: <code>hclust</code></p> <hr/> <p>Data: a distance matrix d representing the distances between all pairs of carepaths</p> <p>Result: clustering dendrogram</p> <ol style="list-style-type: none"> 1 repeat 2 Merge the closest two clusters; 3 Update distance matrix with the new clusters, to reflect the distance between the new cluster and the original clusters; 4 until <i>only one cluster remains</i>; <hr/>	<hr/> <p>Algorithm 2: <code>pam</code></p> <hr/> <p>Data: a set of data points $S = \{x_1, \dots, x_n\}$, a distance matrix d, the number of clusters K</p> <p>Result: a set $M = \{m_1, \dots, m_K\}$ of medoids, a vector c for cluster membership</p> <ol style="list-style-type: none"> 1 foreach $m_i \in M$ do 2 Select K points from S as initial medoids; 3 end 4 repeat 5 Form K clusters: update c by assigning each point to its closest medoid from M; 6 Update the medoid of each cluster; 7 until c has not changed; <hr/>
---	---

Figure 3.1: Pseudocode for the hierarchical clustering and partitioning around medoids algorithms [66].

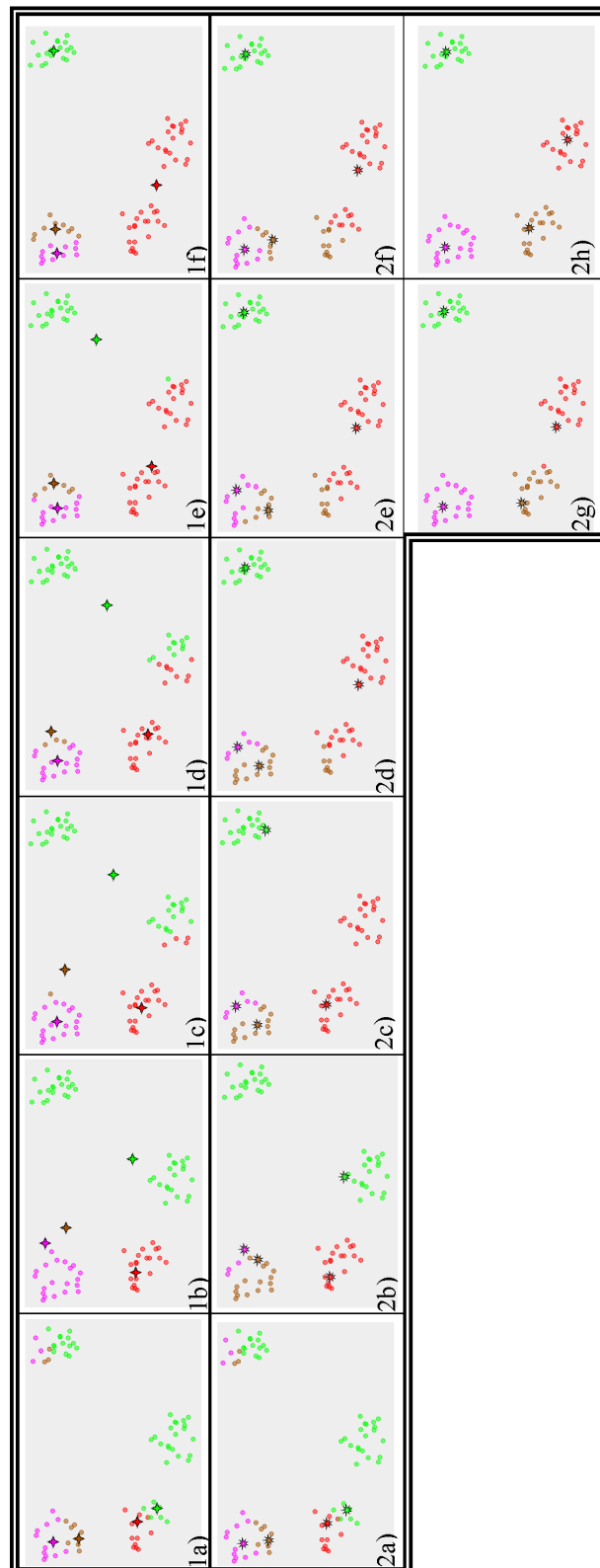


Figure 3.2: K-means versus K-medoids – the small circles are data points, the four-ray stars are *means* and the nine-ray stars are *medoids*. (1a-1f) present a typical example of K-means convergence to a local minimum. (2a) starts with the same initial position of medoids, but the final result (2h) represents the obvious clustering as opposed to the local minimum in (1f).

3.1.1 Clustering using Vector Distances

The clustering algorithms described in the previous section are based on the distances between objects and their specific properties and attributes. In our specific case, we have a collection carepaths: the set of activities patients had to undergo in order to be treated for a specific diagnosis. By creating a *vector of activity frequencies* (Representation 2), we are able to calculate the distance between two carepaths using different distance functions. The advantage of this representation is that it is a fast and clear way to represent carepaths, the downside is that we lose the ordering of the activities performed.

Distance functions describe the difference or dissimilarity between two data objects. These functions or measures have well-known properties for $d(x, y)$. A measure that satisfies all three of the following properties is called a metric [66]:

1. $d(x, x) \geq 0$ for all x and y and $d(x, x) = 0$ only if $x = y$ (Positivity)
2. $d(x, y) = d(y, x)$ for all x and y (Symmetry)
3. $d(x, z) \leq d(x, y) + d(y, z)$ for all x, y and z (Triangle Inequality)

In this study, we explore three well-known distance functions: the Euclidean distance, the Cosine distance and the Tanimoto distance. The latter two functions are derived from their similarity counterparts.

Euclidean distance One of the most commonly used functions is the Euclidean distance d between two points x and y in n -dimensional space, as described by the following formula:

$$d(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2} \tag{3.2}$$

The downside of this function is that the distance between two points can be small, even though they lack any common denominator. Take points $A = \{1, 0\}$, $B = \{3, 0\}$ and $C = \{0, 1\}$ for example, where each vector describes the frequency of two activities. According to the Euclidean distance, points A and C are closer together ($d = 1.4$) than points A and B ($d = 2.0$), even though they do not have a single common activity.

The Euclidean distance can be generalized by the *Minkowski distance* for $r = 2$, as described by the formula:

$$d(x, y) = \left(\sum_{k=1}^n |x_k - y_k|^r \right)^{1/r} \tag{3.3}$$

where n is the number of dimensions, x_k and y_k the k^{th} component of x and y and a parameter $r \geq 1$. The number of dimensions n should not be confused with r , which simply specifies different ways of combining differences in each dimension into an overall distance.

Tanimoto distance A function derived from the Extended Jaccard(EJ) similarity measure. This distance function compares both the similarity and diversity of two vectors, and is defined by the following equation:

$$EJ(x, y) = \frac{x \cdot y}{||x||^2 + ||y||^2 - x \cdot y} \tag{3.4}$$



The Tanimoto distance is a proper metric for binary values (which is actually similar to the Jaccard coefficient), but also proven to be a proper distance metric for positive valued vectors [42].

$$d(x, y) = 1 - EJ(x, y) \quad (3.5)$$

We expect this distance function to offer good performance with regards to carepath clustering, as matching zero-frequencies do not contribute to the similarity (unlike with the Euclidean distance), and the magnitude of frequencies is taken into account.

Cosine distance A function derived from the Cosine similarity: a similarity measure commonly used for text mining. It takes two frequency vectors as input and measures the cosine of the angle between them, using the following formula:

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \|y\|} \quad (3.6)$$

where $x \cdot y = \sum_{k=1}^n x_k y_k$ (dot product) and $\|x\| = \sqrt{\sum_{k=1}^n x_k^2} = \sqrt{x \cdot x}$ (length of vector x).

The distance function is described by the following formula and is a proper metric [66]:

$$d(x, y) = \arccos(\cos(x, y)) \quad (3.7)$$

Similar to Tanimoto, matching zero-frequencies are not considered to contribute to the similarity, as required for mining text or transaction data (the input vectors are often sparse). However, the magnitude of frequencies is not considered, therefore this measure is expected to cluster carepaths with similar ratio's between activity frequencies. Cases where every activity is performed ten times the regular frequency will not be labeled as exceptional, but based on their similar activity ratios they will be clustered together with the shorter paths.

3.1.2 Compression Clustering

In the next section, we describe a number of techniques that take into account the ordering of carepath activities, by using *string representations of activity sequences* (Representation 1). The Kolmogorov complexity gives us a good measure to compare two strings x and y [13]: it describes the length of the shortest computer program that produces x as output (in bits). Compression algorithms offer a suitable heuristic for the calculation of the Kolmogorov complexity of input x (e.g. a string or an image), as their goal is to describe x in the smallest number of bits. Compression offers a useful tool for many data mining tasks, as it can be applied in a generic way and are often parameter-free. The according distance function is described by the following formula [14, 22]:

$$d_c(x, y) = \frac{C_{xy} - \min(C_x, C_y)}{\max(C_x, C_y)} \quad (3.8)$$

where C_x is the size of the compressed string in bits, and xy is the concatenation of x and y .

We expect compression-based clustering to offer added value to the activity frequency clustering techniques described in the previous section, as different compression techniques perform well on recurring subsequences and therefore on similar carepaths. A downside of this technique is that small variations in the ordering of activities may greatly decrease the performance of the compression techniques, and activity frequencies may not have enough influence on the final clustering. In the next section, we explain the technical properties and implementation of our compression-based clustering algorithms.

3.1.2.1 Compression algorithms

In this section we describe two compression algorithms (GZIP, Bzip2) readily available in R. For both of these algorithms, we give a short summary of the idea behind the algorithm and how we expect it to perform on our carepaths. A third algorithm (xv) is also available, but due to the extremely poor performance in compression and computing time, we choose not to cover this technique in more detail.

GZIP Using the DEFLATE algorithm, which is a combination of LZ77 and Huffman coding, compression is achieved in two separate steps: [60]

1. *Matching and replacement of duplicate strings with pointers.* The LZ77 part of the algorithm identifies duplicate substrings and replaces the duplicate by a back-reference to the first occurrence (length, distance). An example is given in Figure 3.3.



Figure 3.3: Example LZ77 compression of *Length* = 4 and *Distance* = 5.

2. *Replacing symbols with new, weighted symbols based on frequency.* The Huffman method creates a tree based on symbol frequency. The resulting bit representation (Dictionary) uses the shortest bit-sequences for the most frequent substrings or symbols.

Compressing two similar or even identical strings gives a lot of possible back-referencing and allows for good compression. In Section 4.4.1.2 we use this feature for clustering the different carepaths.

Bzip2 An open source implementation of the Burrows-Wheeler (BWT) algorithm [63]. It achieves a higher compression rate than GZIP, but also increases computation time and resource usage.

It uses BWT which takes all possible rotations of the input string, sorts the strings and takes the last column. If the original string contains several duplicate substrings, the resulting string will have sorted similar characters together [12]. An example is given in Table 3.1.

Table 3.1: Example of Burrows-Wheeler Transform on “banana-”.

Input	Rotations	Sorted rotations	Output
banana-	banana-	anana-b	bnn-aaa
	-banana	ana-ban	
	a-banan	a-banan	
	na-bana	banana-	
	ana-ban	nana-ba	
	nana-ba	na-bana	
	anana-b	-banana	

The next steps exist of amongst others the *Move To Front* transform and *Huffman coding*. The different transformation steps in Bzip2 allow for recurring substrings to be combined together, which increases compression performance [12]. Similar to GZIP, we hope to exploit this feature in Section 4.4.1.2.



3.1.2.2 Clustering compressed activity strings

In Section 4.4.1.2 we represent individual carepaths using strings. As described in Section 2.1.1.1 there are less than 24 different unique activity classes, therefore a simple alphabet suffices to convert these carepaths into strings using a trivial lookup table like ZPK1 = 'A', ZPK2 = 'B', ... In Section 4.2 we will see that a small subset of activities suffices to represent the general patient careflow. Using the compression algorithms described above, the difference (or *distance*) between two carepaths strings is calculated using the Kolmogorov complexity function described by Equation 3.8. Pseudocode for the algorithm is given in Algorithm 3.

Algorithm 3: calculate Kolmogorov complexity

Data: a set of strings S
Result: a distance matrix describing which strings are most similar

```

1 foreach combination of strings  $x, y \in S$  do
    | // We only store the minimal compressed size of the concatenations  $xy$  and  $yx$ ,
    | // this way we maximize the compression rate.
2 | Calculate size of compressed concatenated strings ;           //  $C_{xy} = \min(C_{xy}, C_{yx})$ 
3 end
4 foreach combination of strings  $x, y \in S$  do
5 | Calculate the distance between  $x$  and  $y$  ;           //  $d = C_{xy} - \min(C_x, C_y) / \max(C_x, C_y)$ 
6 | Store  $d$  in distance matrix ;
7 end
8 return Distance matrix

```

Figure 3.4: Pseudocode for the calculating the compression-based distance matrix.

3.2 Classification

Classification mining techniques, such as decision tree classifiers, describe a systematic approach to building classification models from an input data set. These techniques identify a model that best fits the relationship between the class label and attributes of input data by employing a learning algorithm. Besides fitting the original input data, the goal of the resulting model is to correctly predict class labels for new records. For the purpose of this study, we use a *decision tree* to identify the main characteristics of individual clusters: for each child, the tree splits the value of *one* attribute into two child nodes (a decision) until there are no decision that improve the number of correctly classified nodes. An example decision tree is shown in Figure 3.6.

At the basis of many decision tree induction algorithms is Hunt's algorithm, a recursive fashion for growing a decision tree by partitioning the training records successively per subsets [66]. We use one of these decision tree algorithms, CART, as implemented in R (library `rpart`). The algorithm basically consists of two (recursive) steps, described in Algorithm 4.

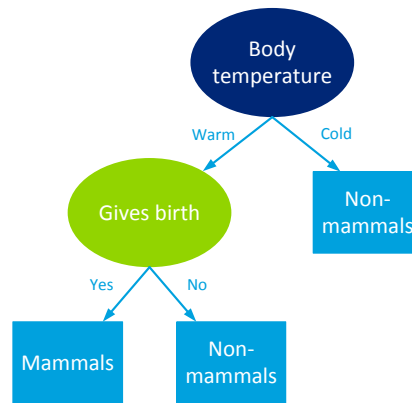
As a result, the algorithm returns a table with five variables, where each record depicts the results for a single iteration (i.e. a specific number of nodes). One of the most interesting variables is the `xerror`,

Root (dark blue) a node that has no incoming edges and zero or more outgoing edges

Internal (green) each of which has exactly one incoming edge and two outgoing edges

Leaf (light blue) each of which has exactly one incoming edge and no outgoing edges. Each leaf is assigned a class label

(a) Three types of nodes



(a) Example tree

Figure 3.6: Overview of a decision tree result. Animals are classified whether they are mammals or not, using only two characteristics. The values for these characteristics are described at the lines between nodes.

Algorithm 4: Hunt’s algorithm

Data: a set of training records D_t associated with node t , a set of class labels $y = \{y_1, y_2, \dots, y_c\}$

Result: a decision tree

- 1 **if** the records in D_t belong to the same class y_t **then**
 - 2 t is a leaf node labeled y_t
 - 3 **end**
 - 4 . **if** the records in D_t belong to more than one class **then**
 - 5 // Partition the records into subsets based on an *attribute test condition*
 - 6 Create a child node for each outcome of the test condition;
 - 7 Distribute the records in D_t based on the outcome of the test;
 - 7 Apply Hunt’s algorithm to each of the child nodes;
 - 8 **else**
 - 9 **end**
-

Figure 3.7: Pseudocode for Hunt’s algorithm, which is the basis of many existing decision tree induction algorithms [66].

as this depicts an estimate of the cross-validated prediction error for different numbers of splits. The number of splits is shown in the `nsplit` column, which indicates the size of the tree. The third useful variable is the complexity parameter (CP): the lack of fit has to decrease by at least a factor of the CP for a split in the tree to be allowed. The main role of this parameter is to save computing time by pruning splits that are obviously not worthwhile: the decision tree algorithm splits each node containing records that belong to more than one class, the CP makes sure that these splits do not continue indefinitely (i.e. until all nodes contain records of just one class), but rather only allow splits that are worthwhile.



3.3 Process Mining

In order to assess the patient careflow described by the individual clusters from the previous sections, we apply two process mining techniques: the Heuristics Miner and Trace Alignment (with Guide Tree). The first is useful to draw a global workflow model for the main process, whilst the latter is able to identify the main process pattern within a cluster. In this section, we provide a general introduction to these techniques. For further specification we would like to refer to the literature on these techniques.

3.3.1 Heuristics Miner

The Heuristics Miner algorithm has proven to be able to deal with noise and low frequent behavior, resulting in workflow models that describe only the main behavior. These models offer a high-level overview visualization of the process [28, 72]. This technique lacks the ability to indicate the number of occurrences of certain paths and dependencies. Since the data we use is high-level and offers a limited number of distinct activities, we expect rather simple models as a result. A few examples are given in Figures 5.3, 5.11 and 5.18. Instead we look at the Trace Alignment for more detail on the process.

3.3.2 Trace Alignment

This technique offers more detail on both the most frequent carepath pattern and the individual deviations within a group of carepaths or *traces* [10]. In an early stage of the analysis, this technique can be used to explore the process, whilst specific questions can be answered at later stages.

A formal definition of Trace Alignment is proposed in [9, 10]. In short, this technique tries to align activities from one trace to the activities described by all other traces using gaps-symbols for activities which do not match. An example for different pair-wise alignments is given in Figure 3.8. The number of possible alignments is high (e.g. for two traces of length 100, the number of possible alignments is approximately 10^{77}), therefore it is infeasible to enumerate all possible alignments.

(a)	(b)	(c)
trace 1: a b c a c - -	trace 1: a b c a c -	trace 1: a b c a c - - - - -
trace 2: a - c a c a d	trace 2: a c a c a d	trace 2: - - - - - a c a c a d

Figure 3.8: An example of different pair-wise Trace Alignments. It is infeasible to enumerate all the possible alignments, this is just an impression of a few possibilities [10].

In order to select the best alignments, the *sum-of-pairs* method is adopted. A succession of pairwise alignments is iteratively constructed (i.e. progressive alignment), where the selection of aligned traces for each iteration is based on their similarity [10].

Note that the Trace Alignment is based on the *Guide Tree*, which performs hierarchical clustering of activity sequences. Since we perform extensive clustering analysis using more traditional clustering techniques, we “skip” this clustering step and instead select one cluster (i.e. the entire set of carepaths from our clustering algorithms described earlier) and align the entire cluster. A few examples are given in Figures D.4 and D.5.

3.4 Conclusion

In this section we provided more detail on a number of different techniques and metrics as described in Section 2.3. During the methodology design phase, we explore these techniques and select the best algorithms to perform clustering, classification and process mining.



Chapter 4

Methodology design

In previous chapters we gave a thorough understanding of the healthcare environment, patient careflow and mining techniques. In this chapter, we apply this information to the different phases of CRISP-DM.

4.1 Business Understanding

The first phase is to define and understand the *objectives* and *requirements* from a business perspective. We translate this knowledge into a data mining problem definition [75]. The main objective of this study as stated in Section 1.2 is:

To explore advanced Process- and Data mining techniques, and to define a methodology that provides insight into patient careflow for specific DBC's in a hospital environment.

In other words, we are looking for a methodology that supports us in gaining insight into patient careflow in order to be able to improve its performance, quality and standardization. Our search for the specific business objectives and requirements is guided by our research questions.

4.1.1 Which insights do we require to assess patient careflow?

Question 2.1: Which criteria (logistic/medical/cost) are used for the assessment? Due to the highly dynamic and flexible nature of the healthcare domain in combination with limited medical expertise involved in this project, we limit our scope to the *organizational* or *logistic* process. In order to understand what we mean by patient careflow we refer to Section 2.1.1.3, which describes how the Dutch reimbursement systems work. This section leads us to the five representations of patient careflow as shown in Table 4.1. Question 2.2: Which elements of a specific patient careflow can we use? For the purpose of this project, the analyses are based on patients within individual DBC-codes from the DBC system. Data on this level is readily available, and patient careflow described by the DBC system offers enough variety whilst being limited to one specialism.

One thing we have learned from previous data mining projects in both healthcare and other domains is that data mining is an interactive and iterative process. Question 2.3: Which parameters/techniques do we need to calculate quality for a specific carepath? The selection of the number of clusters is a compromise between quality/significance and usability. In other words: a high number of clusters

Table 4.1: List of the three different representations of patient careflow.

#	Description
1.	A string representation of the ordered sequences of activities on a ZPK-level.
2.	Counts of the performed activities on a ZPK-level.
3.	Counts of all performed activities on a CTG-level.

can have the best numbers regarding quality, but is often less insightful or even unusable than a lower number of clusters. For this reason, we focus on a variety of *visualization techniques*. Visualizations provide a useful tool to gain insights into results, and are easy to communicate results to medical specialists, nurses and other medical staff. With the right type of visualization, a clear assessment of quality and usability is easily performed. When we look at quality of patient careflow, we look at the level of homogeneity as defined by:

1. The (sub)sequence of events.
2. The number of events (both in total and individually).
3. The total cost of individual carepaths.

Question 2.4: Which elements of a carepath do we have available as input? Attributes like *activity sequences*, *activity counts* and (*estimated*) *cost functions* can easily be derived from DIS and other event logs. The details of the available data are explored in the next phase of CRISP-DM in Section 4.2.

4.1.2 How can we compare, evaluate and advise different carepaths?

Table 4.2: Examples of views 1 and 2 of patient careflow, based on a small eventlog. The ZPK 5 events are both on the same day, and represented by a single activity.

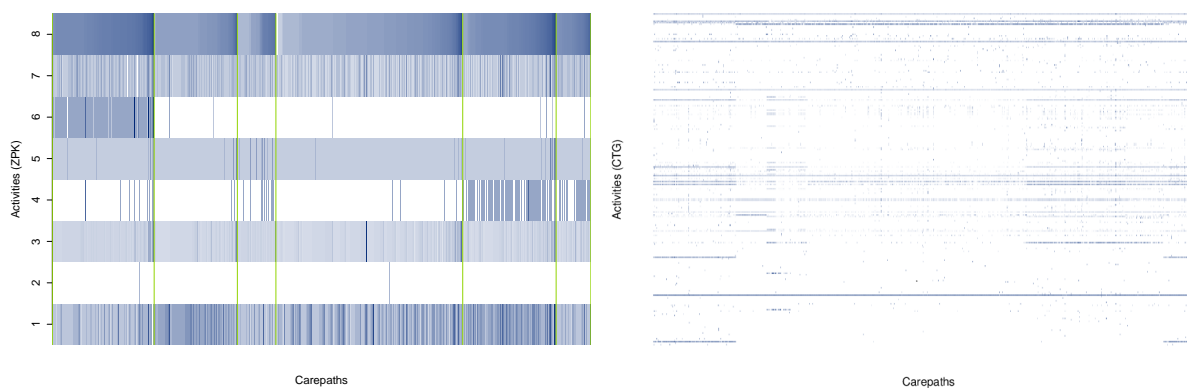
(a) Example event log			(b) Representation 1 and 2								
CTG	ZPK	Date	ZPK	1	2	3	4	5	6	7	8
100	1	01-01-2010	Vector	1	0	0	1	1	0	0	0
400	4	01-01-2010									
500	5	02-01-2010	String	a	d	e	a				
501	5	02-01-2010									
100	1	03-01-2010									

Examples of the representations described in Table 4.1 are found in Table 4.2, based on the event log in Figure 4.2a. Figure 4.2b shows representation 2: a fixed-length vector, where the length is the total number of distinct activities over all carepaths, and representation 1: the sequence of activities per individual carepath are translated into strings, e.g. ZPK 1 = 'a', ZPK 2 = 'b', ..., ZPK 8 = 'h'. For both representations, we see that both ZPK 5 activities (CTG 500 and 501) are represented by a single activity, as these both occurred on the same day. Representation 3 is similar to representation 2, but then counts each individual activity. This produces a longer and sparser vector.



Question 3.1: How do we visualize patient careflow? By scaling the total counts per activity on a level between 0 and 1, we are able to visualize the many different carepaths where the gray-scale represents the relative quantity [74]. This is possible on a ZPK-level, as shown in Figure 4.1a, and similarly on a CTG-level as shown in Figure 4.1b.

Question 3.2: How do we define patient careflow quality? The visualized part of patient careflow *quality* is simple with the use of the previously described visualization techniques. When a group of carepaths shows a high level of logistics homogeneity, its visualization shows little color variation for each row in a column. Vice versa, a high level of color variation implies inhomogeneous patient careflow. Theoretically, the number of activities and total cost should be similar for individual carepaths. An example visualization of one patient careflow on both a ZPK and CTG-level is shown in Figure 4.1. Although some patterns are visible in the CTG overview, the high spread of activities support the generalization to a ZPK-level in order to improve readability.



(a) The variety of performed activities on a ZPK-level (b) The variety of performed activities on a CTG-level

Figure 4.1: Count overviews of patient carepaths. (4.1a) shows Representation 2 for ZPK 1 through 8, where the green lines separate the different hospitals. (4.1b) shows Representation 3 of all individual activities. The high spread of CTG-activities supports the generalization to the ZPK-level.

For a good clustering of patient careflow, we expect to see relatively little color variance within each cluster column (separated by the green lines in Figure 4.1a), and relatively high color variance between clusters. As we have stated before we choose not to specify numerical measures, because for the business application of this methodology a limited number of clusters is preferable. Although it is difficult to create a valid model on too few ZPK-codes, too many patient types become unmanageable from both a modeling and business point of view [32]. Question 3.3: How do we compare different cluster outcomes? We aim to group patients in a small number of clusters, but this is often not statistically optimal, as the variance within the resulting clusters will be relatively high. Therefore we choose to use a variety of visualization techniques for the assessment of clusters, instead of a numerical measure. The list below describes the five different visualization techniques that provide insight on both the cluster quality and the patient careflow described by such a cluster:

Overview plot as shown in Figure 4.1. This overview shows the number of activities (y -axis) of the individual carepaths (x -axis), using color to indicate the frequency – the darker the color, the higher the frequency for this activity. The high spread of CTG-activities supports our decision to

generalize activities to the ZPK-level. Green lines are used to indicate different clusters.

Parallel coordinates for ZPK frequency This is a simple technique to visualize the different activity frequencies per carepath. A complete overview of all paths helps identifying the important factors in clustering, where individual plots per cluster give clear insight in the frequency patterns within a cluster. The latter is useful to compare a group to its representative carepath as defined in Section 4.4.3. An example is given in Figure 4.3.

(Stacked) Barchart for ZPK frequency The overview and parallel coordinates plots offer nice frequency indications per ZPK per carepath. However, neither provide a clear overview of the total number of activities. For this purpose, we use a stacked barchart where each ZPK is labeled with a unique color. The total height of the bar is the total number of performed activities. We also use the green bars to indicate the different clusters similar to the overview plot. An example is given in Figure 4.2.

Histogram for activity frequency A frequency histogram for each activity offers a clear visualization of the distribution over a group or cluster. It is useful to evaluate both data quality and cluster quality, as it allows us to check the activity frequencies overall and the difference of frequencies between clusters. An example is given in Figure 4.4.

Boxplot for costs and frequency A simple though effective *numerical* way to show patient care-flow quality is to calculate the spread in total costs and number of total activities of the individual carepaths in a cluster. A small spread represents high homogeneity within a cluster, and different value-ranges between clusters represents a good clustering. Boxplots regarding total costs and number of activities give an extra level of detail with regards to the difference in cluster quality. An example is given in Figure 5.6.

Once we have identified our preferred clustering, we need to identify the main carepath that represents each cluster. For this purpose, we use the Trace Alignment as described in Section 3.3.2. This tool enables us to preserve ordering, visualize patterns and identify deviations.

4.2 Data Understanding

The second phase is all about data: the selection and collection of available data, discovering first insights into the data, understanding the data and identification of data quality problems. This phase has a close link to the Business Understanding phase, as the identification of business objectives is impossible without at least some understanding of the available data [75].

For the remaining phases in CRISP-DM describing the methodology design, we take a subset of the available data for DBC 305..1701.223 (arthrosis: pelvic/hip/thigh – does not describe a *care type*). This is a relatively routine treatment and allows us to assess the different available techniques.

4.2.1 Collection

In Section 2.1.2 we described the DIS, a national system that is responsible for gathering all DBC data. Every hospital in the Netherlands is legally obligated to provide this system with their DBC data,



describing both the provided and billed care. Since this system offers one single dataset with all DBC's of all Dutch hospitals, this is a suitable source: it supports a standardized analysis methodology and allows benchmarking between hospitals and patient careflow.

Unfortunately, data from the official DIS was not publicly available at the collection phase of this project. Instead, we use similar datasets that are at our disposal that describe DBC data in a similar fashion. These sets were originally used for the analysis of the effect of the implementation of DOT compared to the current DBC system. The quality requirements of these sets were lower than that of DIS, which required a more thorough data audit.

Neither the original DIS, nor the sample datasets used in this study contain specifics about the registered activities and treatments. For these details, we turn to DBC Onderhoud, which provides us with more details on DBC and CTG codes.¹ Although activity prices are not publicly available for a large part of CTG-codes, Deloitte has calculated average prices for over 900 CTG codes in previous projects [74]. Experts also gave an indication of prices on the rough average cost of a ZPK-level activity. An overview of the resulting external datasets is shown in Table 4.3, each of which is linked to our main dataset.

Table 4.3: External data files.

Name	Description	Source
DBC Typification list	A set of DBC codes and their description	DBC Onderhoud
Activities description list	A set of CTG-codes, with their description and the ZPK it is classified to	DBC Onderhoud
CTG Cost table	A set of averaged costs for ± 900 CTG's	Deloitte
ZPK Cost table	A set of roughly estimated prices for a number of ZPK's	Deloitte

4.2.2 Understanding

In order to be able to format and assess the acquired data, it is important to have some understanding of the data: which columns do we have, what do they mean and what values can they acquire? We start with a more thorough overview of the available tables as described in Section 2.1.2 and Table 4.3.

Patients This table contains patient data: a unique identifier (unique on a hospital level), birthdate (DD-MM-YYYY) and sex (1 = male, 2 = female, 0 = missing).

Note that in the DIS system, PatientID's are based on the *Social Security Number*. This means patients can be tracked back to every hospital they visited, which allows for more extensive analysis. In our sample sets, PatientID's are entirely anonymous.

Carepath Each patient is linked to one or more carepaths, and each carepath is directly linked to a specific specialism. The creation of two or more simultaneous carepaths is restricted and allowed only if it satisfies specific rules.

Subcarepath In order to make distinction in diagnosis and treatment, each carepath is subdivided into *subcarepaths*. A side-effect is that these subcarepaths are an administrative feature that

¹http://www.dbconderhoud.nl/index.php?option=com_docman&task=doc_download&gid=2323&Itemid=593

help hospitals declare provided care before the completion of an entire carepath. This table also contains the AGB-code, which represents the hospital the patient was admitted to for that sub-carepath.

We look up the codes describing the diagnosis, treatment and specialism from this table in the “DBC Typification list”, which provides us with a textual description of each DBC.

Activities table Each subcarepath describes a series of activities, and each activity is represented by a CTG-code. As stated before, the only recorded time is the date of the execution of the activity. No specific time or duration is available. This table also contains an *amount*-field, which describes a variety of values: e.g. the number of performed activities, or the amount of milliliters of medicine. It can also contain a 0 for activities that have not been performed even though they are part of the predefined carepath, or when the activity is already booked on a different DBC. In the Data Preparation phase, we describe how we cope with these values.

Based on the CTG-code, we find the description and ZPK-code in the “activities description” list. The ZPK-code can be seen as a higher-level activity description and is useful for generalization of medical activities, especially for the logistics process.

DBC Typification list This list contains descriptions for the DBC components. The descriptions depend on the specialism, and an *As code* describes the component (1 = type, 3 = diagnosis, 4 = treatment).

Activities description list For each CTG-code, this table provides an activity description and ZPK-code.

CTG Cost overview A list of averaged prices in Euro per CTG-code. Note that this list contains less than the top-20% (± 900 out of ± 5100) most occurring activities of all possible CTG-codes, but covers over 97% (± 80650 out of ± 83000) of all the activities described in the event log. The missing values mainly describe anesthetics (ZPK 6) and labtests (ZPK 8), which would have little effect on the total costs of a carepath.

ZPK Cost overview A list of rough prices in Euro per ZPK-code. Note that only for ZPK 1, 2, 3 and 5 pricing is included. For the purpose of this study, the other activities are assumed to be financially insignificant.

In our definition of patient careflow as proposed in Section 2.1.1.3, experts have pointed out that for most DBC's only ZPK levels 1 through 8 are relevant. This is a valuable insight, as this decreases the number of distinct possible activities and immediately decreases the amount of noise. Experts have also indicated that with the arrival of new interns, they see the number of ZPK 8 (labtest) activities increase.² No data is available regarding the involved staff, therefore it is impossible to determine which labtest was requested by whom. This would result in a bias towards the (lack of) experience of the performing staff member. An interesting question is whether statistics on the set of requested labtests can support interns in their decision making, as they have statistic support for which labtests they should request.

On a DBC level, a number of things have to be taken into account. It is possible that one patient is linked to more than one carepath, and each carepath can consist of more than one subcarepath.

²This is explained by their lack of experience. Interns often request more labtests “just to be sure” they include the right test.



In some cases, this implies that one patient has multiple diagnoses. Whilst we have little data on a patient's (medical) background, the number of diagnoses often offers a good indication of the complexity of a patient's required care. Due to the lack of medical background information, it is hard to identify *complication factors*. In combination with the fact that we focus on the *logistics process*, we are not going to try to classify patients in order to predict which type of patient careflow they require, and focus on the evaluation of the logistics properties of carepaths in general.

When it comes to individual activities, especially on a ZPK level, we need to be aware that the registered amount for an activity is not always a frequency. It can also be a measure (often in milliliters) or 0-value. Activities with an amount of 0 have to be ignored – they have not been performed and are there only for administrative purposes. As described for Representation 1 and 2, the logistics aspect of patient careflow allows us to summarize all ZPK-activities on one day in a single event. However, this is not the case for ZPK 3: some hospitals register the number of nursingdays per patient at the day of discharge, therefore we have to generate the activities accordingly.

The CTG Cost overview should offer a more accurate pricing than the ZPK Cost overview with regards to the total costs of individual carepaths. However, for this study this is not always desirable for two reasons. First of all, a more accurate pricing model might differentiate financially between two carepaths similar on both a logistics and a medical level. Since we are not trying to improve the reimbursement system, we do not want to differentiate solely for financial reasons. Secondly, with regards to the summarized activities as described above, multiple CTG-activities might be merged into one ZPK-code for a single day. In order to prevent extensive calculations and assumptions a generalized set of ZPK prices provides reasonable financial insight, even though they are infeasible in practice.

4.2.3 Audit

Quality assessment (or *data audit*) is also a big part of the data understanding phase. If the data is of bad quality or has a large number of missing values, the quality of the data-driven analysis will also be bad.

The origin of data should always be taken into account, as manual registrations have a high risk of typos and wrongful registrations. However, hospitals continuously invest money to prevent erroneous registrations by keeping their HIS and other systems up-to-date and according to protocol. Since the DIS system has set legal quality requirements, we assume that the registered data meets this quality standard.

At the Data Collection phase of this study, the DIS data was unavailable. Instead, we collected DBC data from individual hospitals. Each individual set required extensive modeling to make them conform the DIS structure, which was performed by colleagues during previous projects. Therefore, thorough evaluation of *data completeness* and *data correctness* was readily available.

In the modeling of our final data structure, a number of factors remain to be considered:

1. Some of the subcarepaths were linked to missing carepaths, these are disregarded.
2. Some of the patient data is missing. As stated before, we do not necessarily require these attributes, as a thorough predictive analysis is not part of our research objectives.
3. Some of the diagnosis values were missing. These records are disregarded, as we need the diagnosis to identify the DBC of a carepath.

4. The CTG Cost table describes prices for less than 20% of the total number of CTG-codes. It would be unrealistic to estimate specific activity values, and since we do not have the data to calculate average prices, we also include roughly estimated ZPK prices. These prices are solely used to indicate cost homogeneity of patient careflow, and do not represent a realistic carepath cost indication.
5. Not every CTG can be linked to a ZPK. This is because for some activities a different coding table (e.g. internal) is used. We remove these from the set and count the number of “extra” activities after clustering for validation (a large number of extra activities may indicate that a carepath is no longer representative, as too many activities are removed).
6. Some hospitals register activities in a single record, where the amount describes the number of performed activities. In order to preserve the ordering of activities, these activities need to be split up (this is important only for ZPK 3).

4.2.4 Visualization

In addition to Figure 4.1a, we give a first impression of the data by two different ways of indicating the frequency of ZPK activities: the (stacked) activity frequency barchart and parallel coordinates plot from Section 4.1.2. Figure 4.2 shows the total number of activities performed for each of the 2110 carepaths, sorted by hospital.

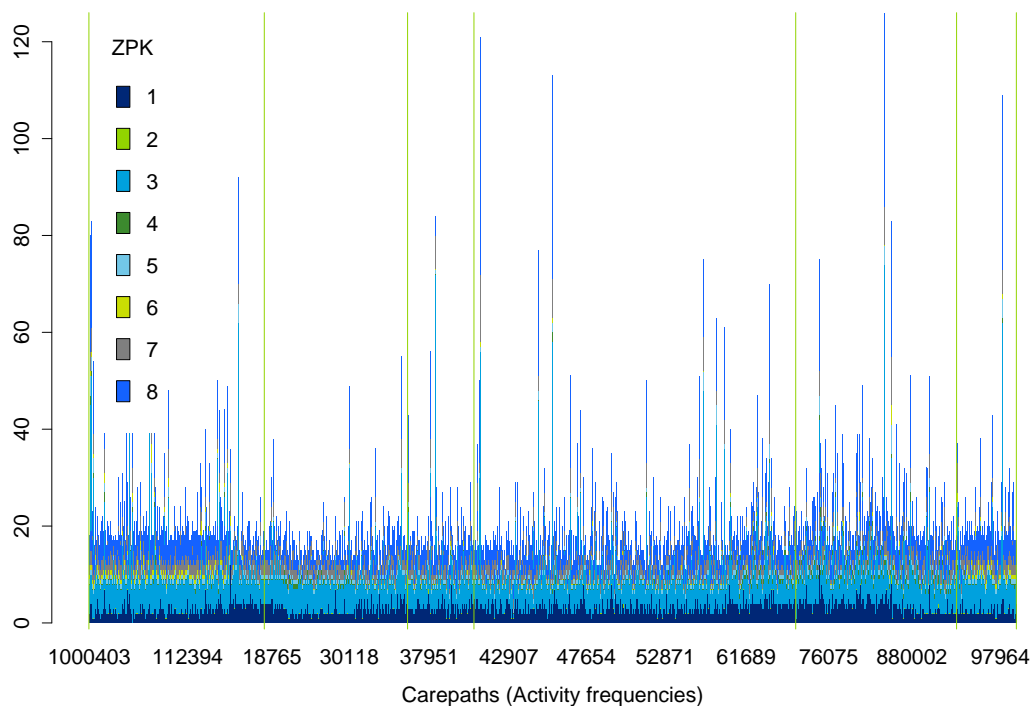


Figure 4.2: Activity frequency Barchart for all carepaths, per hospital (separated by green lines). Each bar represents a single path, each color one type of activity (ZPK).



In the parallel coordinates plot of Figure 4.3, each individual line represents one or more carepaths, and the height of the line represents the frequency for that activity. There is some indication of distinction between *regular* and *top clinical* hospitals (patients in the latter type generally require more surgery and clinic time). However, since the DBC system does not make this distinction with regards to the reimbursement for a single DBC, therefore we do not include the hospital type in our analysis. Instead we cluster all patients and try to differentiate between e.g. *routine* cases and *exceptional* cases. The resulting clusters will be checked to see whether they are dominated by a specific type of hospital.

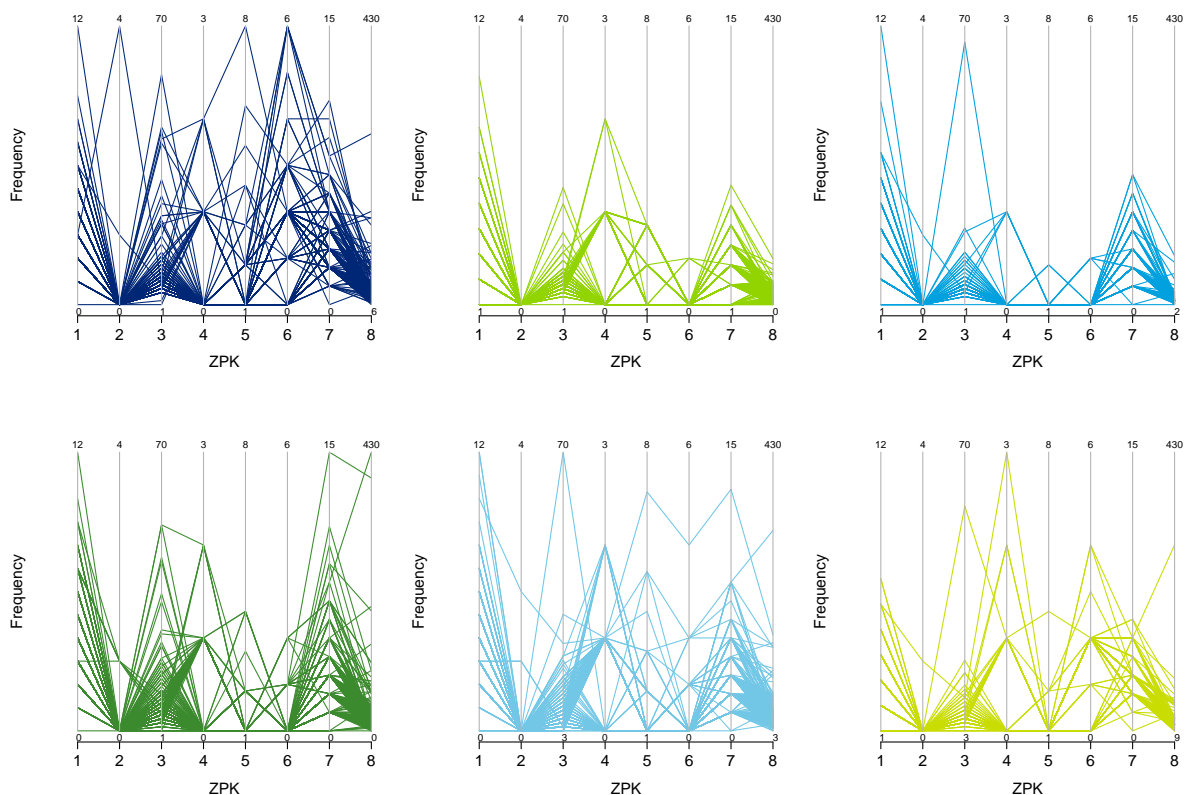


Figure 4.3: Parallel coordinates plot for all hospitals. A line represents one or more carepaths, the height the frequency of an activity.

From the figures above we read that for a small number of cases the total number of activities explodes: about 140 carepaths contained over 60 activities, 11 paths have more than 200 registered activities. ZPK 8 (labtest) is by far the most frequently performed activity, as visualized in more detail by Figures 4.4. This figure also shows a relatively large spread for ZPK 3 – some patients required up to 70 nursingdays. Experts indicated the difference in number of nursingdays is important and therefore interesting, as a larger number of nursing days is expensive and one of the reasons for increased waiting times. They also suggested ZPK 8 activities can be removed from the event log, since they are large in number, but of relatively low importance considering these are short and cheap activities. Besides, earlier we stated that interns apply more tests than experienced medical professionals, which is a known problem and defined as part of the learning process for interns.

Figure 4.4 suggests that ZPK 4 (diagnostic activity) and 6 (other therapeutic activity) are performed

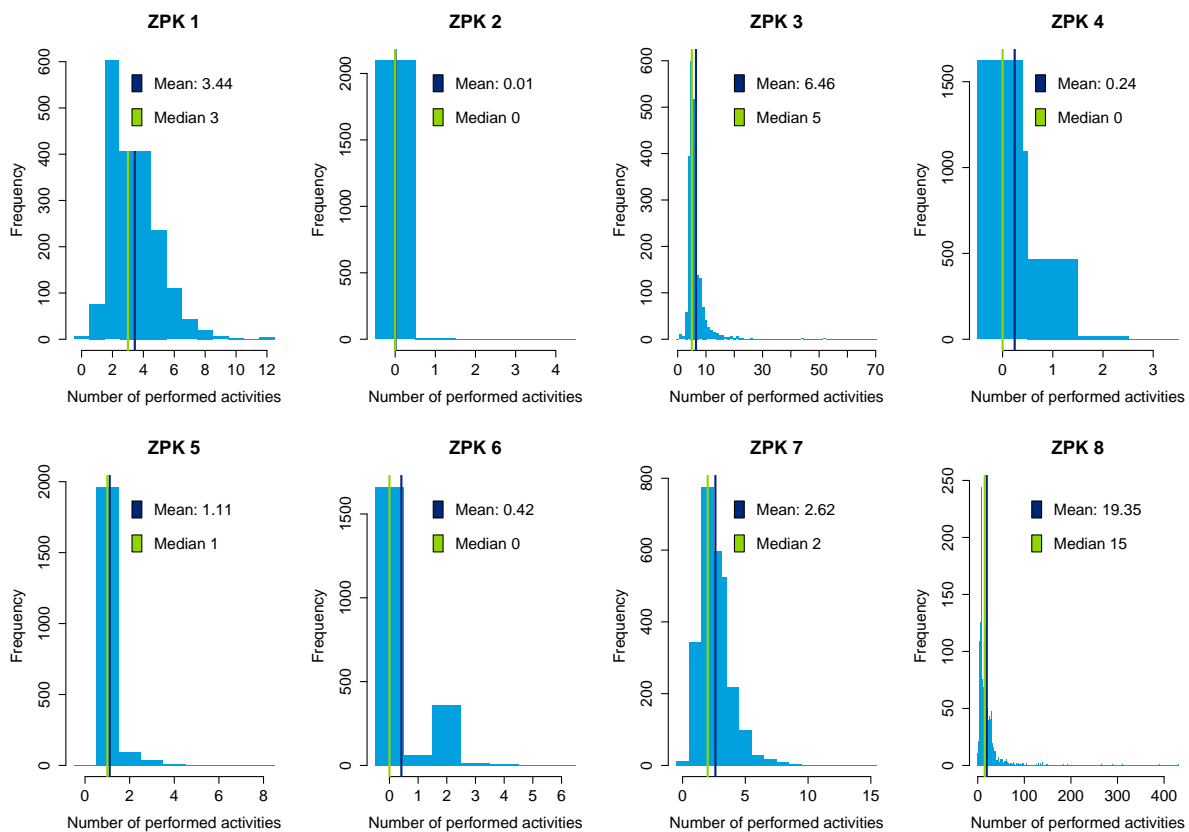


Figure 4.4: Histogram per ZPK for number of activities (over all hospitals).

only in about 21 to 24 percent of the total number of carepaths, as for most carepaths the activity frequency is 0. As especially diagnostic activities are required for most type of treatments, this seems infeasible. However, Figure 4.1a indicates that both these activities are concentrated in just one or two hospitals, which suggests this may be caused by the difference in administrative software and/or habits of hospitals. In future data collection steps extra care should be given to preventing these administrative differences, as especially diagnostic activities (ZPK 4) are an important part of patient careflow.

4.3 Data Preparation

The data preparation phase covers all steps required to transform the raw data into a format that will be fed into the modeling tool(s). Example steps are: selecting the right tables and records, merging multiple sets, cleaning missing data and transforming the overall structure [75]. Multiple different transformations are performed in order to comply with the required data structures for the different analyses.

In our case, data preparation includes a number of *preprocessing* steps: we “simplify” the log by removing the excess of low level activities (i.e. we focus on ZPK-level activities). By aggregating the data to one level, unnecessary details and other noise is removed which improves clarity of the gained insights. Appendix A gives a detailed description of the collected data structures and the different steps required in order to create the final sets. Although this is an extensive and time-consuming task, we summarize the performed steps in the following overview:



1. Import individual files into Microsoft SQL Server 2008, using SQL Server Integration Services. This step includes the formatting of individual records and the addition of (dummy) values.
2. Join individual hospital tables, and add details: DBC description, ZPK-codes and ZPK/CTG prices. This also includes the labeling of Process Instances: a unique ID for each carepath and one for each DBC (or subcarepath).
3. Select the activities for a single DBC, and calculate the (total) number of activities and carepath costs.
4. Export data for import in R, as described in Table 4.4. We start with the largest number of events feasible, where all activities for ZPK-codes 1 to 8 are included.

Table 4.4: Data model for input tables in R.

Carepaths		Patients		Activities	
PK	Process ID	PK/FK	Process ID	PK	Key
	CTG Count		DBC Count	FK	Process ID
	ZPK (aggregated) Count		Sex		Start date
	CTG Costs		Age		CTG-code
	ZPK Costs		Hospital		ZPK-code
	ZPK (aggregated) Costs		Path Start date		
	Other activities Count		Path End date		

5. Import and transform data in R. The tables Carepaths and Patients are imported as-is. The other objects are divided into separate tables for CTG-activities and ZPK-activities. This allows us to filter unwanted ZPK activities and clean duplicate ZPK-activities. The resulting table is used to generate the tables for Strings/Counts/Norm in R, which is more flexible and faster than MS SQL Server. The following list gives describes the resulting tables:

CTGs: a sorted list of PID's with ordered CTG activities.

ZPKs: a sorted list of PID's with ordered ZPK activities, aggregated to one activity per ZPK per day.

Carepaths / Patients: same as input table.

Strings: a list of PID's with String representations of ZPK activities.

Counts: matrix of activity counts (representing either CTG or ZPK activities).

Norm: matrix of activity counts scaled between 0 and 1 (this is used for pretty printing: the closer the value to 1 the darker the color representing a higher number of performed activities. 0-values have not been performed and colored white).

6. Import data in ProM, this includes: create output for ProM (one set per cluster), format data using MS Access macro and create MXML file using ProM Import.

4.4 Modeling

The fourth phase is concerned with the actual modeling of the data. In this phase, we apply various modeling techniques as described in Chapter 3. As some techniques require specific data formats, there is a close link between Data Preparation and Modeling. During the Modeling phase, it is common to identify data (format) problems and discover new ideas which require additional Data Preparation steps. The different model results are assessed, compared, and sometimes even combined [75].

Based on an event log for one unique DBC, our methodology describes three stages:

1. In the first stage, the individual carepaths are clustered based on (dis)similarity measures.

Different mining techniques are explored and evaluated for the clustering of different carepaths into groups that describe similar activity sequences. In a medical environment, individual cases can often be differentiated into categories or types of patients (e.g. *routine*, *re-admitted* and *complicated*). As we have seen in Figure 4.1, there is a lot of variation in carepaths. Our goal is to reduce this variation by dividing patients into similar clusters or categories.

2. The second stage applies classification to identify the main characteristics of each cluster.

3. The third stage identifies frequent recurring patterns and points out important deviations for each category.

This stage is based on the clustering result of the first stage: for each cluster we align activity sequences and identify the pattern as executed for the majority of cases. The visualization also points out deviations from this pattern. The global pattern of a cluster offers a good description for the patient careflow, and provides a more detailed description of the type of patients in each category.

4.4.1 Create Clusters

In Section 2.1 we proposed two different representations of carepaths: using *vectors* for the number of activities or *strings* for the sequence of activities. For either representation, different measures are available to define the (dis)similarity between two individual carepaths. Two main methods describe the available clustering techniques: *hierarchical* and *partitioning*. For hierarchical clustering, we use the standard `hclust` algorithm as implemented in R. For partitioning clustering, we mainly focus on the use of `pam` (partitioning around medoids) in R. In the sections below, we analyze the different possibilities as summarized in Table 4.5 and select the best available option.

4.4.1.1 Using Vector representations

We start modeling the dataset using vector representations for individual carepaths. In Section 4.2.4 we already introduced the issues with the exploding number of ZPK 8 activities. In order to demonstrate its disruptive power, we used the complete dataset for our initial clustering. The resulting clusters are visualized in Figure B.1. The overview plots show large variation within columns, where the activity barcharts only show a noticeable difference between clusters for ZPK 8. Taking the previously described expert's opinion on ZPK 8 into consideration, we choose to completely ignore ZPK 8 for further clustering.



Table 4.5: Different clustering techniques for both *Vector* and *String* representations.

(a) Vector		(b) String	
Clustering	Distance metric	Clustering	Compression
Hierarchical	Euclidean	Hierarchical	gzip
	Cosine		bzip2
	Jaccard	Partitioning	gzip
Partitioning	Euclidean		bzip2
	Cosine		
	Jaccard		

Similarly, none of the different clustering techniques and distance measures are able to cope with the administrative differences between hospitals for ZPK 4 and 6 as described earlier (see Figure 4.1 and Section 4.2.4). With the support of experts, we decided to also remove these activities from our dataset, as these will bias clustering towards differentiating between hospitals. This is unfortunate, because ZPK 4 is an interesting activity to use for analysis.

In order to remove further noise from the dataset, we filter each carepath to one activity per ZPK per day: ZPK 1, 2 and 3 can only occur once a day anyway³, and although multiple surgeries (ZPK 5 activities) may be registered on a single day, in a logistics sense this is similar to one activity.⁴ Similarly, ZPK 7 can be regarded as a single activity. Note that these steps are actually part of the Data Preparation phase, and show a perfect example of the iterative nature of CRISP-DM.

(a) Overview plot (sorted per Hospital – green line)

(b) Activities barchart (sorted on total count)

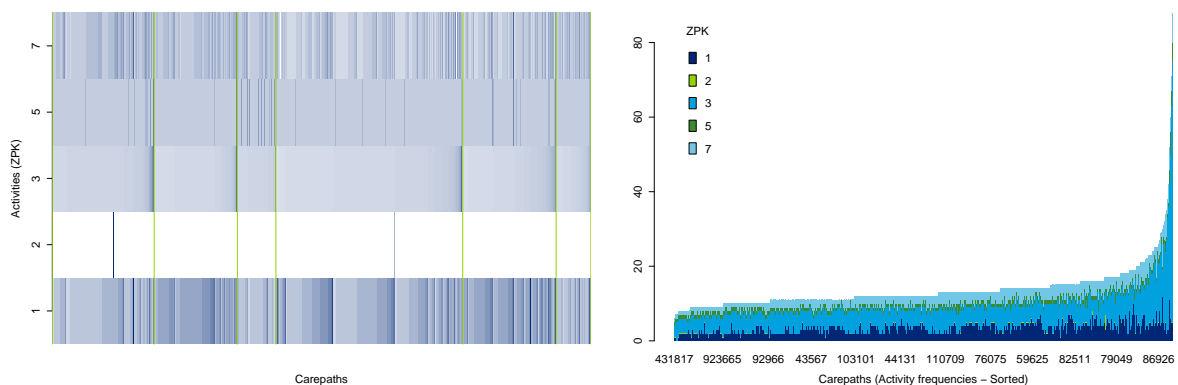


Figure 4.5: Initial overviews of DBC 305..1701.223 for ZPK 1, 2, 3, 5 and 7.

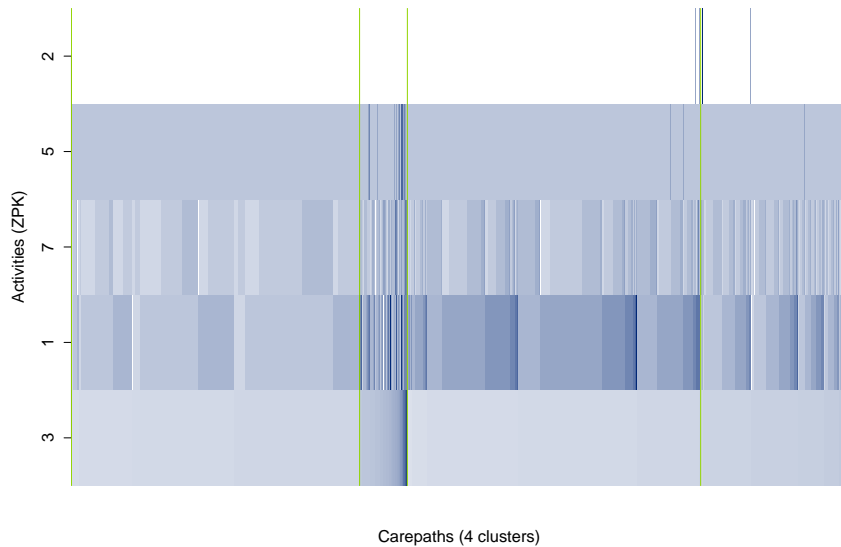
The remaining ZPK-activities are sorted per hospital and shown in Figure 4.5a. Figure 4.5b shows a stacked barchart for the same activities, sorted on the total number of activities.

In Section 4.1.2, we stated that a limited number of clusters is required to provide a workable so-

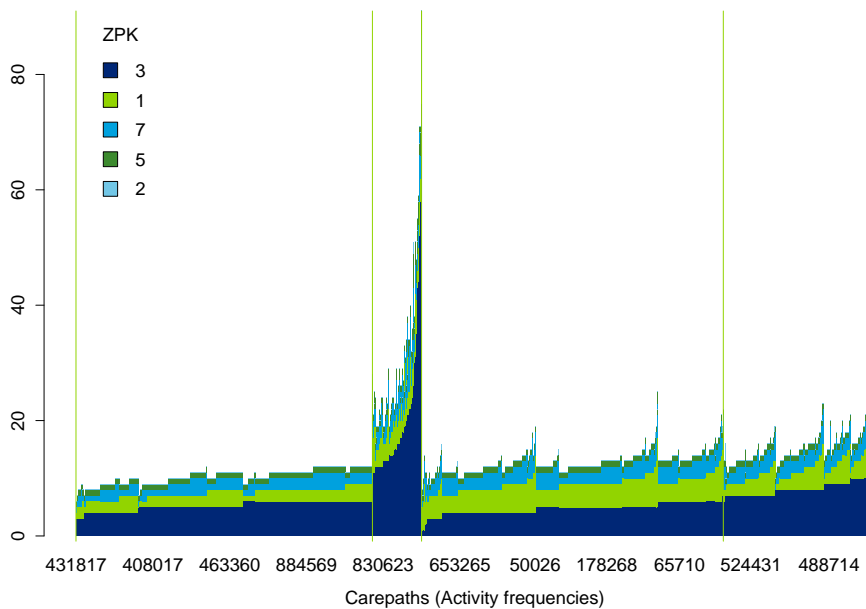
³ZPK 2 and 3 describe a day at the hospital.

⁴A patient can only undergo surgery once a day (which may include multiple activities), unless it is an emergency. The latter would imply a different DBC which would be filtered from our dataset.

lution. Based on expert input, we set the amount of preferable clusters to a number between 3 and 6. In Appendix B, we provide a complete overview of the results for all combinations of distance functions and clustering algorithms proposed in Table 4.5. However, we only show the results for four and five clusters. Using less clusters lacks sufficient distinction between different groups of carepaths, whilst the gain in cluster quality of more clusters does not outweigh the loss of its practical applicability.



(a) Clustering overview plot, showing different color variations between the four clusters.



(b) Activities barchart sorted per cluster, showing different color variations and height of the bars between the four clusters.

Figure 4.6: Four clusters using `pam` with the *Tanimoto* distance.



The best model is shown in Figure 4.6, creating four relatively balanced clusters with the Tanimoto distance using `pam`. The selected number of clusters is based mainly on the activity frequencies visualized by Figure 4.6b and activity frequency histograms per cluster. From these visualizations we can see whether we have found distinct clusters, as the silhouette value returned by `pam` indicates the highest clustering quality is for a large number (about twenty) of clusters. The `pam`-algorithm has a risk of reaching local optima, but since multiple executions of the algorithm using different random initializations of its medoids return the same visualizations, we will assume for the rest of this study that one execution of the `pam`-algorithm is sufficient.

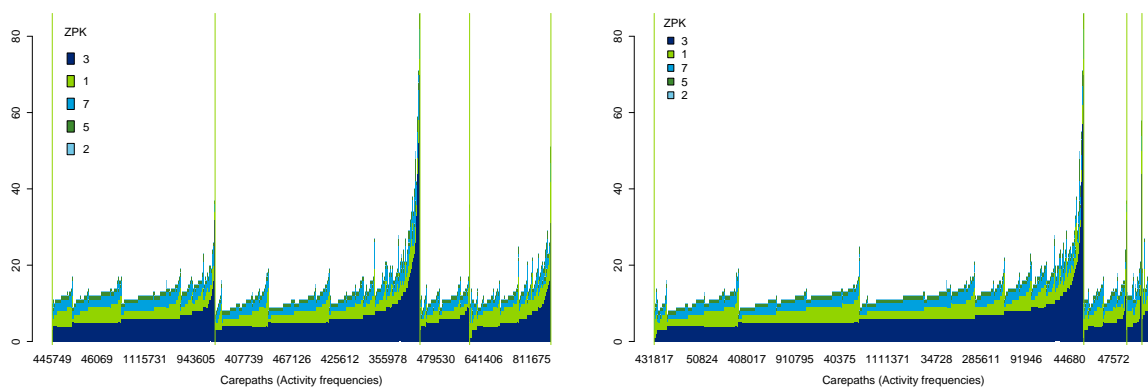
The Cosine distance completely fails to separate the longer carepaths from the shorter paths. Earlier, we explained that the Cosine distance is often used in text mining due to its ability to ignore common 0-values. This advantage is lost however since we are working with such a small, filtered subset of the initial activities. The Euclidean distance appears to offer better performance than the Cosine distance, as it groups a small number of long (and thus exceptional) carepaths together. However, the distinction between other clusters is less obvious, and the number of exceptions caught by Tanimoto is larger.

Although it seems that `hclust` offers a more balanced clustering than `pam`, this is at the expense of the distinction between individual clusters. The latter technique clearly distinguishes the exceptional cases, and offers a clear distinction between number of nursingdays (ZPK 3) and outpatient department (ZPK 1) for the remaining clusters.

The addition of an extra cluster (five instead of four) results in a more specific clustering of exceptional cases, but – similar to our argument to disregard the results for six clusters – this does not outweigh the gain in overall clarity of a lower number of clusters, as the gain in distinction between the remaining clusters is negligible. More details on the specific clustering results for this DBC are presented in the case study in Section 5.1.

4.4.1.2 Using String representations

String clustering is based on the same dataset as used for vector clustering in the previous section. The advantage of string clustering is that the ordering of activities is maintained when comparing individual carepaths, the disadvantage that this is at the expense of cluster quality.



(a) GZIP compression for 4 clusters (`hclust` in R)

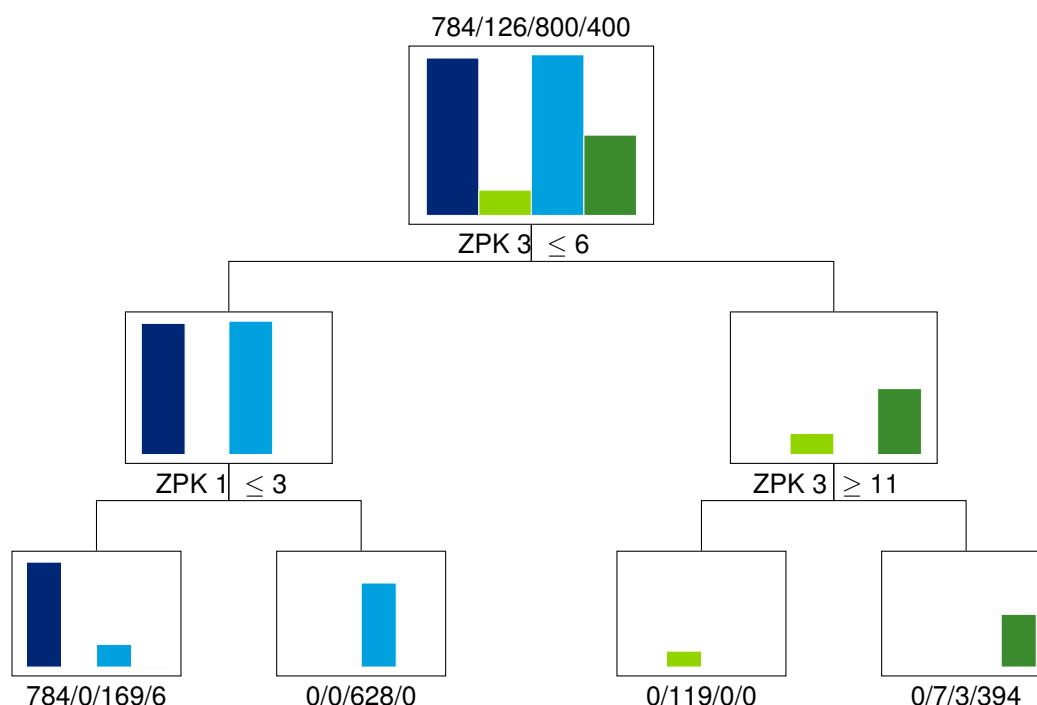
(b) Bzip2 compression for 5 clusters (`pam` in R)

Figure 4.7: Two examples of bad clustering.

Regardless of the compression algorithm and clustering method, the complete set of results shows a large spread in total number of activities within different clusters, similarly to the results as shown in Figure 4.7 (see Appendix B for more visual impressions). This is undesirable, as this number is what defines both financial homogeneity (more activities are more expensive) and logistics homogeneity (e.g. the length of hospitalization and number of operations) as described in Section 2.1.1.3.

4.4.2 Classification of resulting clusters

In the second modeling stage, we try to identify the main characteristics of the different clusters gained from the pam model for the Tanimoto distance. We use a standard R library rpart as described in Section 3.2 to build our tree. First, we grow the entire tree on the entire input data we used for clustering, and then we will evaluate the different cutoff points in search for the best tree. For the many different



(a) Classification tree for ZPK 1 and 3: the description between the nodes represent the decision-rule (the left child satisfies the rule, the right does not), the barcharts represent the number of paths for each of the clusters in the node, and the numbers at the leafs the final number of paths per cluster, per leaf-node.

(b) rpart error

CP	nsplit	xerror
0.3840	0	1.046
0.0908	2	0.232
0.0019	3	0.141
0.0015	5	0.143
0.0008	8	0.140
0.0000	13	0.134

(c) Cluster labeling based on Classification tree

#	Label	Description
1	Short	6 or less nursingdays, 3 or less outpatient department visits
2	Long	Exceptional cases up to 70 nursingdays
3	Short	6 or less nursingdays, relatively more outpatient department visits
4	Medium	Between 7 and 11 nursingdays

Figure 4.9: Classification of resulting clusters.



resulting trees it becomes obvious that ZPK 1 and ZPK 3 are the most important attributes to discriminate between clusters, which corresponds to what we see in the figures in Appendix D.1. Figure 4.6 also supports these conclusions, and also show the importance of the ratio between ZPK 1 and 3. For the classification tree however, adding this ratio does not improve the quality.

Figure 4.8b shows the error rates for the tree built with ZPK 1 and 3. In many cases, the absolute lowest value for the `xerror` would indicate the best possible CP-value. However, a lower CP-value indicates more splits, which may result in a single cluster classified by multiple nodes, which is not preferable as this implies clusters are described by multiple sets of characteristics. Instead we look for the smallest number of `nsplit` for the lowest value of `xerror`. With only two activities (ZPK 1 and 3) required to predict the according cluster, we have an insightful tree using only *seven nodes* as opposed to the *thirteen splits* the lowest error rate offers. Three out of four clusters are classified almost entirely in a single leaf node, which is preferable because this allows us to describe a cluster by a single set of characteristics. Only cluster 1 and 3 are similar, and cannot easily be distinguished without additional information or attributes (which is also visible in Figure 4.6). In total, less than 9 percent of all paths is incorrectly classified, which experts considered acceptable considering the simple general description of different carepaths provided by the tree.

The simplicity in this classification model offers an easy overview of the characteristics of the different clusters. The four clusters the carepaths are divided into can be labeled by length of hospitalization according to Table 4.8c. A detailed analysis of the resulting classification is given in Section 5.1.

4.4.3 Identification of frequent patterns

In the third and last modeling stage process mining is applied to identify frequent patterns for each cluster, resulting from the data mining process as described in the previous sections. We export the Process ID's for each cluster and create a subset of the original data. The four individual subsets are read into ProM and analyzed using the Trace Alignment plugin. Note that we use a single cluster for the Guide Tree in order to enable the plugin whilst preserving our original clustering. We sort the alignment in an effort to visualize the most important patterns, e.g. by putting the longest traces at the top. The Trace Alignment also returns the most frequent pattern (i.e. the pattern described by more than half of the traces) and provides an insightful visualization for deviations. The results are shown in Appendix D.2.

As a final step, a small visualization of the Heuristics Miner gives a global view of the possible flow. As we expected in Section 3.3.1, the resulting models are extremely simple. Since they lack the number of repeated activities etc., its usability is limited to the global comparison of clusters. The models for the patient careflow described in the previous sections are shown in Figure 5.3, together with the main process pattern found by the Trace Alignment plugin.

4.5 Conclusion

The development of the analysis proposed in this study is based on CRISP-DM, which offers a solid basis for any type of data analysis project. This includes the application of our methodology as described in Figure 4.10. However, as the main goal of this methodology is one-fold, the first stage of CRISP-DM does not have to be repeated: a thorough Business Understanding of the healthcare domain is

described in Sections 2.1 and 4.1, which also define the main business objective: *to provide insight into patient careflow on a logistics level*. Besides, this methodology is designed for a collection of standardized DIS-data and a number of readily available external sources providing more detail on individual activities. A detailed description of these data is offered in Section 4.2 (Data Understanding). It is at the physical Data Collection stage that we start with the description of our methodology – we call it *Step 0*, as it needs to be performed only once for multiple analyses.

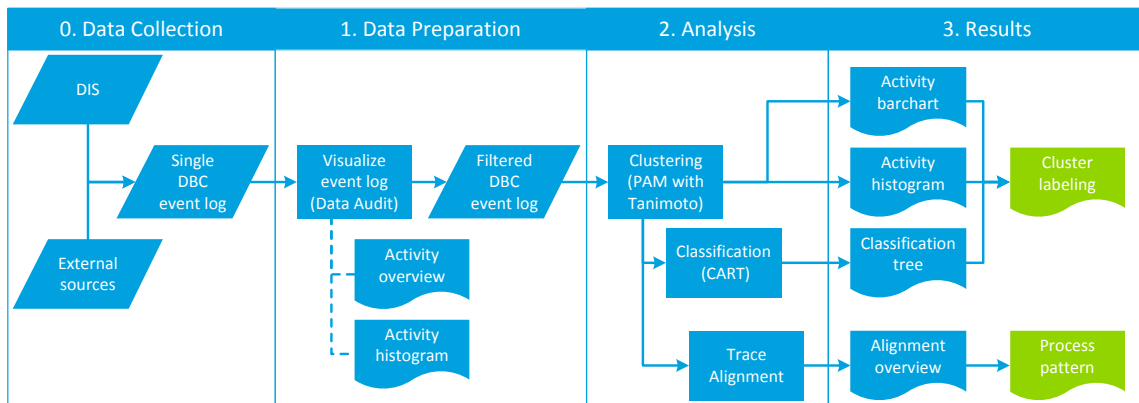


Figure 4.10: Methodology overview.

Step 0 Data Collection

The analysis in this methodology is based on standard data structures: DIS-data containing patient information and activity sequences or carepaths, description lists of DBC’s and activities, and cost overviews on both a CTG and ZPK level. The required data modeling phase is thoroughly described in Sections 4.2, 4.3 and Appendix A. Once it is set up however, the database can be used for numerous analyses.

Step 1: Data Preparation

The focus of this step is on the filtering and selection of the right activities used in the analysis. It also entails the filtering of infeasible or distorting carepaths. The choices for filtering and selection are based on the activity overview plot and activity histograms (e.g. Figures 4.1a and 4.4, which indicate data differences between hospitals and infeasible activity frequencies). More examples are provided in the case studies of Chapter 5.

Step 2: Analysis (Modeling)

The thorough exploration in the previous section identified three suitable and complementary mining techniques that provide valuable insights:

Partitioning Around Medoids (using Tanimoto distance) This clustering technique shows good performance in the *grouping* of different carepaths into groups with similar paths.

Classification (using CART) The clustering obtained with `pam` is easily classified using trees, which enables us to *describe* the different groups by their characteristics.



Trace Alignment (ProM plugin) The third step allows us to *identify* the main process for each cluster, and to identify deviations in individual carepaths.

Step 3: Results (Evaluation / Deployment)

The second part of our research objective as stated in Section 1.2 asks for “*insight into patient careflow*”. In the previous section we managed to develop a number of models and visualizations that provide us with detailed insights. More specifically, the results lead us to a number of possible process improvements:

1. Using the average cost of the individual clusters, it is straightforward to identify the most preferable (i.e. cheapest) carepath. This is a legitimate argument in favor of trying to increase the share of patients treated according to this path. However, the lack of detailed medical insight included in the analysis implies that it is still up to medical specialists to make the final decision on the provided care for individual patients.
2. The Trace Alignment provides a detailed overview of the (logistics) process for a specific type of patients. It also shows *many* deviations from the global path within such a cluster. This overview is a valuable insight to improve cluster carepath standardization, which is proven to result in an increase of care quality and decrease of care costs (see Section 1.1).
3. Now that a limited number of different patient types are defined, a more robust operating schedule is feasible by applying sophisticated optimization techniques to schedule patients by category. Previous work has pointed out the value of categorizing patients and scheduling accordingly, as it allows for even workload distribution, prevents over-utilization of resources and decreases under-utilization of beds [71]. However, without further medical knowledge on patients, little statistical support is found on predicting the category a patient belongs to. This is often compensated with an educated guess based on years of experience of the medical specialists.

In the next chapter, we validate our methodology by applying it to a number of Case Studies. The first study is based on the same event log used to develop this methodology, but offers more detail on the final results. Two more case studies are performed to validate that the methodology is not overfitted to the case used during the development.



Chapter 5

Evaluation (Case Studies)

This chapter tests the previously developed methodology to a number of cases. The first case is used during the modeling stage: hip arthrosis (or more commonly: hip replacement). This surgical procedure is considered standardized and involves hospitalization of patients. In addition to the results shown during the modeling phase, a more detailed look at the clustering results with regards to e.g. costs and cluster description is provided. The second case study is similar to the first: an orthopedic surgical DBC that describes arthrosis carepaths, commonly known as a knee replacement. The third case is known medically as a malignant breast neoplasm (or more commonly: breast cancer). These cases are all considered standardized surgical procedures, and entail patient hospitalization.

5.1 Arthrosis (hip) - surgical/clinical with joint prosthesis

During the modeling phase of Chapter 4 we described the first two steps of our methodology, *Step 0: Data Collection* and *Step 1: Data Preparation* based on this DBC. In order to be thorough, we start with a summary of Step 1 before continuing to the actual modeling stages.

5.1.1 Step 1: Data Preparation

As described in Section 4.4 we limit the dataset to ZPK's 1, 2, 3, 5 and 7. Figure 4.1a showed that ZPK 4 and 6 only occur in one or two hospitals, which would bias clustering towards differentiating between hospitals. We also decided – with the support of our medical experts – to ignore labtests (ZPK 8) because of their high frequency (Figure 4.4 shows an average of almost 20 activities for ZPK 8, compared to 6.5 nursingdays (ZPK 3) as second-most frequent activity), whilst they have little effect on the logistic process.

5.1.2 Step 2: Analysis

Clustering The first step is to group the different carepaths into similar groups, by applying `pam` using the Tanimoto Distance. The selected clustering result from Section 4.4 is shown in Figure 4.6b. An overview of the different patterns described by the clusters is shown in the parallel coordinates plot of activities within clusters (Figure 5.1).

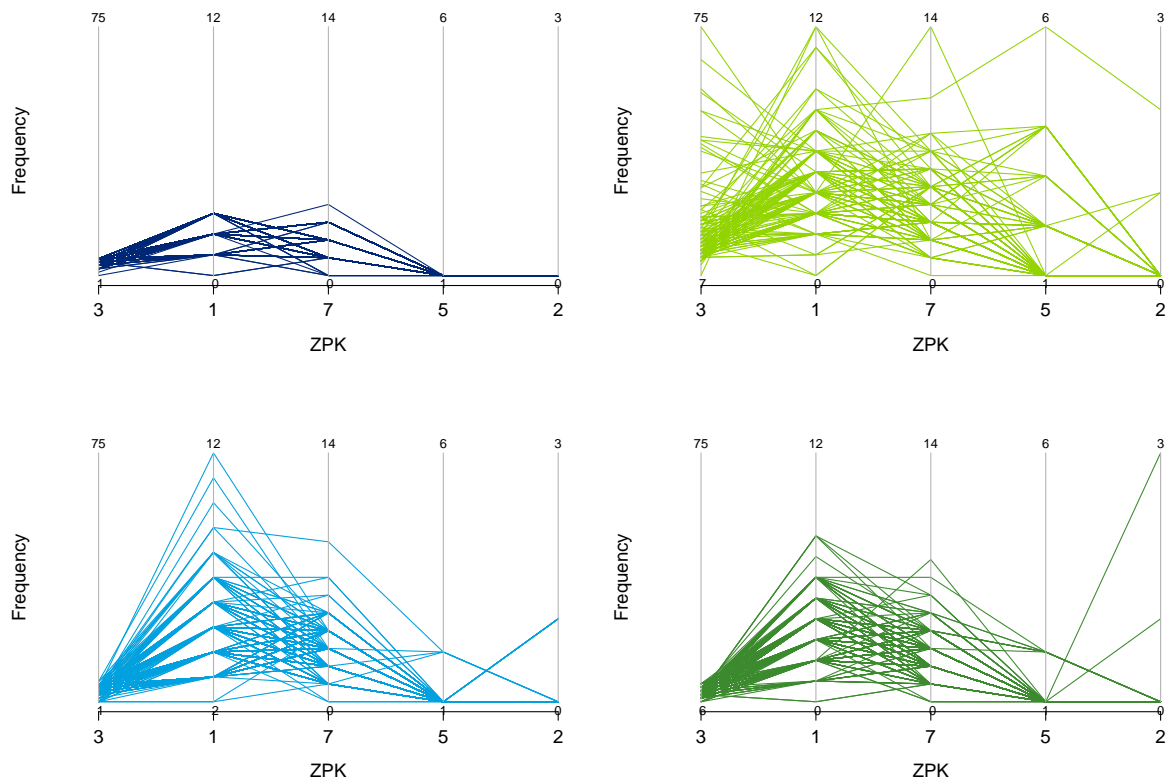


Figure 5.1: Parallel plot per cluster (clusters 1 to 4 are plotted from left to right, top to bottom), with the exceptional cluster 2 (right-top) showing the highest frequencies for all activities. It also shows that although cluster 3 (left-bottom) appears short, it has many outpatient department visits (ZPK 1).

In addition to the parallel coordinates plot, we look at the actual activity frequency distributions within clusters per ZPK, as shown in Figure 5.2. This visualization shows the actual number of exceptions and deviations, something the parallel coordinates plot does not offer.

Combined, the different views from Figures 4.6b, 5.1 and the histograms in Appendix D.1 give some indication of the type of patients described by each cluster. The longest carepaths are grouped in cluster 2 as these paths have the highest number of nursingdays, diagnostic activities and surgeries performed. Cluster 1 and 3 appear similar, except that patients in the latter visited the outpatient department more often. In the next section, classification is applied for a thorough analysis of the differences between these clusters.

Classification The second model classifies the clustering above as shown in Figure 4.9. As described in Section 4.4.2, we try to find a tree that identifies each cluster with preferably just one or two different descriptions. Unlike with many other classification models, this means we do not necessarily select the tree with the best fit (or lowest `xerror`), but rather use common sense to select a good (or at least sufficient) fit.

Using just two activities over three splits, the tree selected for this analysis is able to correctly classify over 91 percent of all carepaths. The tree contains exactly four leaf nodes, where each leaf represents one cluster. The best fit spreads the four clusters over 14 nodes, with an increase

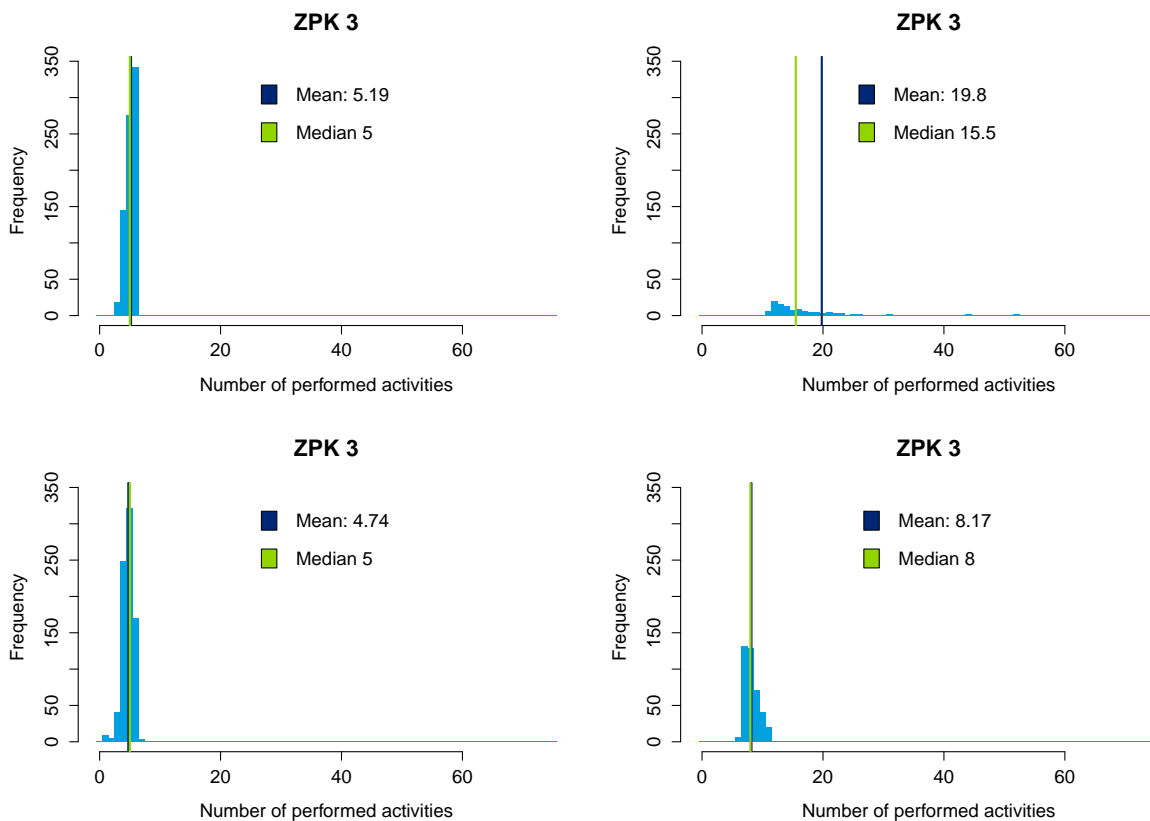


Figure 5.2: Nursingday (ZPK 3) frequency distributions per cluster: the clusters are given from left to right, top to bottom. The histograms for the other ZPK-activities are described in Appendix D.1.

of only about 1 percent in correctly classified paths. If we take into account all available activities, we can correctly classify the entire set using 25 nodes. Experts agreed that this increase of fit does not outweigh the increased complexity of the tree.

Table 5.1: Cluster labeling based on classification tree. These labels are created by medical experts, and give some indication of the type of care profiles a cluster describes.

#	Label	Description
1	Short	6 or less nursingdays, 3 or less outpatient department visits
2	Long	Exceptional cases up to 70 nursingdays
3	Short	6 or less nursingdays, relatively more outpatient department visits
4	Medium	Between 7 and 11 nursingdays

Process Mining When the Heuristics Miner is applied to the same event logs, we get the process models as depicted in Figure 5.3. The Trace Alignment offers insightful images, although unfortunately the formatting of the images offers little flexibility for larger clusters and longer carepaths. From the different plots above we could already conclude that cluster 2 is relatively small (about 6 percent of all carepaths) describing the exceptionally long traces, which is supported by the alignment in Figure D.4. Clusters 1, 3 and 4 actually contain too many traces to plot clearly on a single

page in Figure D.5, though support the short patterns as indicated by previous visualizations.

For overview and comparison purposes, we combine the results of the Heuristics Miner and the main traces from the Trace Alignment in Figure 5.3.

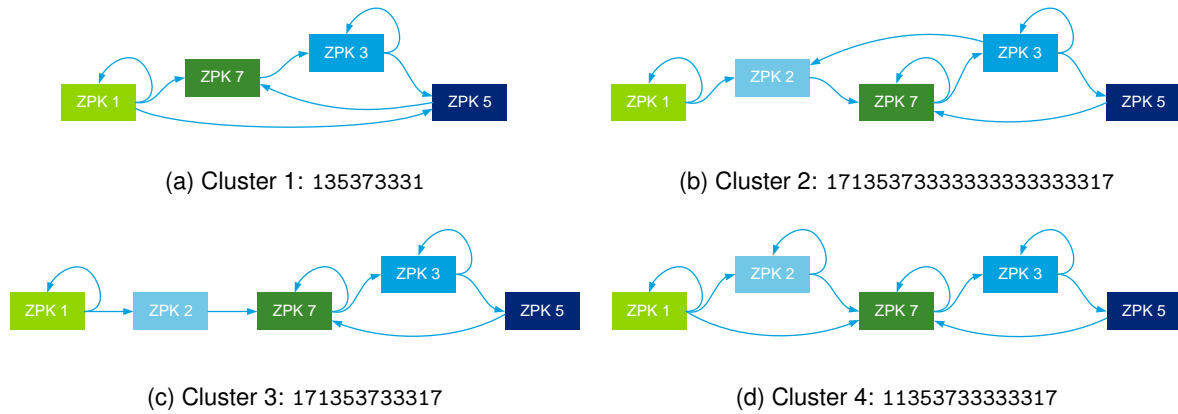


Figure 5.3: Heuristics Miner process models with main patterns from the Trace Alignment. These main patterns describe ZPK-activity sequences that occurred for at least half of the carepaths. Each activity is represented by the number of its class.

5.1.3 Step 3: Results

The deliverables for this analysis consists of two main items based on the actual process: first, we have the clustering result as visualized in Figure 4.6, and the results from the classification tree in Table 5.1. Combined, these visualizations offer a clear division of carepaths into four groups with specific characteristics. The parallel coordinates plot in Figure 5.1 and the activity histograms in Appendix D.1 provide a more detailed view of these characteristics. Secondly, the main activity patterns and process models are described in Figures 5.3 and Appendix D.2. The figures in the appendix also give a good indication of deviations within the standardized processes.

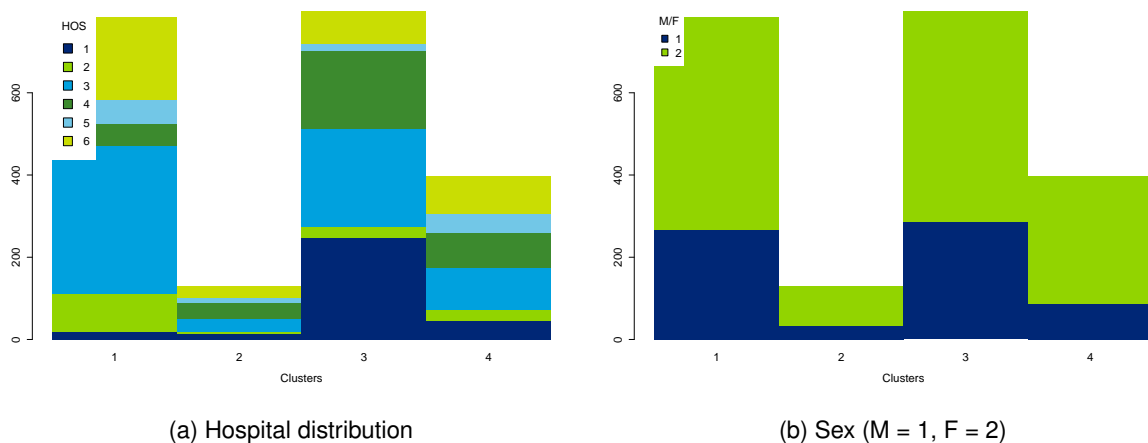


Figure 5.4: Neither a specific hospital type nor the patient's sex are dominant in any cluster.



Additionally, we have data on patients' age, sex and the type of hospital. Figure 5.4a indicates that none of the clusters is dominated by a specific type of hospital, although hospitals 1 and 4 have a relatively high number of patients in cluster 3. The available data on patients was limited and is plotted in Figure 5.5 and 5.4b. The fact that there are more women (obviously) does not add any insights to the clustering. Age however is an interesting factor: we have seen that patients in clusters 1 and 3 require less nursingdays, Figure 5.5 shows that these patients are generally younger than patients from clusters 2 and 4. This type of insight could offer support for predictive modeling, which may help improve scheduling and resource management in future projects.

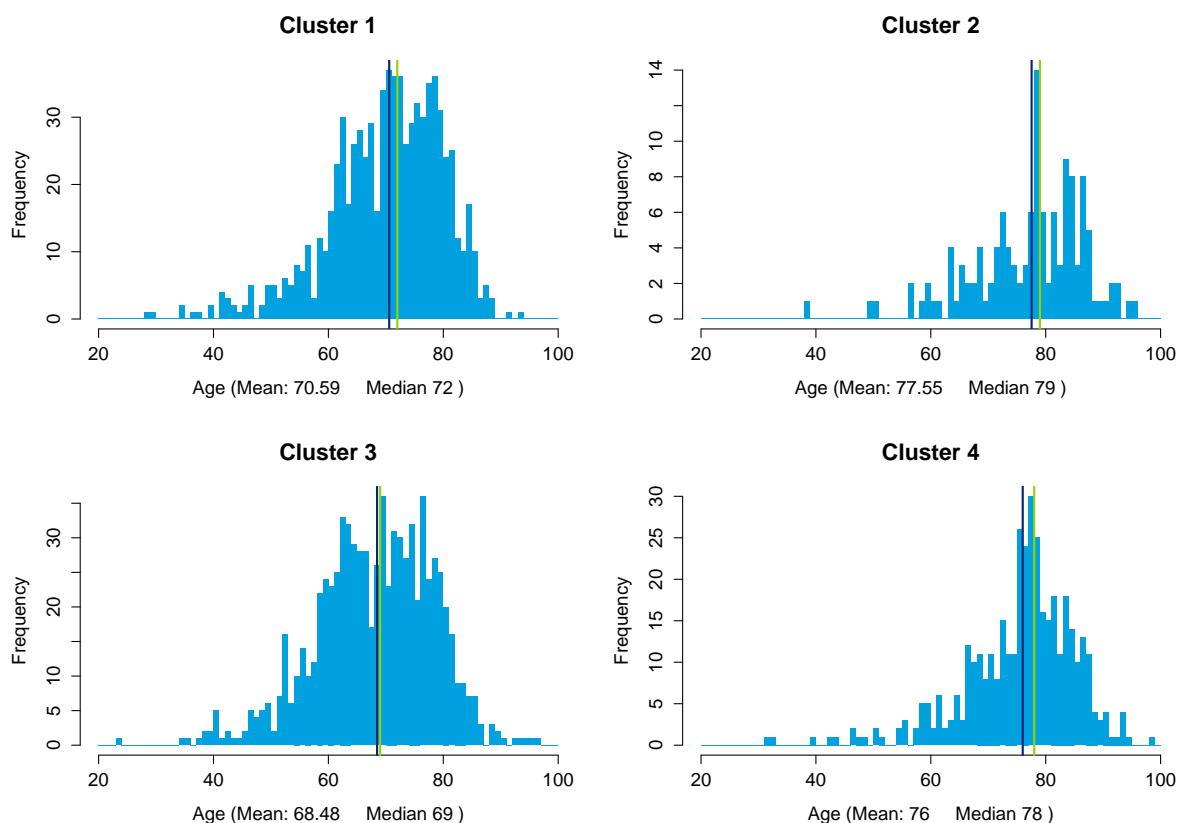


Figure 5.5: Although the mean and median ages differ between clusters, each cluster contains patients of a broad range of ages, which implies age is not a predictive or decisive factor in the selection of a carepath.

During the Data Preparation phase we have performed extensive abstraction and filtering of the original dataset. An interesting insight gained from Figure 5.6 is that although the total number of activities has decreased with almost 60 percent using distinct ZPK-activities instead of individual CTG-activities, the relative number of activities is still similar to the original CTG-activities. The same goes for the “real” price versus the roughly estimated price: even on a limited set of activities and prices, the relative *estimated* cost per cluster is similar to the *real* cost per cluster. Another interesting point is the number of “other activities” (the activities that could not be linked to a ZPK-code), these are low overall. Cluster 2 shows the highest number of outliers, which could be explained by the fact that these patients require special care (such as new and innovative – and therefore unregistered – operations).

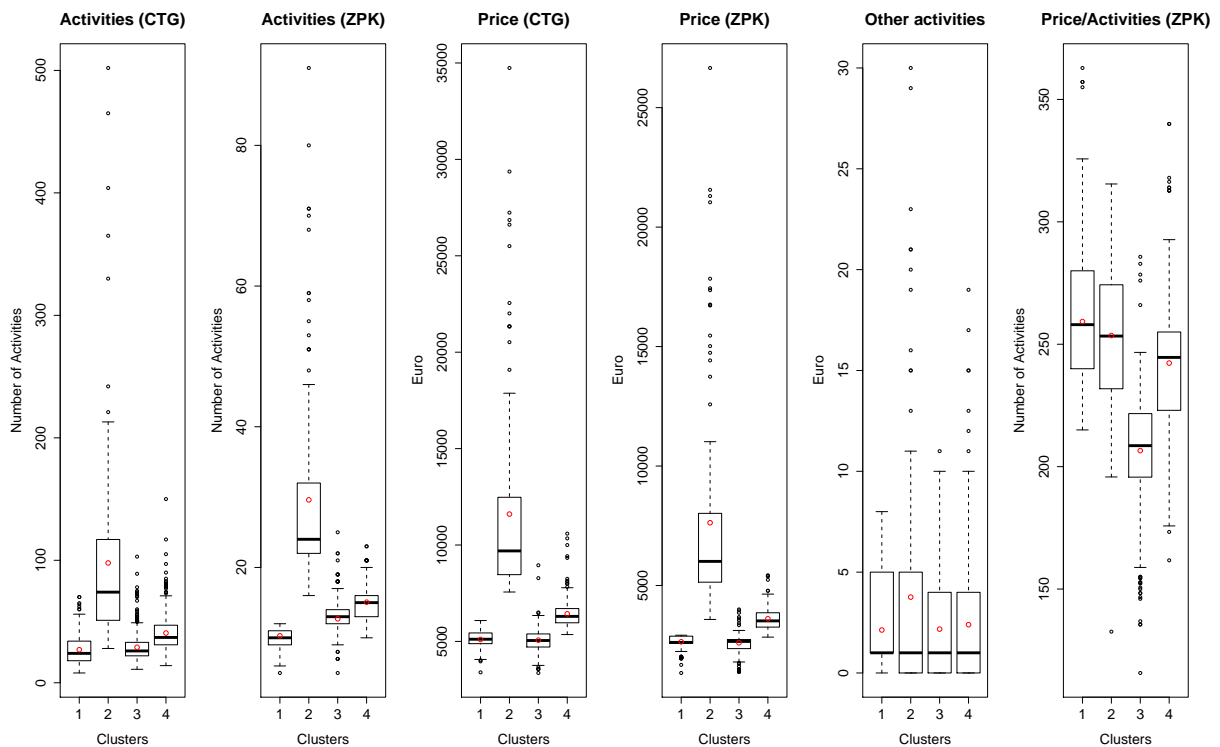


Figure 5.6: Boxplots for number of activities, carepath costs (real and estimated) and their ratio. The red circle denotes the average value for each column, whereas the thick black line in the middle of the plots denotes the mean. The box covers the values for 50 percent of the entire sample, whilst 95 percent of is covered by the area described by the dotted line. This allows us to evaluate the spread of values for each cluster in one overview. The remaining 5 percent are considered outliers, and their values are denoted by the black circles.

The last column of Figure 5.6 shows the relative price of the carepaths described by the four different clusters. Although the total number of activities for cluster 1 and 3 is similar, the latter has lower cost due to the higher number of (cheap) outpatient department activities, whilst cluster 1 entails more nursingdays. This insight indicates a possible financial profit from having patients come in at the outpatient department, instead of hospitalizing them. The downside of this carepath is that – statistically speaking – it shows a higher risk of requiring “extra” activities, as the number of outliers in the first two columns is higher for cluster 3 than for cluster 1.



5.2 Arthrosis (knee) - surgical/clinical with joint prosthesis

The next case study is similar to the first case: treating *knee* arthrosis by surgically implanting a joint prosthesis, which requires the hospitalization of patients. As the initial step in our methodology, *Step 0: Data Collection*, has already been performed, the required data is easily exported. We start the analysis at the Data Preparation stage.

5.2.1 Step 1: Data Preparation

The initial dataset is visualized in Figure 5.7a and 5.7c. These plots point out a number of data issues:

1. This DBC is *surgical*, but (5.7a) shows that the fifth hospital has recorded a number of carepaths without surgical activities (ZPK 5). We return to Step 0 to remove these paths from the dataset.
2. According to (5.7a), diagnostic activities (ZPK 4) are mainly registered in hospitals 5, 6, and a little in 4. This is unfortunate, as experts expect this analysis may lead to valuable insights that support the decrease in number of diagnostic activities. Similarly to the previous case study however, we disregard this activity at this stage.

Similarly, other therapeutic activities (ZPK 6) are mainly registered in hospital 2. For this DBC, experts indicated these activities are not relevant and they are also removed from the set.

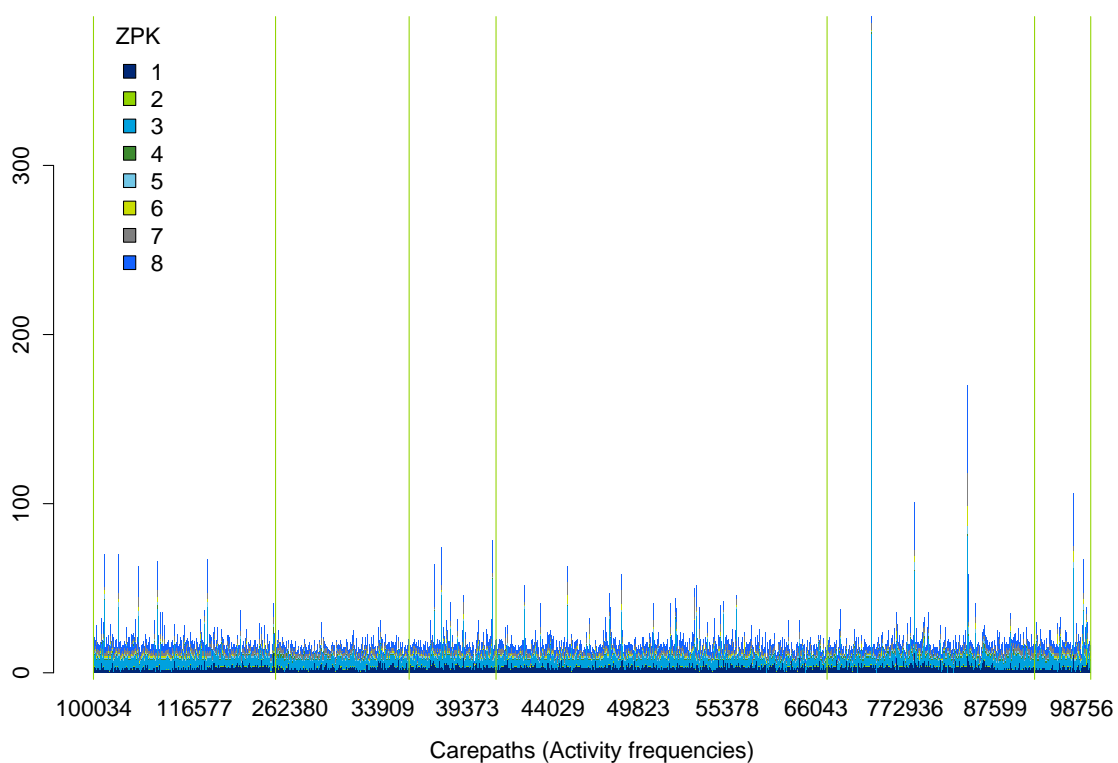
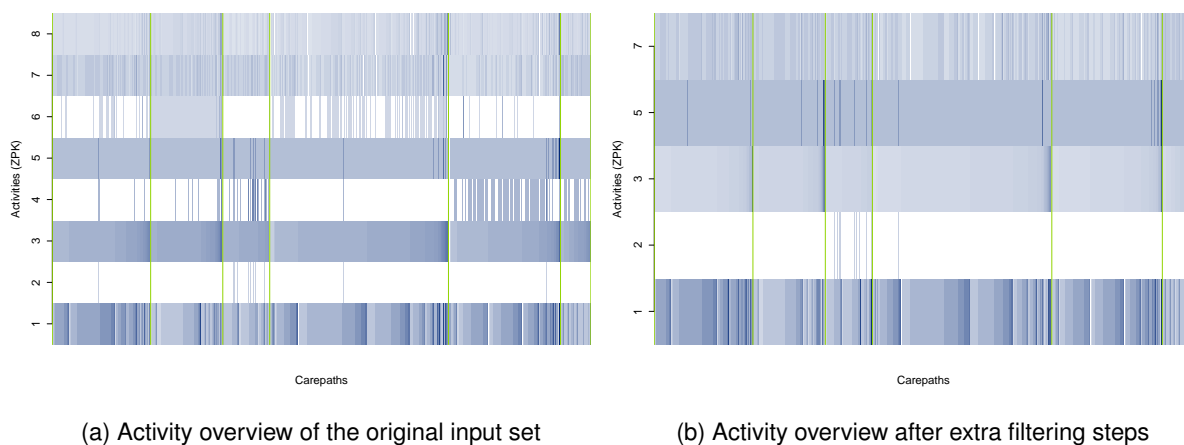
3. In (5.7c) we notice a number of exceptional cases. Although some of these exceptions are realistic, one carepath shows hospitalization of more than 366 days. Not only does this appear unlikely, the DBC system restricts the length of one DBC to a maximum of one year, which forces us to return to Step 0 to remove this entire path from the dataset.

A visual impression of the remaining dataset is given in Figure 5.7b. This figure shows little distinct differences between the different hospitals, and only contains the same activity classes as the previous case study. We now continue to perform the actual analysis based on this filtered and audited dataset.

5.2.2 Step 2: Analysis

Clustering For the first step, we use the Partitioning Around Medoids (`pam` in R) technique, in combination with the Tanimoto distance between individual carepaths. The input for this step is the event log resulting from the previous audit and filter step (Figure 5.7b). In order to find the optimal clustering, we apply the algorithms for a range of clusters. The stacked barchart visualizing activity frequencies is used to assess and compare the different clusterings. Figure 5.8 shows the activity frequency barchart for 4, 5 and 6 clusters: 4 clusters results in distinct categories, more distinct than using 6 clusters (cluster 6 contains carepaths similar to the paths in clusters 1 and 2). However, 5 clusters (Figure 5.8c) shows the most distinct clusters with the highest level of detail, and this is the model we use for the rest of the analysis steps.

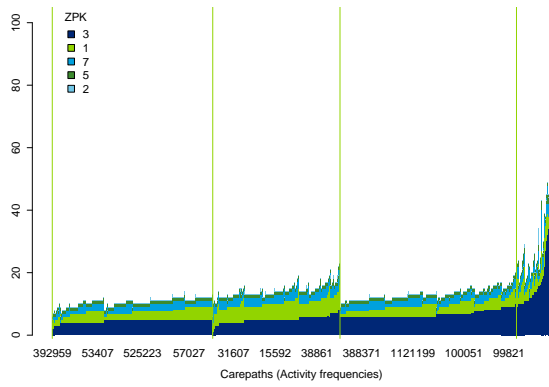
To provide further validation of the quality of the selected clusters we zoom in on individual activities. Like for many carepaths, ZPK 1 and 3 are the most variable in both number and cost. Although a ZPK 5 activity (the actual surgery) is more expensive, this is generally performed only



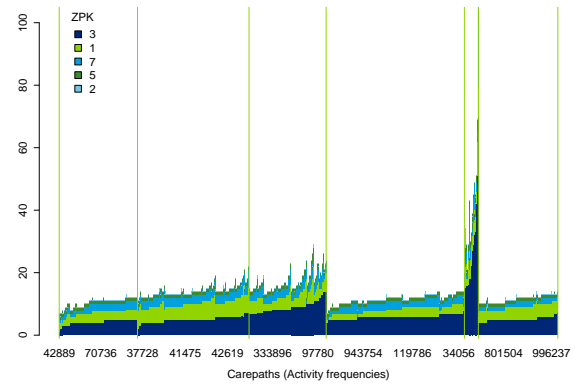
(c) Activity frequency barchart

Figure 5.7: Carepath overviews per hospital

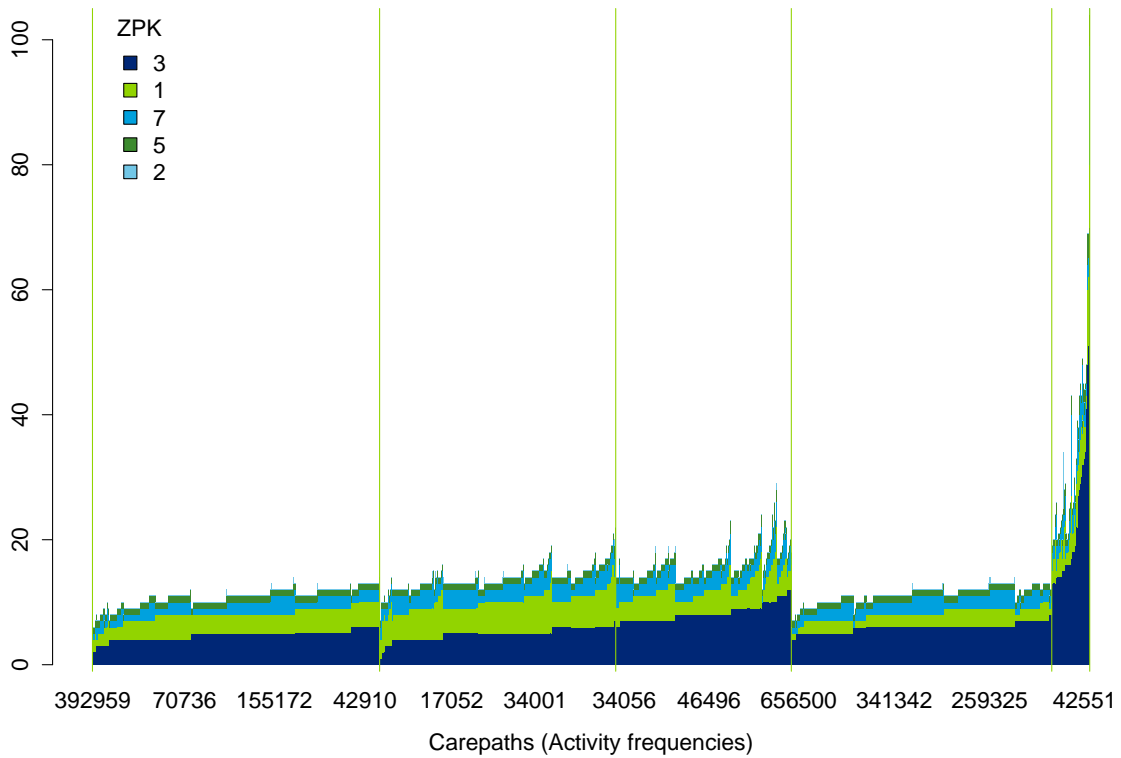
once. Besides, experts indicated that a second surgery is never optional, but an absolute requirement, and it is not something we can improve by e.g. standardization. The detailed descriptions for ZPK 3 are shown in Figure 5.9, the remaining activity histograms are described in Appendix E.1.



(a) 4 clusters



(b) 6 clusters



(c) 5 clusters

Figure 5.8: Clustering results for 4, 5 and 6 clusters.

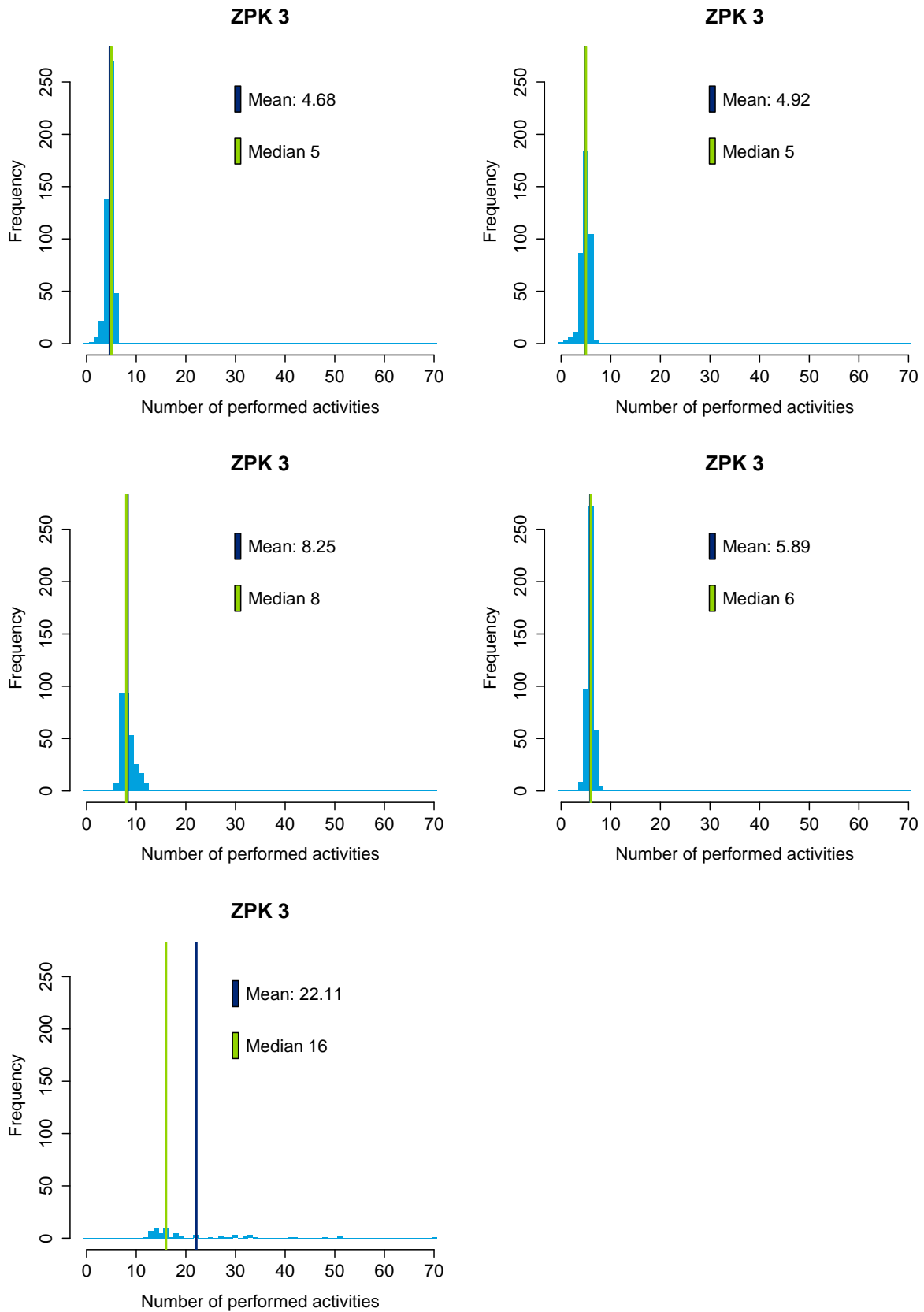


Figure 5.9: Histograms for nursingday frequencies per cluster (clusters are ordered from left-to-right, top-to-bottom). The fifth cluster contains the exceptional cases with longest hospitalization.



Classification In order to automatically identify the characteristics that define a specific cluster we apply the Classification and Regression Tree (CART) as explained in Section 3.2. As we stated before, CART returns a table with x_{error} -values for different tree sizes. The minimal value for x_{error} provides the best fit¹. For the purpose of this project however, we aim to describe the general patient careflow at a high level. Therefore, we try a number of different parameters (e.g. with or without individual ZPK's, or including the ratio between two different ZPK's) until the resulting tree shows the right² balance between tree-size and x_{error} -value.

After thorough analysis of the trees resulting from the different attributes, one of the best performing trees is based on ZPK 1, 3, 7 and the ratio between ZPK 1 and 3, as shown in Figure 5.10 (x_{error} -values are shown in Table C.1b). Especially the ratio between outpatient department visits and number of nursingdays proved to be an important indicator for the separation of clusters 1 and 2 from clusters 3, 4 and 5. With seven leaf nodes, just two clusters are divided over multiple nodes. Overall, this tree managed to correctly classify over 91 percent of the total number of paths. Experts agreed that the simplicity of the tree provides valuable insights and outweighs the incorrect classification of a small set of carepaths.

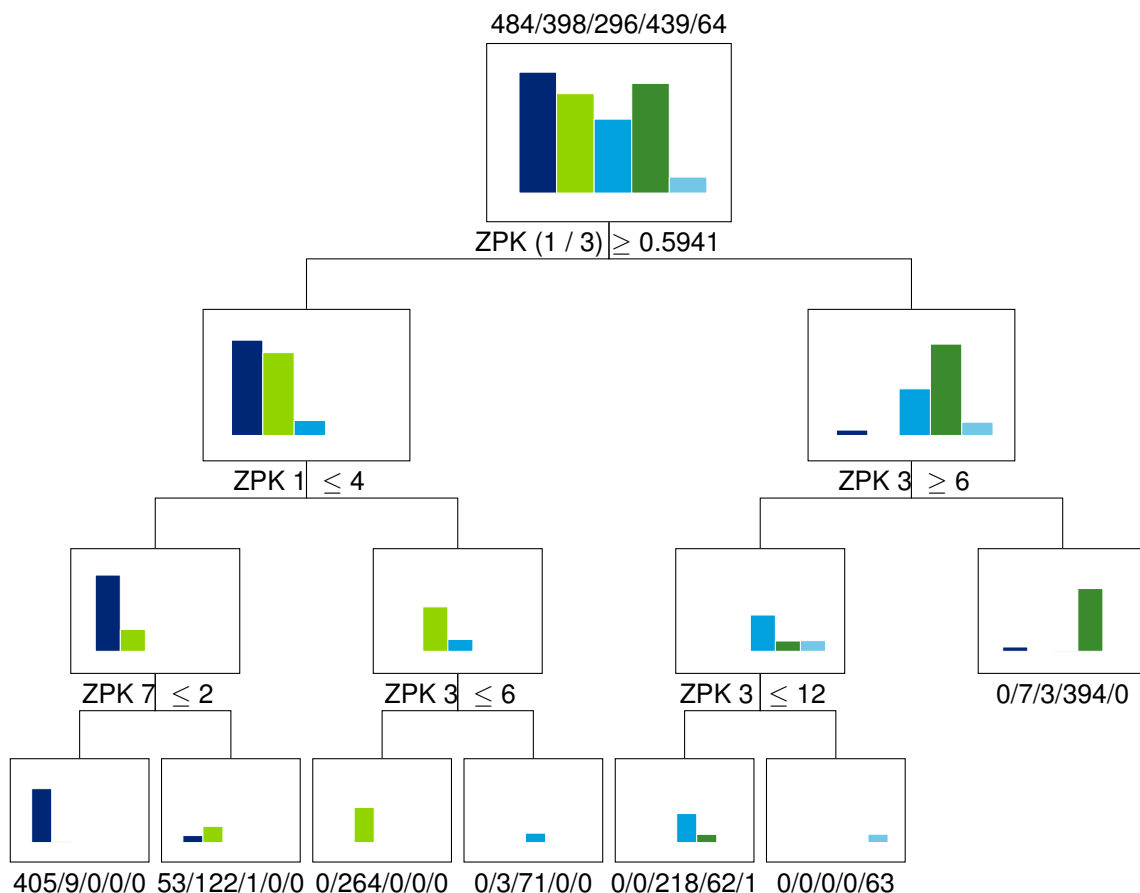


Figure 5.10: Classification tree based on ZPK 1, 3, 7 and the ratio between ZPK 1 and 3

¹As a statistically *objective* measure, the “best fit” contains the lowest number of erroneously classified carepaths.

²The value of “right” is a *subjective* measure: we look for the lowest number of erroneously classified paths in the smallest possible tree, in an optimal tree each cluster is concentrated in exactly one leaf-node.

Each cluster is labeled according to the choices made by the decision tree, which lead us to the categories listed in Table 5.2.

Table 5.2: Cluster labeling based on classification tree.

#	Label	Description
1	Short	4 or less outpatient visits with similar number of nursingdays, 2 or less images
2	Short	5 or more outpatient visits, 6 or less nursingdays
3	Medium	Between 6.5 and 12.5 nursingdays
4	Medium	More than 6.5 nursingdays with relatively many Outpatient visits
5	Long	More than 12.5 nursingdays

Process Mining Figure 5.11 shows the process models discovered by the Heuristics Miner. The process models also show the main activity patterns discovered by the Trace Alignment plugin. For readability purposes, we visualized the actual alignments in Appendix E.2.

A noticeable downside of the Heuristic Miner becomes obvious in Figure 5.11: although every carepath starts with a visit to the outpatient department (ZPK 1), this is not clearly indicated by three out of five process models that start with surgery (ZPK 5). patient careflow that starts with surgery before any type of imaging (ZPK 7) and visit to the outpatient department is infeasible – at least for this DBC. Only the models (5.11d) and (5.11e) are similar to the process models from the previous study, whereas in fact the trace alignment indicates all types of carepaths from both DBC’s are similar. This is sensible, as they are similar types of orthopedic DBC’s.

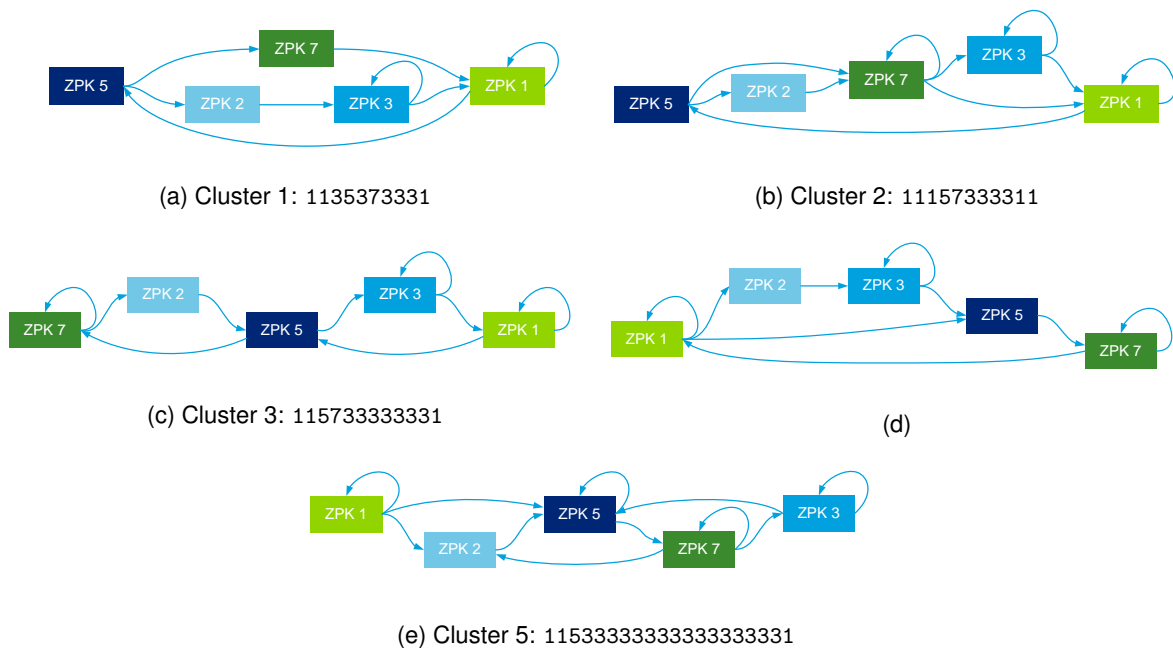


Figure 5.11: Heuristics Miner process models with main patterns from the Trace Alignment.

Similarly to cluster 2 from the previous case study, cluster 5 contains the “exceptional cases”: a small set – less than four percent of the total number of paths – of relatively long carepaths.



5.2.3 Step 3: Results

The clustering (Figure 5.8c) and classification (Table 5.2) results combined offer a clear division of carepaths into five groups with specific characteristics for each cluster. A more detailed view of these characteristics is visualized in Appendix E.1. The main activity patterns, together with a good indication of deviations from these patterns are visualized per cluster in Figures E.6, E.7 and E.8. Figure 5.11 combines the main activity pattern with the heuristic process model in a single visualization.

As previously stated, we have additional data on patients' age, sex and the type of hospital. None of the clusters is dominated by a specific type of hospital in Figure 5.13a, although hospitals 2 and 6 are mainly concentrated in cluster 4. The share of male patients per cluster is similar for all clusters (on average about 30 percent) and does not give any indication of differences between males and females. Every cluster contains patients of the ages between 40 and 98. Yet, the average age for the clusters with more nursingdays is higher (cluster 3: 74, cluster 5: 75) than for the shorter clusters (cluster 1: 69, cluster 2: 70, cluster 3: 71). This indicates that on average, older patients require longer hospitalization.

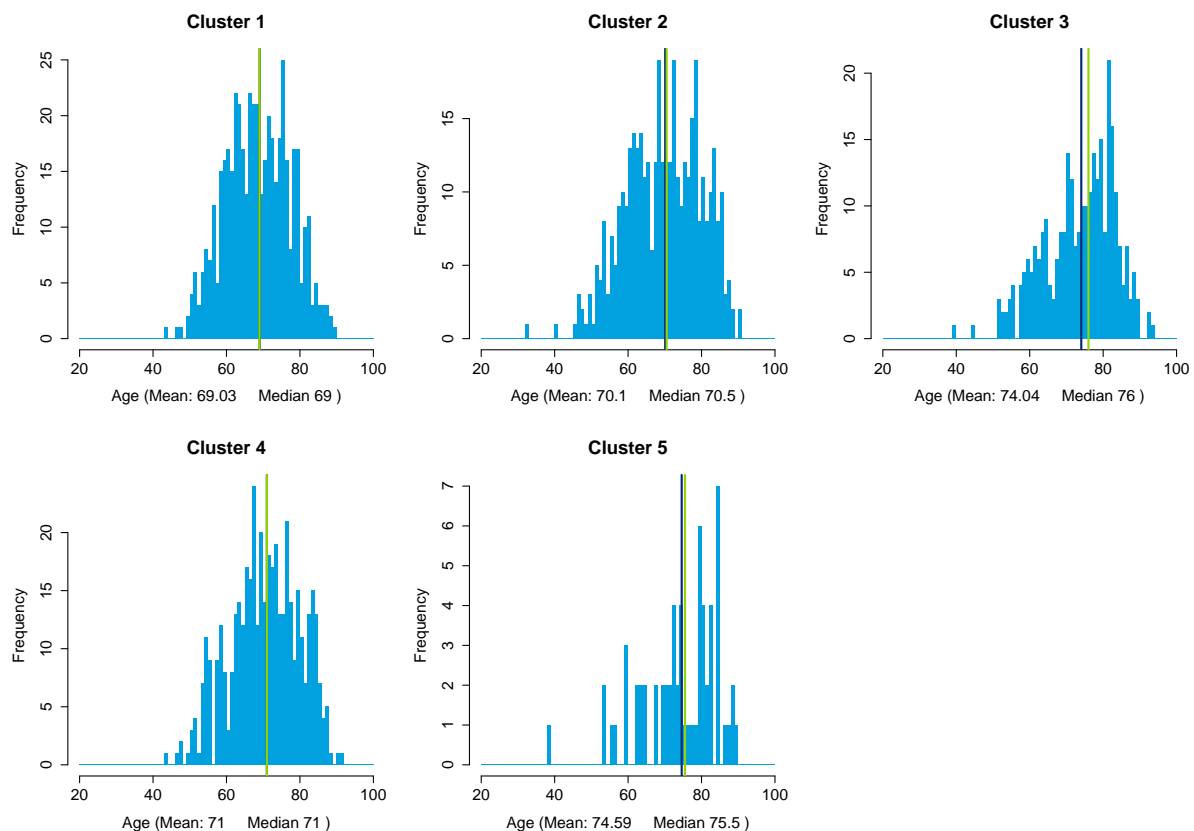
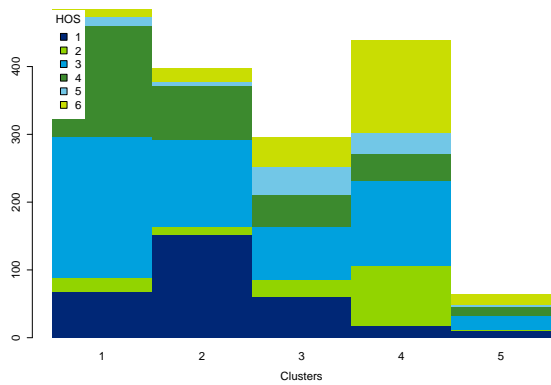
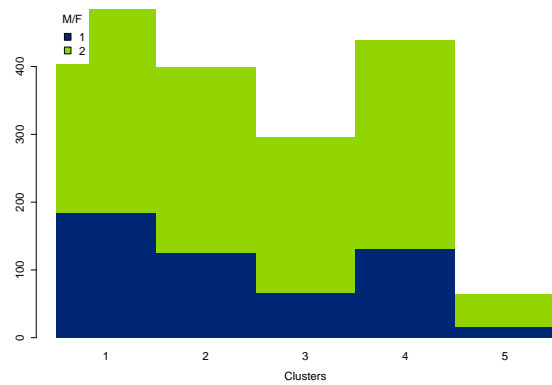


Figure 5.12: Similarly to the previous case study, age is not a decisive factor as each of the clusters shows a similar range in patient's ages.

Similar to the previous case study, the abstraction of activities to the level of activity classes decreased the number of activities by more than 55 percent, without changing the relative number of activities and costs. Also, the cluster containing "exceptional" carepaths shows the highest variance in number of both regular activities and "other activities". The overall average for the latter is generally low with about 2 other activities per individual path.



(a) Hospital distribution



(b) Sex (M = 1, F = 2)

Figure 5.13: Neither a specific hospital type nor the patient's sex are dominant in any cluster.

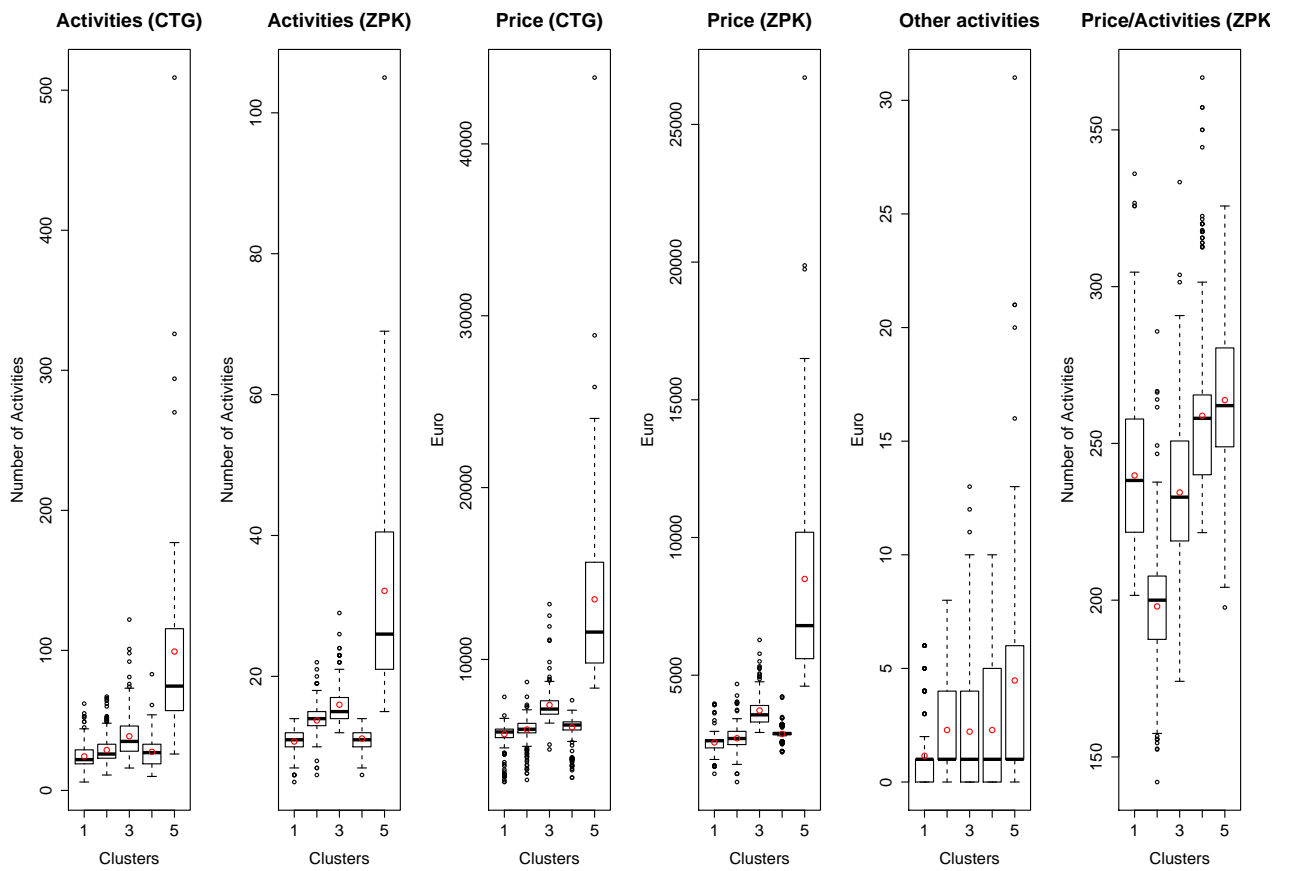


Figure 5.14: Number of activities, carepath costs (real and estimated) and their ratio.



5.3 Malignant breast neoplasm - surgical/clinical

The common description of our third case study is the surgical removal of breast cancer, which requires the hospitalization of patients. As the initial step in our methodology, *Step 0: Data Collection*, has already been performed at the development stage of this study, the required data is easily exported. We start the analysis at the Data Preparation stage.

5.3.1 Step 1: Data Preparation

The initial dataset is visualized in Figure 5.15. Both plots point out a number of data issues:

1. This DBC is *surgical*, but (5.15a) shows that three carepaths do not contain any surgical activities. We return to Step 0 to remove these paths from the dataset.
2. In (5.15c) we notice a number of exceptional cases. As none of these exceptional cases seem excessive (unlike the path from the previous case study), no further action is required.
3. According to (5.15a), hospital 2 has a relatively high concentration of carepaths with multiple surgical activities. As this hospital is one of the top-clinical hospitals, this is plausible and no further action is required.
4. For the first time, ZPK 4 and 6 registrations are not disproportionately dominated by one or two hospitals, which allows us to include them in our analysis.
5. Considering the relatively low activity frequency, ZPK 8 appears to be less dominant for this DBC compared to the previous case studies. However, during the modeling phase we observed a decrease of clustering quality when we include this type of activities, and we decided to ignore it nevertheless.
6. During our initial modeling step, we identified a number of male patients. As the activities registered for these patients were not exceptional, we simply assumed these were clerical errors and manually altered their sex.

After the data corrections and filtering of the DBC dataset, we continue to do the actual analysis.

5.3.2 Step 2: Analysis

Clustering The input for this step is the event log resulting from the previous audit and filter step (Figure 5.15b). In order to find the optimal clustering, we apply the p_{am} algorithm using the Tanimoto distance for a range of clusters. The stacked barchart visualizing activity frequencies is used to assess and compare the different clustering results. Figure 5.16 shows the barchart for 4, 5 and 6 clusters: both 4 and 6 clusters results in distinct categories, but lack the cluster for easy diagnosis as visible in (5.16c). The solution obtained with 5 clusters shows the most distinct clusters with a separate cluster for patients with little diagnostic imaging, and this is the model we use for the rest of the analysis steps.

To provide further validation of the quality of the selected clusters we zoom in on individual activities. Like for many carepaths, ZPK 1 and 3 are the most variable in both number and cost. For this

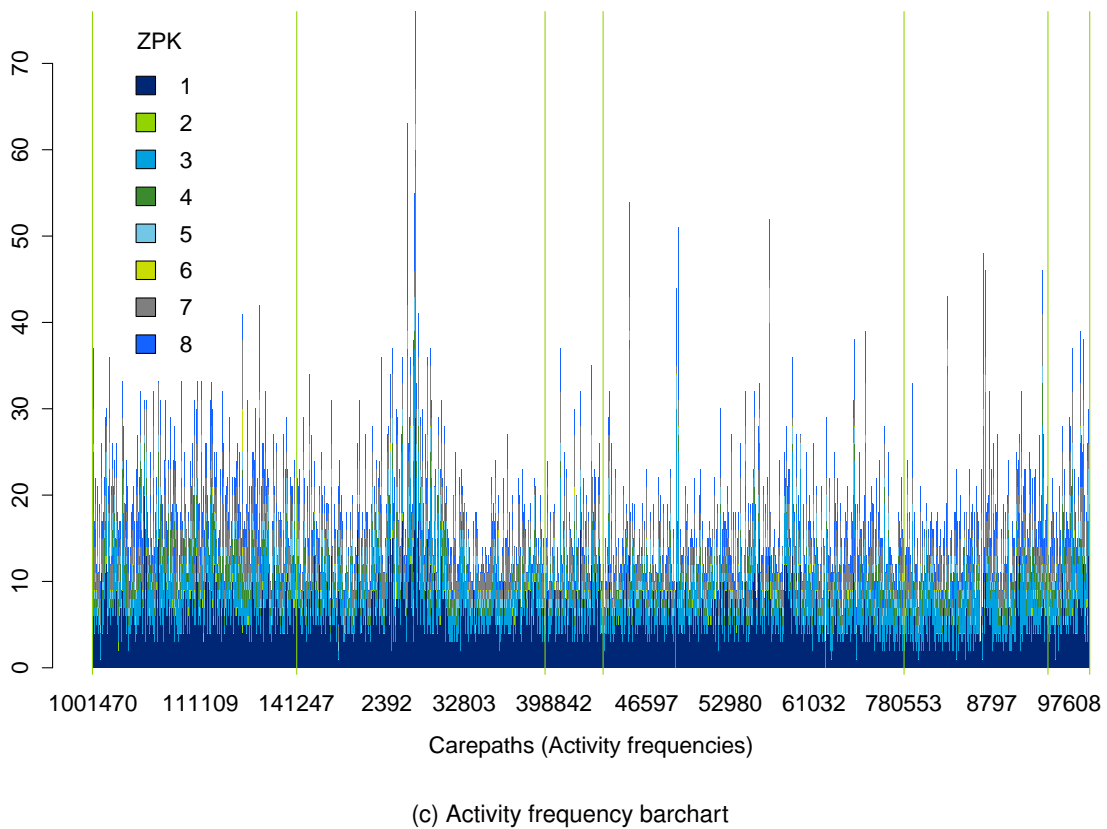
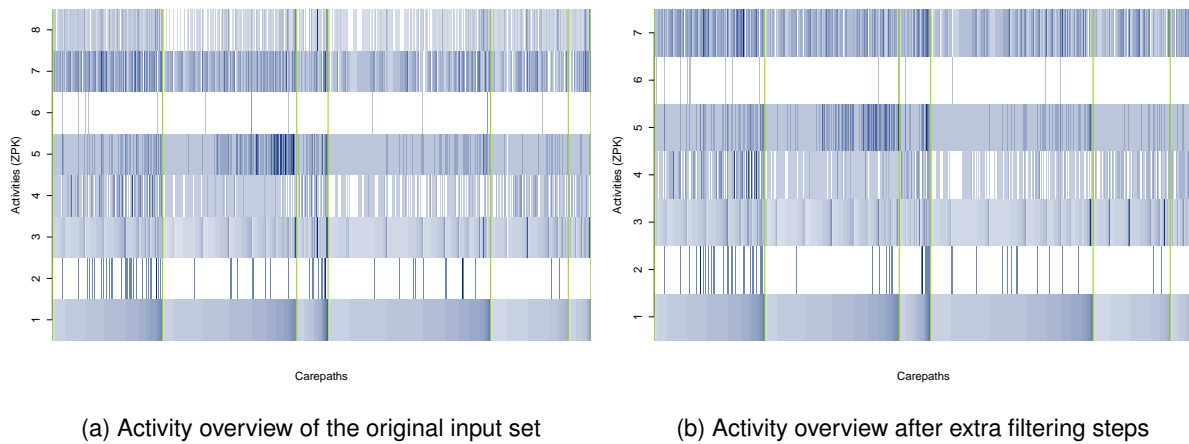
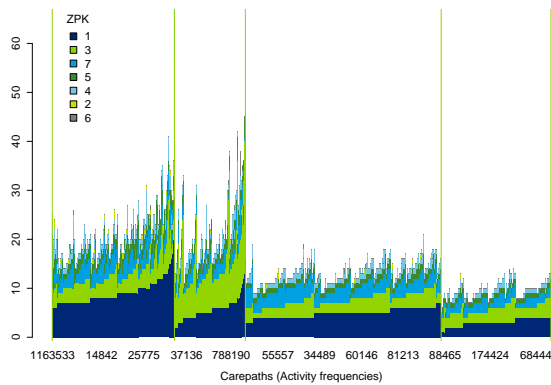


Figure 5.15: Carepath overviews per hospital

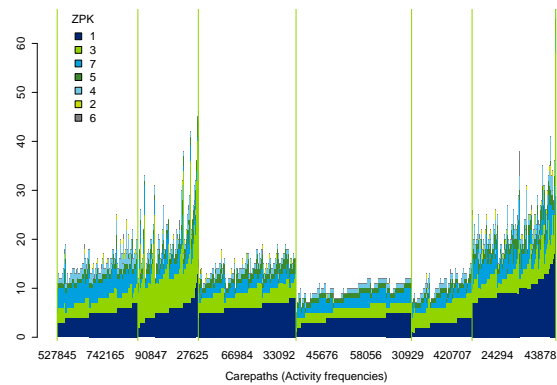
case study however, ZPK 7 (diagnostic imaging) and 5 (the actual surgery) activities also show cluster characterizing properties. These figures are shown in Appendix F.1.

A good example for these properties is cluster 4³: in most cases only one diagnostic image was required for the cancer to be discovered and surgically removed. This leads to a low number of total surgeries and nursingdays. A second cluster with similar numbers for surgery and nursingdays

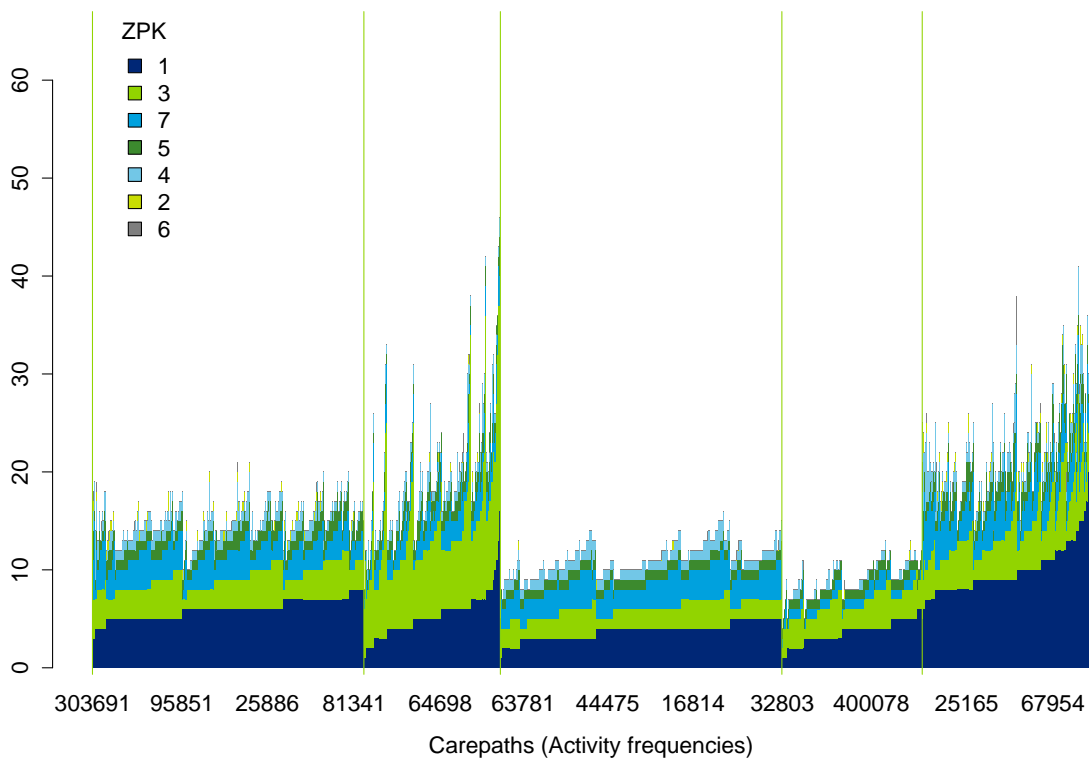
³Note: this cluster was one of the main reasons to choose 5 clusters, this paragraph points out its significance.



(a) 4 clusters



(b) 6 clusters



(c) 5 clusters

Figure 5.16: Clustering results for 4, 5 and 6 clusters.

does have an average number (about 3) of diagnostic images taken. The detailed descriptions for these ZPK's are shown in Appendix F.

Classification In order to automatically identify the characteristics that define a specific cluster we apply the Classification and Regression Tree (CART) as explained in Section 3.2. As we stated before, CART returns a table with `xerror`-values for different split numbers. The minimal value for `xerror` provides the best fit (objective measure). For the purpose of this project however, we aim to describe the general patient careflow at a high level. Therefore, we try a number of different attributes (e.g. with or without individual ZPK's, or including the ratio between two different ZPK's) until the resulting tree shows the right balance between tree-size and `xerror`-value (subjective measure).

After thorough analysis of the trees resulting from the different attributes, one of the best performing trees is based on ZPK 1, 3, 7 and the ratio between ZPK 1 and 3, as shown in Figure 5.17 (`xerror`-values are shown in Table C.1c). Especially the ratio between outpatient department visits and number of nursingdays proved to be an important indicator for the separation of clusters 1 and 2 from clusters 3, 4 and 5. With seven leaf nodes, the majority of clusters are characterized by a single leaf-node, leaving just two clusters divided over multiple leaves. Overall, this tree managed to correctly classify over 86 percent of the total number of paths. A small increase of less than 2 percent would require two extra leaf nodes, and the smallest number of incorrect classifications (3 out of 1286) would require 33 leaves. Experts agreed that the simplicity of the tree provides valuable insights and outweighs the incorrect classification of a relatively small set of carepaths.

Each cluster is labeled according to the choices made by the decision tree, which lead us to the categories listed in Table 5.3.

Table 5.3: Cluster labeling based on classification tree.

#	Label	Description
1	Medium	6 or more outpatient visits, 5 or less nursingdays
2	Long	7 or less outpatient visits, 6 or more nursingdays
3	Short	4 or less outpatient visits, 5 or less nursingdays, multiple images
4	Simple	Similar to 3, with only one image required for surgery
5	Outpatient	8 or more outpatient visits

Process Mining In the previous case study we observed process models from the Heuristics Miner that did describe patient careflow in the way we would expect or find intuitive. Similar results were gained from the event log for this case study. Although the different process models gave correct results, the model offering the most intuitive overview of the corresponding carepaths was created for cluster 5 in Figure 5.18b. This model shows patients come in at the outpatient department (ZPK 1), then continue for diagnostic activities and imaging (ZPK 4, 6 and 7) during hospitalization (for either a single day or a number of nursingdays – ZPK 2 and 3). Alternatively, patients are admitted for surgery (ZPK 5) after visiting the outpatient department, and before (or actually during) being hospitalized (ZPK 2 and 3). The addition of ZPK's 4 and 6 gives a good indication of the difference between surgical and non-surgical days.

The main patterns from the Trace Alignment plugin are given in Table 5.18a. Although the alignment was unable to find main patterns for the additional ZPK's 4 and 6, the overall pattern gives a

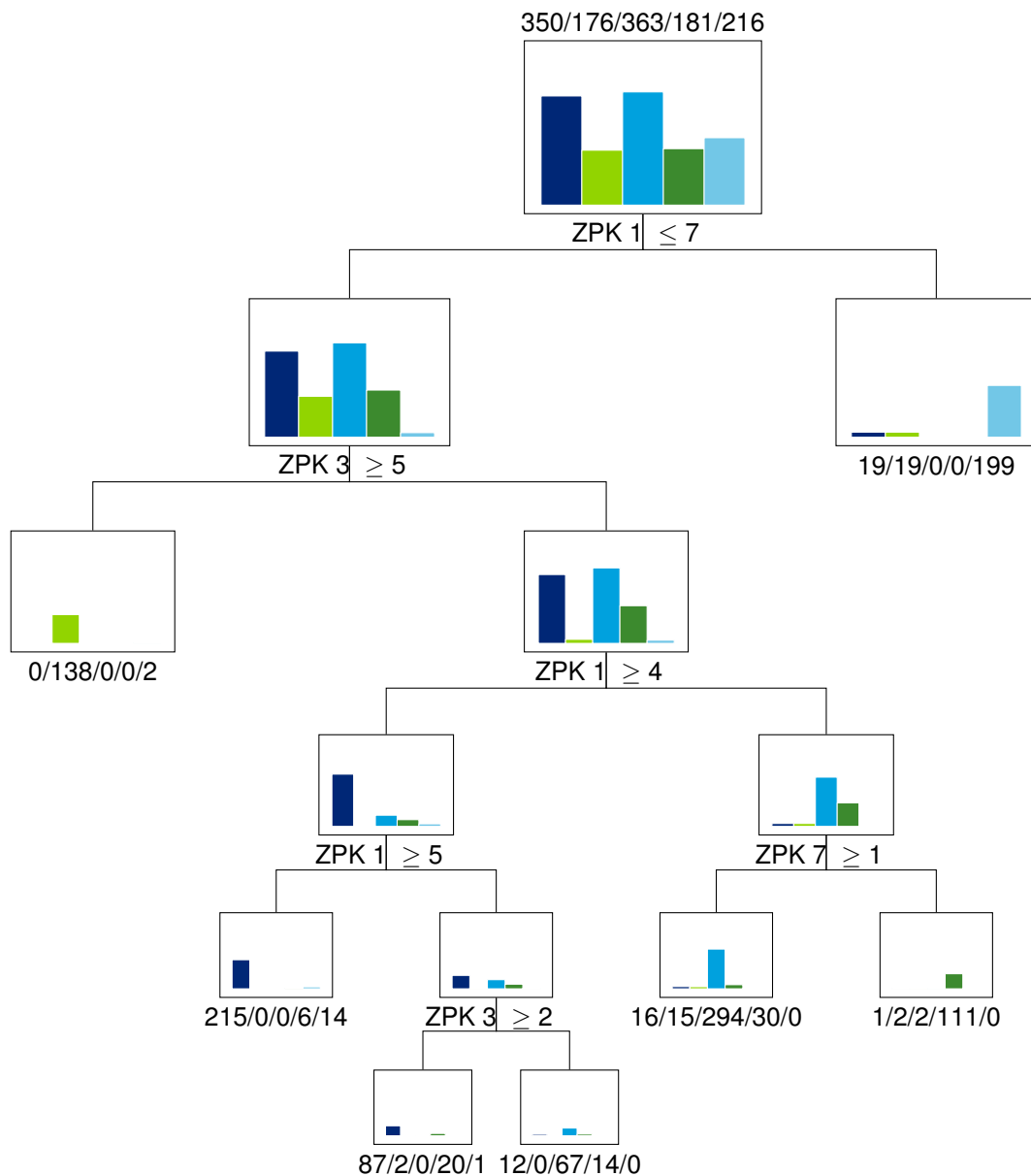


Figure 5.17: Classification tree based on ZPK 1, 3, 7.

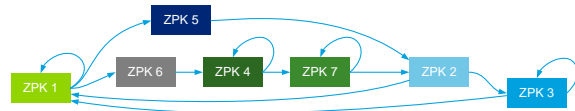
clear description of the main pattern, and the length of the patterns is consistent with the characteristics identified by the classification tree.

5.3.3 Step 3: Results

Combined, the clustering (Figure 5.16c) and classification (Table 5.3 and Figure 5.17) results offer a clear division of carepaths into five groups with specific characteristics as visualized in Appendix F.1, which displays histograms for the most frequent ZPK's (i.e. 1, 3, 7 and 5). The Trace Alignments, showing the main activity patterns and deviations from these patterns, are visualized per cluster in Appendix F.2. From these visualizations we derived the main process patterns as shown in Figure 5.18, which also shows the most insightful process model, describing two different paths from the outpatient

#	Main pattern
1	1135373331
2	11157333311
3	11573333331
4	135333331
5	115333333333333331

(a) Main patterns (all clusters)

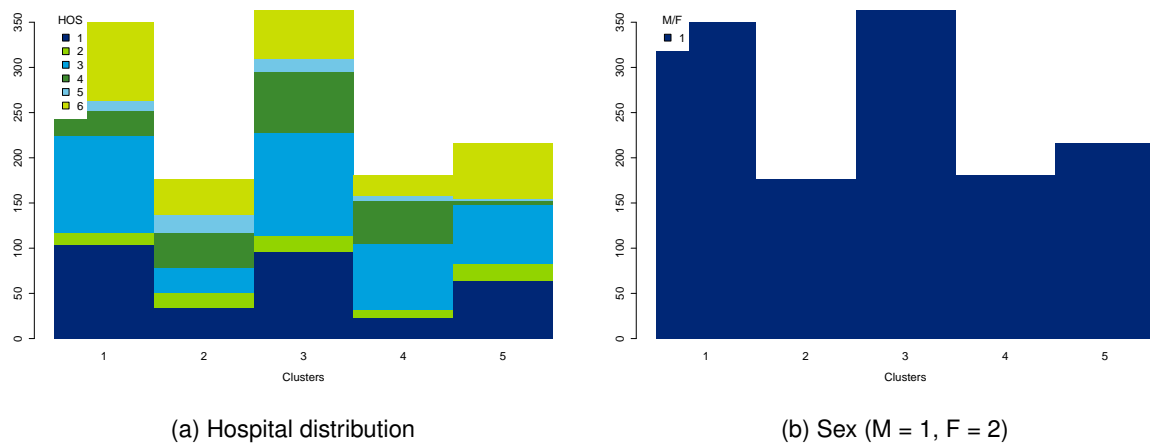


(b) Heuristics Process Model (cluster 5)

Figure 5.18: Example Heuristics Miner process model with main patterns from the Trace Alignment.

department to the actual hospitalization – one with surgery, and one with different types of diagnostic activities.

As previously stated, we have additional data on patients' age, sex⁴ and the type of hospital. Like with the previous case studies, none of the clusters is dominated by a specific type of hospital in Figure 5.19a. An interesting observation, however, is that hospitals 4 and 5 have few patients (about 3 percent) in the exceptional cluster 5. This does not add valuable insight to the type of patients or surgery per type of hospital, as 4 is a top-clinical hospital whilst 5 is a general hospital. Similar to the previous cases, the longest type of carepaths have the oldest patients on average, but again – with just 44 exceptions – the ages range from 40 to 90 for all clusters.



(a) Hospital distribution

(b) Sex (M = 1, F = 2)

Figure 5.19: Neither a specific hospital type nor the patient's sex are dominant in any cluster.

Unlike the other case studies, the extensive abstraction and filtering of activities from our original dataset does show its effect on the relative number of activities. Not only did the total number of activities decrease with over 50 percent, also the variety in number of activities decreased. With the support of experts, we consider this to be a useful side-effect, as we target the logistics process of patient careflow on a daily basis: the resulting dataset results in a higher level of distinction (i.e. less overlap) in number of activities per cluster. The resulting variation in number of activities per cluster is similar to the variations described in the filtered datasets from the previous case studies.

The relative cost per activity is relatively high for cluster 4, although the average total carepath cost

⁴Note that this DBC entails female patients by definition, and male entries are considered to be clerical errors.

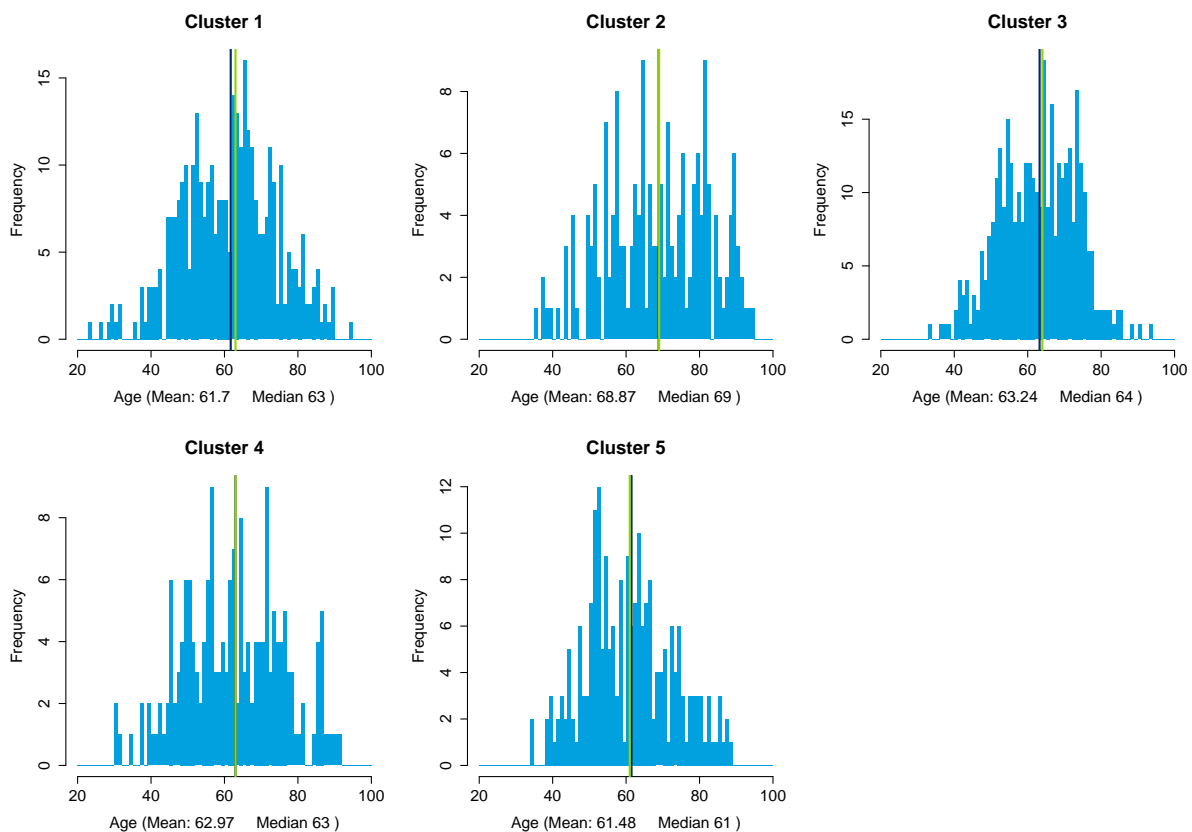


Figure 5.20: Although the mean and median ages differ between clusters, each cluster contains patients of a broad range of ages, which implies age is not a predictive or decisive factor in the selection of a carepath.

is similar to the costs of cluster 3. This is an example of why it is important to take both the number of activities and the ratio of cost per activity into account: in this case the reduced number of images for cluster 4 implies that there are less activities to spread the costs over. One could consider this cluster to have more cost efficient activities, and we would prefer the carepath of cluster 4 over that of cluster 3.

5.4 Summary

In the previous sections we analyzed three different types of patient careflow, and managed to identify a number of distinct carepath types. Using a variety of visualizations and mining techniques, we were able to identify the main process pattern differences, and characterized each cluster based on their main attributes. Also, the total number of activities and costs give insight into which carepath type is preferable on both a logistic and financial level. The type of hospital did not offer support in categorizing the myriad of paths, but the age of patients does provide the insight that older people generally do require more nursingdays to recover. The latter insight cannot be used as a general rule of thumb however, as the variety of patients' age has a similar distribution for all clusters, ranging between the ages of 40 and 90.

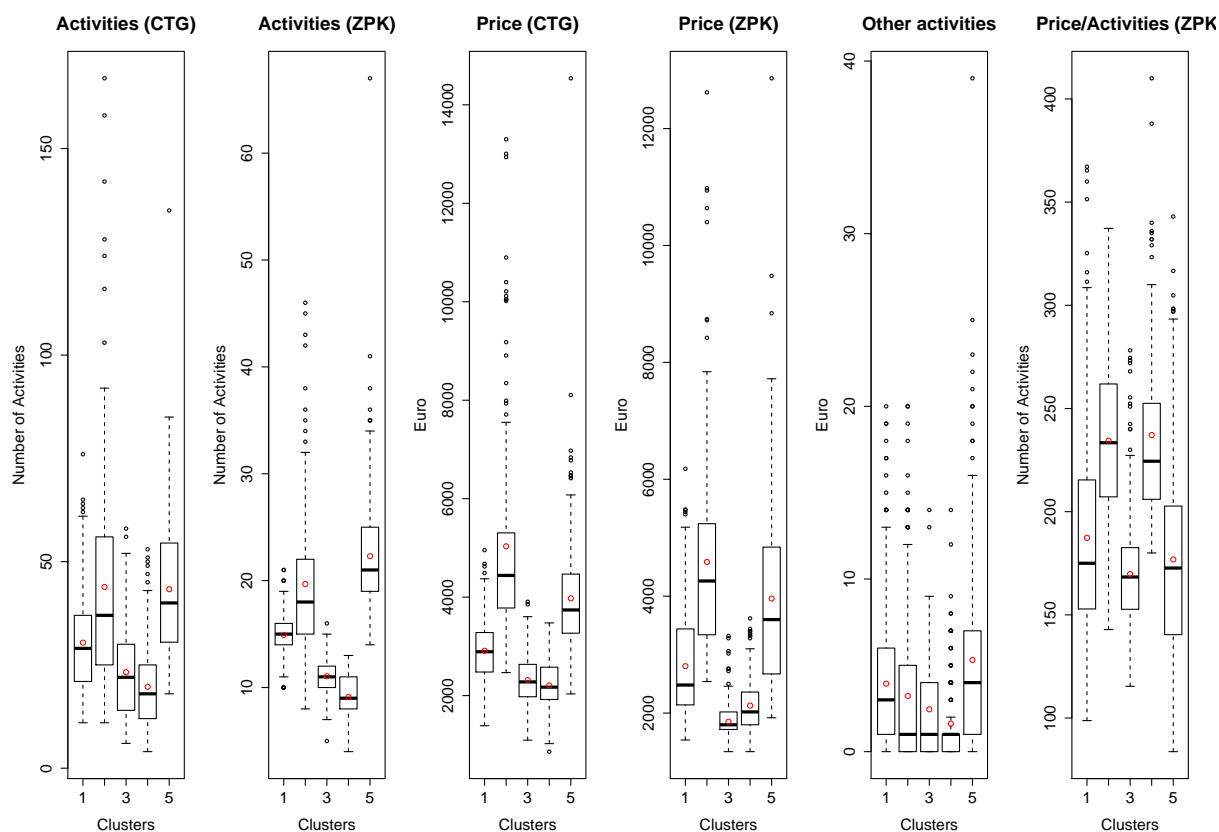


Figure 5.21: Number of activities, carepath costs (real and estimated) and their ratio.

A selection of insights gained from the previous sections is summarized in Table 5.4. Together with the visualizations used to gain these characteristics, this table provides a valuable overview of the different types of patient careflow described by a single care product.



Table 5.4: Summarized tables. Each of the table provides a short overview of the case studies presented in the previous sections. For each of the clusters, these tables describe: the size of the clusters in number of carepaths, the label our experts gave to the clusters describing the general patient careflow, the main activity pattern supported by at least half of the paths within a cluster as identified by the Trace Alignment plugin, the average total number of ZPK activities, the average number of ZPK 1 and 3 activities, the average costs of a carepath, the spread in size and/or cost of paths within a cluster (++ represents a low spread, -- a high spread) and the average age of patients within a cluster.

(a) Arthrosis (hip) - surgical/clinical with joint prosthesis, 4 clusters

Cluster	Label	Path	ZPK#	ZPK1	ZPK3	Cost	+/-	Age
1 (784)	Short	135373331	10	2.2	5.2	€ 5 100	+	71
2 (126)	Long	1713537333333333333317	30	4.5	19.8	€ 11 603	--	78
3 (800)	Short	171353733317	13	4.7	4.3	€ 5 064	+–	69
4 (400)	Medium	11353733333317	15	8.2	3.3	€ 6 432	+–	76

(b) Arthrosis (knee) - surgical/clinical with joint prosthesis, 5 clusters

Cluster	Label	Path	ZPK#	ZPK1	ZPK3	Cost	+/-	Age
1 (484)	Short	1135373331	11	3.3	4.7	€ 5 652	+	69
2 (398)	Short	11157333311	14	5.1	4.9	€ 5 924	+–	70
3 (296)	Medium	115733333331	16	4.0	8.3	€ 7 345	+–	74
4 (439)	Medium	135333331	11	2.2	5.9	€ 6 052	+	71
5 (64)	Long	11533333333333333331	32	4.8	22.1	€ 8 496	--	75

(c) Malignant breast neoplasm - surgical/clinical, 5 clusters

Cluster	Label	Path	ZPK#	ZPK1	ZPK3	Cost	+/-	Age
1 (350)	Medium	7411533111	15	6.0	3.2	€ 2 911	+–	62
2 (176)	Long	411453333331	20	5.3	8.8	€ 5 032	–	69
3 (363)	Short	741145331	11	3.8	2.2	€ 2 315	+	63
4 (181)	Simple	115331	9	3.6	8.8	€ 2 207	+	63
5 (216)	Outpatient	41411453311111	22	9.9	4.7	€ 3 976	--	61



Chapter 6

Discussion and Conclusions

In the previous chapters, we tried to develop a methodology to gain insight into patient careflow using a variety of data analysis and visualization techniques. The first phase was getting a thorough understanding of the healthcare domain, and we proposed our own definition of patient careflow. In the second phase, we collected a number of standardized datasets and identified the useful attributes these sets contained. Based on the prepared datasets from the third phase, we started our exploratory modeling in phase four and developed our final methodology. This methodology was tested on three case studies in the fifth, final phase. The next section provides a recap on these five phases. We describe why we did what, what went wrong and what is next.

6.1 Business Understanding

It is not an easy task to define useful though specific business objectives in a broad, dynamic and complex domain such as healthcare. Although this area of research is increasingly popular due to both an increasing pressure of governments to reduce costs and an increase in demand for care, it is still relatively new. The same goes for the area of process mining; although it is rapidly gaining popularity, few organizations have successfully implemented this type of analysis.

In order to define patient careflow on a suitable level for this study, we proposed a simplified view of the healthcare logistic process: medical activities are described on a high level by their ZPK, and only a limited set of eight different classes is included in the analysis. This view of the logistics process supports the analysis of care products from both the old DBC and the new DOT reimbursement systems, as the process is described on a high level. Especially DOT is designed to supersede individual specialisms; its products may describe medically dissimilar care profiles that are similar only in the logistic process.

In the end, we think this abstract view of the healthcare process allows us to perform a valid statistical analysis without extensive medical knowledge – both on the details of individual activities and patient's specific requirements. Note however, that although the visualizations provided in this methodology are considered statistically sound, it is up to the actual medical specialists to change and improve their decision making during patient treatment. Our analysis is restricted to providing insights, and is not designed to automate medical decision making.

The data collection, auditing and preparation described in the next two phases, Data Understanding

and Data Preparation, are closely related and consisted of numerous iterations. Therefore, we dedicate the next section to both phases simultaneously.

6.2 Data Understanding and Preparation

These phases required a significant amount of domain knowledge in order to gain a thorough understanding of the available fields and data types. One of the main advantages of this study is that it uses a number of standardized datasets, which means that for every implementation and analysis the data collection and preparation phases are similar. After the data is prepared once, new data – whether it is from a new hospital or an addition to an existing set – can be loaded directly into the main database and used for numerous analyses.

For the purpose of this study we used a sample of the available data extracted directly from a number of HISs, structured similarly to the official DIS. However, as these sample sets did not undergo auditing by the DIS, the data quality could not be guaranteed without validation. We do note that DIS only considers care profiles that are consistent with a specific set of rules regarding the type and number of activities they entail; the ordering of activity sequences is not an element validated by this system. A thorough evaluation of the data described and audited by DIS is described in [64].

The Data Understanding and Preparation phases required a large amount of time – about two-thirds of the time spent on this study – partially due to a lack of experience in data modeling and auditing. No research has been done on the selection of database environments, as this was not part of our scope. Instead, we used a readily available Microsoft SQL Server 2008 R2 environment. With the support of technical experts we were able to automate the data preparation process. This helped to gain insight and control over the entire process, and proved to be a worthy investment of time due to the iterative nature of CRISP-DM. We were also able to identify inefficient processes in SQL Server: carepath representations required hours to generate. Instead, we decided to export the event log directly and generate both the vector and string representations in R, which required seconds to generate.¹ This also provided us with more flexibility regarding the filtering of activity classes.

The auditing performed for this study assures that the remaining dataset described the abstract view of the logistic process with the same level of granularity for all treatments and activities, and removed both type problems and quality problems (see Section 2.2.3). Despite this thorough and time-consuming auditing phase, the case studies in Chapter 5 pointed out some issues remained undiscovered until the actual modeling phase. This iterative property of our methodology is typical for data-based analysis, and is also described by CRISP-DM. The fact that we discovered the flaws in the data using a structured methodology and insightful visualizations help prove the completeness of our methodology.

For example, the amount of activities caused confusion in the initially collected data: for some activity types, this field indicated the number of activities performed, whilst others described volumes of medicine. During the auditing stage, we identified one hospital that booked nursingdays on a single day. As a response, experts suggested hospitals did this on purpose, as it would require fewer actions to register the contents of a carepath on the day of discharge. It was not until the final modeling stage that we realized registrations were booked on the first day of hospitalization instead of the last day, and

¹ Apart from the fact that R is better equipped to perform these type of calculations, it only performs these calculations on data from one DBC for the filtered set of activities, instead of the entire database.



we had to return to the data understanding and preparation phase to correct the sample data. With this example, we conclude the Data Preparation phase and continue to the modeling phase.

6.3 Modeling

Prior to the modeling stage, we reviewed the literature for algorithms and techniques used in previous research. Based on this literature we realized that despite the potential of process mining techniques, the dynamic nature of healthcare processes has proven to be too difficult to be tackled by a single technique. Therefore, we decided to split the analysis of these processes into different steps: first, carepaths are clustered and classified based on their (high level) activity content using traditional data mining techniques. Secondly, the activity sequences for individual clusters are analyzed using relatively new process mining techniques.

6.3.1 Data Mining

The clustering and classification performed in this study are complementary. The first tries to identify groups with similar carepaths based on the individual activity frequencies and labels each path with a cluster number. The latter technique takes these cluster labels, and tries to identify the main characteristics for each cluster by building a decision tree. This section recaps some of the decisions made during the development of the model and evaluates the results.

6.3.1.1 Clustering

Traditional clustering techniques come roughly in two flavors: hierarchical and partitional. One of the advantages of hierarchical clustering is that the entire clustering hierarchy is built, before the user has to decide on the best number of clusters k . This immediately describes an important disadvantage: the hierarchy is built on choices made in previous layers of the hierarchy, whilst these choices may not be optimal for this specific value of k .

Alternatively, we need to define the number of clusters before applying the partitional clustering algorithm. Trying a range of clustering values implies running the algorithm for each individual value of k . The advantage is that the clustering result is a local optimum for each k . A good example for this feature is given in Figure 5.16, where the clustering for five clusters identified a care profile the other clustering results ($k = 4$ and $k = 6$) did not. Since hierarchical clustering combines two entire clusters in each layer, this technique would never have identified this profile.

Although K-means is one of the most popular techniques for partitional clustering, we decided to use the Partitioning Around Medoids (*K-Medoids* or `pam`) algorithm as implemented in R. This algorithm takes an existing data point to represent the average value of a cluster, which implies carepaths are clustered based on their distance to an existing path, instead of a fictional value. `pam` is able to use any type of (dis)similarity measure between carepaths, which allowed us to explore a number of metrics:

Compression Clustering The first distance metrics are based on the Kolmogorov complexity as explained in Section 3.1.2. The theory is simple: compression algorithms work better for carepaths with recurring subsequences, therefore a larger decrease in size of compressed string representations of combined carepaths, suggests these paths are similar. Although the clustering results

show this type of distance metrics turned out to be able to identify carepaths with similar subsequences, they were proven unable to identify paths with similar activity frequencies. Especially the longer activity sequences with a limited number of different possible activities performed relatively well with compression, which resulted in a spread of the exceptional cases over the different clusters.

For the purpose of this research, the number of activities represented by a cluster is more important than common subsequences, because the number of nursingdays says more about the total costs of a carepath than a subsequence of diagnostic activities. Therefore we chose not to include this type of analysis in our final methodology. However, it would still be interesting to see whether common subsequences and the total number of activities could somehow be combined in comparing individual carepaths.

Vector Clustering The next set of distance measures is based on vector representations of carepaths.

Both the literature on the development of the original DBC system, and the clustering results as described in Section 4.4.1.1 indicated the high quality of the Tanimoto distance. This distance measure is able to identify a large set of exceptional cases, and known to be able to cope with matching 0-values (i.e. it does not count missing activities from two carepaths as a similarity).

The Euclidean distance shows similar performance, although its clustering appears to show more extremes in clusters: the group of exceptional cases is smaller than for the clustering with Tanimoto, resulting in a higher spread of hospitalization duration for the remaining clusters. Besides, the Euclidean distance has an obvious disadvantage: the distance between two vectors can be small even without sharing any common variable. However, as every DBC of the three case studies analyzed in this Chapter 5 contains at least a number of outpatient department visits and nursing days, the probability of this type of paths existing is negligible for this set of activity classes. When future research tries to target activities on a higher level of detail, the Cosine distance becomes more interesting as it is able to cope with many matching 0-values. The currently selected set of activity classes is too small to have a risk of many matching 0-values, and the clustering resulting from this distance measure turned out to be incapable of identifying similar care profiles.

6.3.1.2 Classification

By applying a classification algorithm to the clustering obtained, we automate the process of identifying main characteristics for each cluster. The Classification And Regression Tree algorithm as implemented in R first builds the entire tree, up to the point where an extra split does not improve the classification error. We can then select the preferred number of splits, based on the `error`-value; the lower the value, the better the carepaths are classified. For this study, however, it is more important to have a small and simple tree describing each cluster with only a few characteristics, rather than having a large tree with multiple characterizations for each cluster. With our case studies, we managed to correctly classify about 90 percent of the individual carepaths, which experts considered a reasonable percentage taking into account the dynamic and complex nature of healthcare processes.

The size and classification error of the tree also give some indication over the clustering quality: if a small tree is able to easily separate one cluster from the other, this implies the clusters are distinct in a small number of activity frequencies. However, as the distances between carepaths are calculated over



a number of activities, it is not realistic to conclude the clustering quality is bad when a larger tree is required to describe the different types of care profiles.

6.3.1.3 Software package

As software analysis was not part of the scope of this project, we selected R for the development of our models mainly because of our (successful) previous experiences. Both the internal documentation and the online community proved to be valuable resources during the modeling phase, and we were able to develop an almost fully automated tool that supports and executes the first steps of our methodology. We were also able to (automatically) create insightful visualizations in a single environment, even though R is not a Business Intelligence or Visual Analytics tool such as Qlikview and MagnaView. Another advantage is that the generic properties of the code generated allows us to re-use it for purposes other than patient careflow discovery. In short, the flexibility and extensibility of R proved valuable for this research.

6.3.2 Process Mining

In Section 2.3.2 we mentioned a number of techniques that have been applied to the healthcare domain in previous research. Especially the Heuristics Miner has been proven to perform well for event logs that contain a lot of noise, which is often the case in healthcare processes. Also the Trace Alignment plugin was expected to provide the insight we were looking for.

For the purpose of this study however, we removed the noise from the event log during the Data Understanding and Preparation phases by generalizing activities by their ZPK, resulting in small lists of five to eight activities. The Heuristics Miner derived similar process models for each of the clusters, and provided little insight into the different carepaths clustered together. Only for the third case study, the process model from Figure 5.18b resulted in a useful insight: on a single nursingday, a patient had either surgery or a number of diagnostic and/or therapeutic activities. The remaining models all described a global process, but lacked information on the number of iterations and start or end activity of a carepath.

The Trace Alignment turned out to be a more valuable tool in visualizing and identifying patterns within a cluster of carepaths. The visualization offers a clear overview of the main patterns by aligning the most frequent activities. As each row in the visualization is unique, the number of occurrences of a carepath is indicated by a number, which helps identifying the patterns for the most frequent paths. A significant downside of this plugin is its coloring function: for each analysis random colors are assigned to the activities. It takes a lot of manual input to re-use the same coloring scheme for multiple analyses.

Note that before we were able to run the Trace Alignment, we had to execute the Guide Tree Miner. However, the different clustering functionalities for this plugin did not provide the means to create clusters of the same quality as our R clustering offered, and we did not want to create a larger set of clusters. Therefore, we decided to ignore the functionality of this plugin entirely by selecting a single cluster per clustering, which enabled us to preserve our original clustering results as input for the Trace Alignment plugin.

The previous paragraph offers a good example of some of our struggles in using ProM, due to the lack of experience with the tool. Although the resulting visualizations can be interpreted by medical experts without specific analytical knowledge, ProM does not offer a user-friendly process mining tooling. Instead, it is a technical though extensible tool containing numerous experimental techniques, and with

the support of experienced users and developers it turned out to be a useful tool for the purpose of this study.

6.4 Analysis results

The case studies in Chapter 5 provide a complete overview of the different visualizations required to perform patient careflow Discovery. For each case study, we started with plotting the entire event log grouped per hospital, which enabled us to audit the data to remove both bias towards hospitals and infeasible carepaths. After we performed clustering for a range of clusters, the different results are plotted in an activity barchart visualizing the different patterns described by each cluster. For each of the clusters, we can evaluate individual activity patterns in activity histograms which help to evaluate the different results. After selecting the optimal number of clusters, a classification tree is built identifying the main characteristics of each cluster. By visualizing the number of cluster elements in each node of the tree, the tree gives a clear overview of the quality of the splits and indicates whether we have identified the main characteristics describing each group. Together with the results from the Trace Alignment, the summaries in Table 5.4 provide a clear and useful insight into patient careflow. These results help improving patient careflow by increasing standardization, selecting the care profile with the lowest cost for the highest quality of care and in the future for predictive modeling, allowing the development of robust and optimal operating schedules.

6.5 Future Work

For the purpose of this study, we selected a single care product (or DBC) for each case study. However, since we are looking at the logistic process of patient careflow, medical differences between different DBC's and specialisms are not considered, which should enable us to combine patient data from different DBC's and even different specialisms to find homogeneity within the logistic patterns. To allow this type of analysis in future work, one has to take care when selecting the DBC products, as the different products should have similar activity profiles regarding the activity classes. Another dangerous assumption is that registrations are done in a uniform way throughout numerous hospitals and specialisms, as this study has already pointed out differences in the way hospitals register their diagnostic activities. With the introduction of the new DOT system, we hope the quality of registrations will increase such that the data supports the analysis of a combination of care products.

We limited the scope of this research to the analysis of the logistic process of patient careflow, which enabled us to perform a statistically sound analysis. This study offers a good starting point for the analysis of the medical processes in future work, where thorough understanding of the analytical process is combined with extensive medical domain knowledge. To enable this, one should consider collecting more medical data on patients such as whether the patient is a diabetic², or a description of his or her medical history. The latter is already made possible in the DIS, as it stores patients by a unique identifier based on their social security number. This means patients can be tracked to every hospital they visited in recent years, and might provide insight on e.g. exceptional cases that are transferred from a different hospital because they need special care.

²The glucose levels for diabetes patients are checked every day, which would explain a higher number of labtests.



In order to take into account the type of hospital in future research, collecting data from a larger number of hospitals would also allow the analysis of whether top-clinical hospitals describe different types of carepath than general hospitals. The dataset used in this study contained only six hospitals, which proved to be too small to find any type of insight.

Finally, future research can enable the incorporation and implementation of this type of analysis tooling into modern WfMS, similarly to the study in [43]. It would be interesting to see how these analyses could be leveraged by incorporating them into day-to-day decision making [43].

6.6 Conclusion

Despite the extensive amount of research performed on the application of data-based methods in the healthcare domain, no single technique or methodology has been found able to cope with its extensive dynamic properties. We developed a methodology that uses a combination of data mining and process mining techniques, and is able to provide valuable insight into patient careflow that supports an increase in quality of care with a decrease in costs. This answers our first research question, “*Can data- and process mining techniques be applied to gain insight in patient careflow?*”, with a solid “yes”.

In the previous chapters we have proven that by using standardized data, this methodology is able to identify clusters with similar care profiles, classify the important characteristics of these profiles, and discover process properties with the following techniques:

Clustering Based on the Tanimoto Distance between vector representations of individual carepaths, the Partitioning Around Medoids (PAM) algorithm identifies distinct clusters describing specific types of care profiles.

Classification Using the Classification And Regression Tree (CART) algorithm, a decision tree identifies the main characteristics for each cluster.

Trace Alignment This plugin in ProM offers a clear overview of the different paths contained in a single cluster, and identifies both the main process pattern and deviations from these patterns.

In order to answer the second research question, “*Which insights do we require to assess patient careflow?*”, we targeted the organizational process of the healthcare domain. This allowed us to focus on an abstract view of the *logistic* process, where individual activities were generalized into activity classes (ZPK’s). This approach automatically dealt with the fact that limited patient data as well as medical knowledge was available for this study. Based on a number of insightful visualizations, the clustering of these generalized carepaths was best achieved using the smallest number of clusters, describing the highest number of distinct care profiles.

The answer to our third research question, “*How can we compare, evaluate and advise different carepaths?*”, lies with the visualizations and subjective measures mentioned in previous sections. The decision on the final number of clusters is mainly based on a stacked barchart, where each bar shows the total activity frequency of an individual path, and colors indicate the individual activities. A good clustering shows little variation for the activities within a cluster, whilst each cluster shows distinct patterns for their average activity frequencies (example are given in Figures 4.6b, 5.8c and 5.16c). Individual activity histograms per cluster visualize these frequencies in more detail (examples are given in Appendices D.1, E.1 and F.1). After classification, which enabled us to identify the main characteristics of each cluster, we performed the Trace Alignment. The data transformation from R into ProM required extensive manual interaction, but the results were valuable: matching activities were aligned amongst individual paths and provided a clear overview of the major patient careflow per cluster. Deviations were also clearly visible, as these were represented by unaligned activities.



Together with a number of statistics on patient's age, number of activities and carepath costs, the mining results and different visualizations offered valuable insight in the different types of patient care-flow for individual DBC's. Based on these insights, a number of process improvements are possible:

1. Medical specialists can be stimulated to treat patients according to the carepath with the highest quality of care and lowest costs.
2. Standardization can be improved for the different types of carepaths within one care product or DBC.
3. These results can be used as input for predictive modeling, in order to provide an optimal and robust operating schedule.

The next step for this study is the last phase of CRISP-DM: Deployment. By including medical specialists from a live hospital environment, the insights gained from this methodology can be used to realize the improvements described above.



Bibliography

- [1] 2009 by the Process Mining Group. Process Mining research tools application. <http://www.processmining.org>, 2009.
- [2] Wil M. P. van der Aalst. Business alignment: using process mining as a tool for delta analysis and conformance testing. *Requirements Engineering*, 10:198–211, 2005.
- [3] Rakesh Agrawal, Dimitrios Gunopulos, and Frank Leymann. Mining process models from workflow logs. In *EDBT*, pages 469–483, 1998.
- [4] J. Alapont, A. Bella-Sanjuán, C. Ferri, J. Hernández-Orallo, J. D. Llopis-Llopis, and M. J. Ramírez-Quintana. Specialised Tools for Automating Data Mining for Hospital Management. In *First East European Conference on Health Care Modelling and Computation*, pages 7–19, 2005.
- [5] Álvaro Rebugue and Diogo R. Ferreira. Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems*, 37:99–116, 2012.
- [6] A. Barbieri, K. Vanhaecht, P. Van Herck, W. Sermeus, F. Faggiano, S. Marchisio, and M. Panella. Effects of clinical pathways in the joint replacement: a meta-analysis. *BMC Medicine*, 7(1):32, 2009.
- [7] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies and Techniques*. Data-Centric Systems and Applications. Springer, 2006.
- [8] Carol Beaver, Yuejen Zhao, Stewart McDermid, and Don Hindle. Casemix-based funding of Northern Territory public hospitals: adjusting for severity and socio-economic variations. *Health Economics*, 7(1):53–61, 1998.
- [9] R. P. Jagadeesh Chandra Bose and Wil M. P. van der Aalst. Trace alignment in process mining: Opportunities for process diagnostics. In *BPM*, volume 6336 of *Lecture Notes in Computer Science*, pages 227–242, 2010.
- [10] R. P. Jagadeesh Chandra Bose and Wil M. P. van der Aalst. Process Diagnostics Using Trace Alignment: Opportunities, Issues, and Challenges. *Information Systems*, 37(2):117–141, 2012.
- [11] J.C.A.M. Buijs. Mapping Data Sources to XES in a Generic Way. Master's thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2010.
- [12] M. Burrows and D. J. Wheeler. A block-sorting lossless data compression algorithm. Technical report, 1994.

- [13] Bilson J. L. Campana and Eamonn J. Keogh. A compression based distance measure for texture. In *SDM*, pages 850–861, 2010.
- [14] Bilson J. L. Campana and Eamonn J. Keogh. A compression-based distance measure for texture. *Statistical Analysis and Data Mining*, 3(6):381–398, 2010.
- [15] Andrzej Ceglowski, Leonid Churilov, and Jeff Wassertheil. Knowledge Discovery through Mining Emergency Department Data. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)*.
- [16] Andrzej Ceglowski, Leonid Churilov, and Jeff Wassertheil. Data driven process modelling for a hospital emergency department. In *Computer Supported Activity Coordination*, pages 61–70, 2004.
- [17] Shelton C.P. Human Interface / Human Error, aug 2005.
- [18] DBC-informatiesysteem. Ga sz 6.0 v 4.3: Standaard voor dis gegevensaanlevering dbc door aanbieders ziekenhuiszorg. Technical report, DIS, 2011.
- [19] DBC-Onderhoud. Diagnose Behandeling Combinatie Onderhoud. <http://www.dbconderhoud.nl/>, 2011.
- [20] DIS. Het landelijke DBC-informatiesysteem. <http://www.dbcinformatiesysteem.nl/>, 2011.
- [21] Marlon Dumas, Wil M. P. van der Aalst, and Arthur H. ter Hofstede. *Process Aware Information Systems: Bridging People and Software Through Process Technology*. John Wiley & Sons, Inc., 2005.
- [22] Christos Faloutsos and Vasileios Megalooikonomou. On data mining, compression, and kolmogorov complexity. *Data Min. Knowl. Discov.*, 15(1):3–20, August 2007.
- [23] Usama Fayyad, Gregory Piatesky-shapiro, and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17:37–54, 1996.
- [24] Diogo R. Ferreira. Applied Sequence Clustering Techniques for Process Mining. In Jorge Cardoso and Wil M. P. van der Aalst, editors, *Handbook of Research on Business Process Modeling*, Information Science Reference, pages 492–513. IGI Global, 2009.
- [25] E. van Ginneken, W. Schafer, and M. Kroneman. Managed competition in the netherlands: an example for others? *Eurohealth*, 16(4):23–26, 2010.
- [26] GNU Project. The R Project for Statistical Computing. <http://www.r-project.org/>.
- [27] Christian W. Günther and van der Aalst Wil M. P. Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics. In *BPM*, pages 328–343, 2007.
- [28] S. Gupta. Workflow and Process Mining in Healthcare. Master's thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2007.
- [29] F. Hasaart. *Incentives in the Diagnosis Treatment Combination payment system for specialist medical care*. PhD thesis, Maastricht University, 2011.



- [30] D. Iacobucci. *Methodological and Statistical Concerns of the Experimental Behavioral Researcher*. Journal of Consumer Psychology. Lawrence Erlbaum Associates, Publishers, 2001.
- [31] Nederlandse Zorgautoriteit i.o. (CTG/ZAio en CTZ). Declaraties Beter Controleerbaar. Technical report, August 2006.
- [32] Mark W. Isken and Balaji Rajagopalan. Data Mining to Support Simulation Modeling of Patient Flow in Hospitals. *J. Med. Syst.*, 26:179–197, 2002.
- [33] Rob Karel. Stop Trying To Put A Monetary Value On Data Its The Wrong Path. http://blogs.forrester.com/rob_karel/11-03-29-stop_trying_to_put_a_monetary_value_on_data_its_the_wrong_path, 2011.
- [34] Eamonn Keogh, Stefano Lonardi, Chotirat Ann Ratanamahatana, Li Wei, Sang-Hee Lee, and John Handley. Compression-based data mining of sequential data. *Data Min. Knowl. Discov.*, 14(1):99–129, February 2007.
- [35] Martin Lang, Thomas Bürkle, Susanne Laumann, and Hans-Ulrich Prokosch. Process Mining for Clinical Workflows: Challenges and Current Limitations. In Stig Kjr Andersen, Gunnar O. Klein, Stefan Schulz, and Jos Aarts, editors, *MIE*, volume 136 of *Studies in Health Technology and Informatics*, pages 229–234. IOS Press, 2008.
- [36] Remco Lassche. Care Pathway Analysis and Redesign: a Methodology. Master's thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2010.
- [37] Richard Lenz and Manfred Reichert. IT Support for Healthcare Processes. In Wil van der Aalst, Boualem Benatallah, Fabio Casati, and Francisco Curbera, editors, *Business Process Management*, volume 3649 of *Lecture Notes in Computer Science*, pages 354–363. Springer Berlin / Heidelberg, 2005.
- [38] Richard Lenz and Manfred Reichert. IT support for healthcare processes - premises, challenges, perspectives. *Data & Knowledge Engineering*, 61(1):39–58, 2007.
- [39] J. Li, R. P. Jagadeesh Chandra Bose, and Wil M. P. van der Aalst. Mining context-dependent and interactive business process maps using execution patterns. In Michael zur Muehlen and Jianwen Su, editors, *BPM 2010 International Workshops and Education Track, Revised Selected Papers*, volume 66 of *Lecture Notes in Business Information Processing*, pages 109–121, 2010.
- [40] Jing-Song Li, Hai-Yan Yu, and Xiao-Guang Zhang. Data Mining in Hospital Information System. In Prof. Kimito Funatsu, editor, *New Fundamental Technologies in Data Mining*, 2011.
- [41] W. Lidwell, K. Holden, and J. Butler. *Universal Principles of Design: 125 Ways to Enhance Usability, Influence Perception, Increase Appeal, Make Better Design Decisions, and Teach Through Design*. Rockport Publishers, 2010.
- [42] Alan Lipkus. A proof of the triangle inequality for the tanimoto distance. *Journal of Mathematical Chemistry*, 26(1):263–265, Oct. 1999.
- [43] R.S. Mans. *Workflow Support for the Healthchare Domain*. PhD thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2011.

- [44] R.S. Mans, Wil M. P. Aalst, N.C. Russell, and P.J.M. Bakker. Flexibility schemes for workflow management systems. In Danilo Ardagna, Massimo Mecella, Jian Yang, Wil Aalst, John Mylopoulos, Michael Rosemann, Michael J. Shaw, and Clemens Szyperski, editors, *Business Process Management Workshops*, volume 17 of *Lecture Notes in Business Information Processing*, pages 361–372. Springer Berlin Heidelberg, 2009.
- [45] R.S. Mans, Helen Schonenberg, G. Leonardi, S. Panzarasa, Anna Cavallini, S. Quaglini, and Wil M. P. van der Aalst. Process Mining Techniques: an Application to Stroke Care. *MIE 2008*, 136:573 – 578, 2008.
- [46] R.S. Mans, M.H. Schonenberg, M. Song, Wil M. P. van der Aalst, and P.J.M. Bakker. Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital. In Ana Fred, Joaquim Filipe, and Hugo Gamboa, editors, *Biomedical Engineering Systems AND Technologies*, volume 25 of *Communications in Computer and Information Science*, pages 425–438. Springer Berlin Heidelberg, 2009.
- [47] R.S. Mans, Wil M. P. van der Aalst, N.C. Russell, A.J. Moleman, P.J. Bakker, and M.W. Jaspers. *Modern Business Process Automation: YAWL and its Support Environment*. Springer, 2009.
- [48] A. K. Alves De Medeiros and A. J. M. M. Weijters. Genetic Process Mining. In *Applications and Theory of Petri Nets*, volume 3536, pages 48–69. Springer-Verlag, 2005.
- [49] Vasileios Megalooikonomou. Kolmogorov incompressibility method in formal proofs - a critical survey, 1997.
- [50] E.M. Mirkes. K-means and K-medoids: applet, 2011.
- [51] Mark Norton. a new approach to systems development: Decisioning. <http://www.idiomsoftware.com/>, 2008.
- [52] IEEE Task Force on Process Mining. Process Mining Manifesto. <http://www.win.tue.nl/ieeetfpm/>, 2011.
- [53] M. Poullymenopoulou, F. Malamateniou, and G. Vassilacopoulos. Specifying Workflow Process Requirements for an Emergency Medical Service. *J. Med. Syst.*, 27:325–335, August 2003.
- [54] K. Putters and Raad voor de Volksgezondheid & Zorg. *Anticiperen op marktwerking: achtergrondstudie uitgebracht door de Raad voor de Volksgezondheid en Zorg bij het advies over marktwerking in de Medisch specialistische zorg*. RvZ, 2003.
- [55] Leonardo Torres Ramos. Healthcare Process Analysis: validation and improvements of a data-based method using process mining and visual analytics. Master's thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2009.
- [56] H.A. Reijers, A.J.M.M. Weijters, B.F. van Dongen, A.K. Alves de Medeiros, M. Song, and H.M.W. Verbeek. Business process mining: An industrial application. *Information Systems*, 32, 2007.
- [57] Hajo A. Reijers and Wil M. P. van der Aalst. The effectiveness of workflow management systems: Predictions and lessons learned. *International Journal of Information Management*, 25(5):458–472, 2005.



- [58] Fu ren Lin and Shien chao Chou. Mining time dependency patterns in clinical pathways. *International Journal of Medical Informatics*, pages 11–25, 2001.
- [59] Patrick Riemers. Process improvement in Healthcare: A data-based method using a combination of process mining and visual analytics. Master's thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2009.
- [60] D. Salomon. *Data Compression: The Complete Reference*. Number v. 10. Springer, 2007.
- [61] Jan Staal. Using process and data mining techniques to define and improve standardization in a healthcare workflow environment. Master's thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2010.
- [62] Inc. StatSoft. Electronic statistics textbook, 2012.
- [63] Julian Steward. bzip2: Home. <http://www.bzip.org>, 1996.
- [64] P.C.E. van Stijn. Data Quality and Usefulness of the Dutch DBC Information System. Technical report, Utrecht, The Netherlands, 2012.
- [65] Kenji Takegami, Yonei Kawaguchi, Hiroshi Nakayama, Yoshiro Kubota, and Hirokazu Nagawa.
- [66] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison Wesley, us ed edition, 2005.
- [67] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition*. Third edition.
- [68] Sjoerd van der Spoel. Outcome and variable prediction for discrete processes. Master's thesis, University of Twente, Enschede, The Netherlands, 2012.
- [69] B.F. van Dongen, A. K. A. de Medeiros, H. M. W. Verbeek, A.J.M.M. Weijters, and Wil M. P. van der Aalst. The ProM Framework: A New Era in Process Mining Tool Support. In Gianfranco Ciardo and Philippe Darondeau, editors, *Applications and Theory of Petri Nets 2005*, volume 3536 of *Lecture Notes in Computer Science*, chapter 25, pages 444–454. Springer Berlin / Heidelberg, 2005.
- [70] Jan Vissers. A logistics approach for hospital process improvements. In Randolph W. Hall, editor, *Patient Flow: Reducing Delay in Healthcare Delivery*, volume 91 of *International Series in Operations Research & Management Science*, pages 393–427. Springer US, 2006.
- [71] K.P.A. van Wanrooij. Operating Theatre Scheduling – Improving bed utilization with an optimized Master schedule. Technical report, Utrecht, The Netherlands, 2011.
- [72] A. J. M. M. Weijters and A. K. Alves De Medeiros. Process Mining with the HeuristicsMiner Algorithm. Technical report, 2006.
- [73] Lijie Wen, Wil M. P. van der Aalst, Jianmin Wang, and Jiaguang Sun. Mining process models with non-free-choice constructs. *Data Min. Knowl. Discov.*, 15:145–180, 2007.
- [74] Machiel Westerdijk, Joost Zuurbier, Martijn Ludwig, and Sarah Prins. Defining care products to finance health care in the Netherlands. *The European Journal of Health Economics*, 2011.

- [75] Rudiger Wirth and Jochen Hipp. CRISP-DM: Towards a Standard Process Model for Data Mining. In *Proceedings of the Fourth international Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pages 22–39, Manchester, UK, 2000.
- [76] I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools And Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufman, 2005.
- [77] Workflow Management Coalition. Terminology and Glossary Document Number WFMC-TC-1011 - Issue 3.0. http://www.wfmc.org/standards/docs/TC-1011_term_glossary_v3.pdf, 1999. definition appears on page 8.
- [78] W. Zhou. Acquiring objective process information for healthcare process management with the CRISP-DM framework. Master's thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2009.



Glossary

care product A single care product describes a set of treatment activities for a specific DBC.

care profile (In Dutch: *zorgprofiel*) See care product.

carepath The specific process or activity sequence a patient goes through during treatment.

clinical pathway See patient careflow.

CRISP-DM (Cross Industry Standard Process for Data Mining) A commonly used technique for DM projects in business applications.

CTG-code Each code represents a specific treatment activity and its fixed prices. The *Nederlandse Zorgautoriteit (NZa)* oversees the healthcare tariff agreements between care providers and funders for each of these codes.

data mining The process of discovering new patterns from large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. Also: *Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases.*

DBC (Diagnosis Treatment Combination – in Dutch: *Diagnose Behandeling Combinatie*) The Dutch refund system for hospital care products. Each code describes a care product. Similar to the Diagnosis-related group (DRG) system.

DIS (DBC Information System) National Information System that records all DBC information – both declared and performed by Healthcare providers (<http://www.dbcinformatiesysteem.nl> – Dutch).

distance function A function that defines the distance between point x and y .

DOT (DBC Towards Transparency – in Dutch: *DBC Op weg naar Transparantie*) An improved refund system, based on the former DBC system, trying to provide a more transparent and clear refund system.

DRG (Diagnosis-Related Group) A classification system for hospital cases, defining homogenous groups of database-stored cases according to medical costs. The partitioning is based on medical and surgical treatments.

event log A collection of recorded events.

grouper A grouper is an algorithm that combines all activities and information on a patient into a billable care product [19].

healthcare processes These general kinds of processes can be subdivided into *medical* and *organizational* processes as shown in Figure 2.1.

HIS (Hospital Information Systems – in Dutch: *ZIS*) A digital Information System for Hospitals, containing patient data (e.g. personal details, insurance provider and general practitioner), logistic process descriptions (e.g. who treated the patient, which activities) and details used for treatment reimbursement.

Kolmogorov complexity The shortest description of a string in a fixed description language – $K(x)$.

medical treatment processes Describes the diagnostic and therapeutic procedures to be carried out for a particular patient, i.e. the diagnostic-therapeutic cycle.

metric A (distance) function that satisfies the properties for *positivity*, *symmetry* and *triangle inequality*.

organizational processes Organizational business processes of healthcare organizations, required to coordinate interoperating healthcare professionals, disciplines and departments.

overdeclaration When a hospital or medical specialist registers an inappropriate number of DBCs for one patient, in order to increase the total refund.

patient careflow (Also: clinical pathway) A group of similar care episodes: when patients undergo similar activities in their path from intake until discharge.

process mining A collection of techniques that allow the extraction of information from event logs. A family of a-posteriori analysis techniques.

upcoding When a higher-paying service (DBC) is chosen, even when this is not medically necessary, this is referred to as upcoding. It is a fraudulent practice where providers try to cheat the system to increase their reimbursement without performing the suggested services and activities.

WfMS (Workflow Management System) A system that provides an environment to automate and assist in the management of tasks and the flow of workitems from one task to another. These systems require a process model and their main function is to ensure that all the activities are performed in the right order and by the right resource [57].

ZPK (activity class) Each individual activity (CTG-code) can be described by a higher level class. A list of different classes is given in Appendix G.



Appendix A

Database structure

At the data collection stage for this project, the DIS system was unavailable for third-parties. Therefore, we use datasets readily available describing patient data in a similar fashion. These datasets however do require extra preparation to become DIS-compliant.

Unlike for the DIS system, the hospitals were unable to guarantee the quality for these sets, they were a quick data dump to use for a quick analysis. We came up against quite some cleaning necessities, e.g.: one hospital did not have a single opening clinic activity, only repeating and repeating-open-new-dbc clinics. Some activities were doubly registered and multiple ID's were missing. This section describes the way we build our final analysis sample set in different stages in data preparation and cleaning.

A.1 DIS_Import

Tablename	description	Unique
*_patient	Hospitalized patients	no
*_zorgtraject	Carepath-ID's for each hospital	yes
*_subtraject	Subcarepath-ID's for each hospital	yes
*_verrichtingen	Activities for each hospital	yes
dbc.AGBCodes	Used for numerical reference to a hospital	yes
dbc.Kostprijzen	Average prices of over 900 unique activities	yes
dbc.Specialismes & _Omschrijving	Description for Specialism-Diagnosis-Treatment combinations	no
dbc.Zorgklasse	Classes for most Activities	no
dbc.ZPKPrijzen	Estimated indicative prices for a handful of Activity Classes	yes

The initial database is a direct import of the complete datasets supplied by a number of hospitals. These sets contain incomplete carepaths and missing or incorrect values, which will be removed or fixed at following stages. Data preparation steps at this stage consist only of:

1. Importing data from source files (using SSIS into MS SQL Server)
2. The creation of numerical PatientID's for certain hospitals. The DIS system uses randomized strings. For clarity and speed purposes, we prefer using numerical ID.

3. The addition of missing patient values with fictional details (birthdate = 01/01/0001, sex = 3).
4. Removal of carepaths without diagnosis for certain hospitals. Since we target specific Diagnosis-Treatment combinations, these carepaths are not going to be in our analysis set.
5. Addition of activities: some hospitals booked the total number of nursingdays on a single day (see Figure A.1). Experts pointed out that this was impossible: a patient for hip surgery rarely stays in the hospital for just one day. In order to use the ordering of activities, we have created individual events for each individual nursingday. Using Trace Alignment we realized that the total number of nursingdays was not recorded on the last day as initially assumed, but rather on the first day of hospitalization.¹

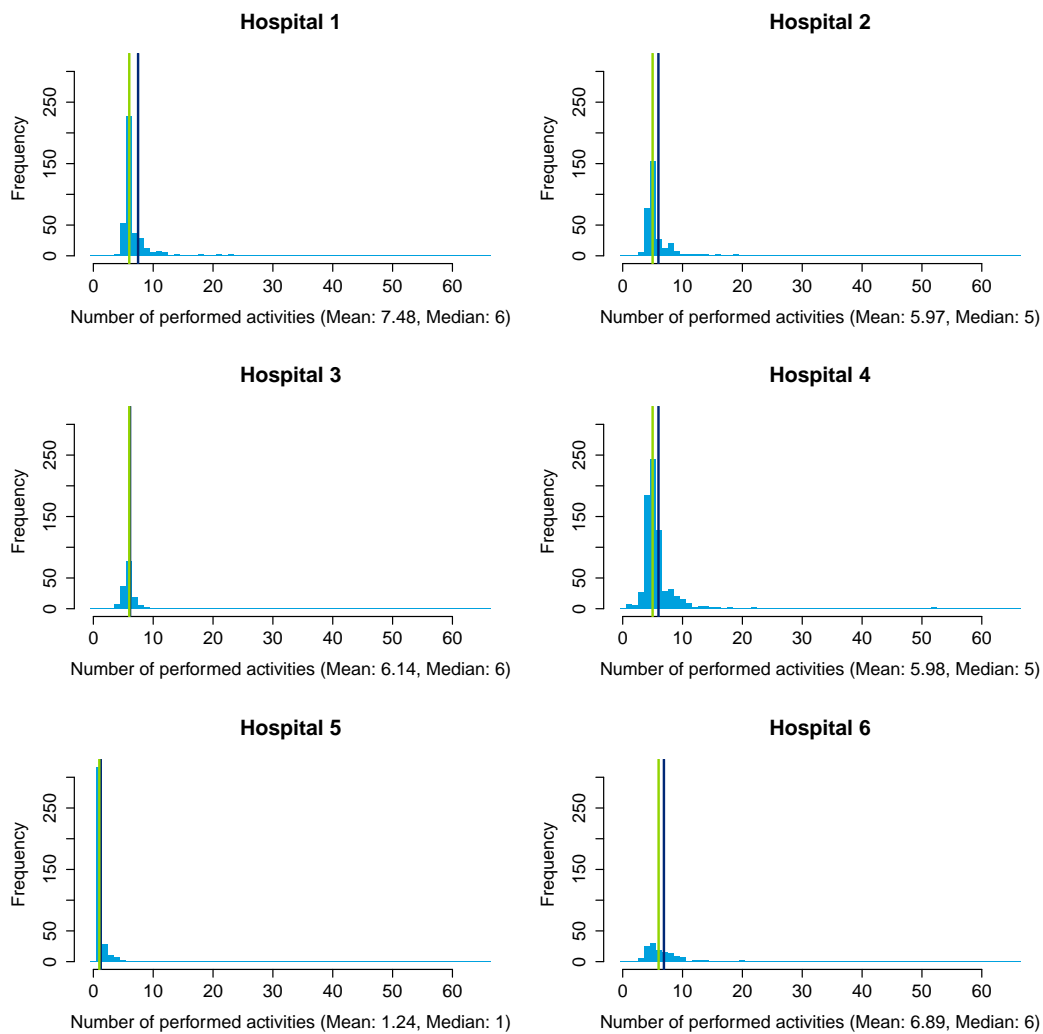


Figure A.1: Overview per hospital: histogram of frequencies for ZPK3. Hospital 5 shows for most patients a single nursingday is recorded. Experts pointed out this is not possible for this DBC.

¹This was a clear example of the iterations in CRISP-DM: in the modeling phase we found that many patients for hip-replacements spent one single day at the hospital.

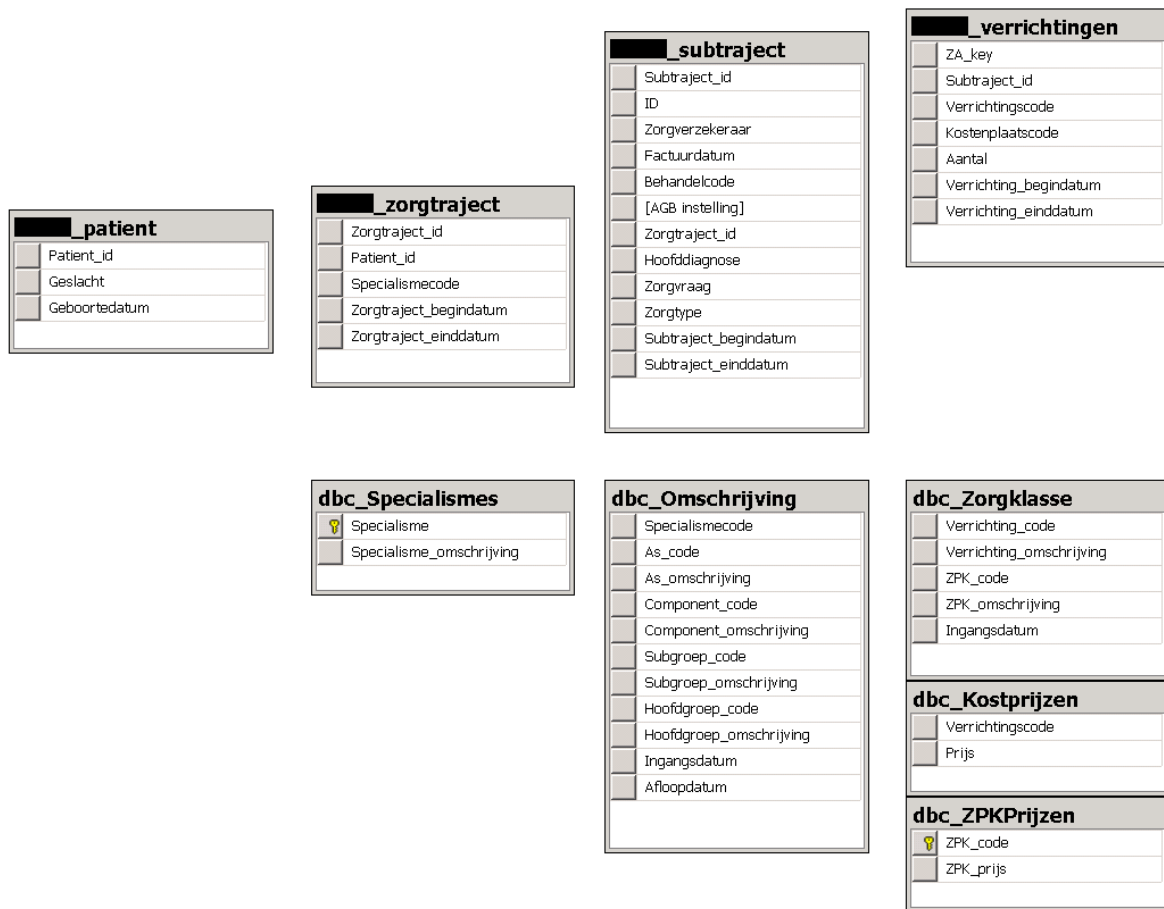


Figure A.2: DIS Import

A.2 DIS_Stage1

Tablename	description	Type
Activities	Inner join on table	Table
PID-Activities	Adding ProcessID's for Careflows and SubCareflows	View

First cleaning stage: an inner join for each hospital on all four tables (Patients, Careflows, SubCareflows, Activities) to remove incomplete flows and activities.

- Some Activity-codes contain trailing Characters. Our Class-codes do not contain this level of Activity detail, for which reason we simply remove the trailing characters. Also note that our remaining set will still contain activities without a link to a Class-code.
- Update Number of Activities. In the DBC-system, each Activity (e.g. a Nursing Day) can only be linked to one DBC, patients on the other hand can be hospitalized for multiple DBC's. We want to prevent losing Activities due to official registration restrictions, therefore we update the NoA.
- Add ProcessID's on a Careflow level (one unique key for each Careflow/Patient combination) and on a DBC or SubPath level (one for each Careflow/Patient/DBC combination) – this is done in a simple View using Partition.

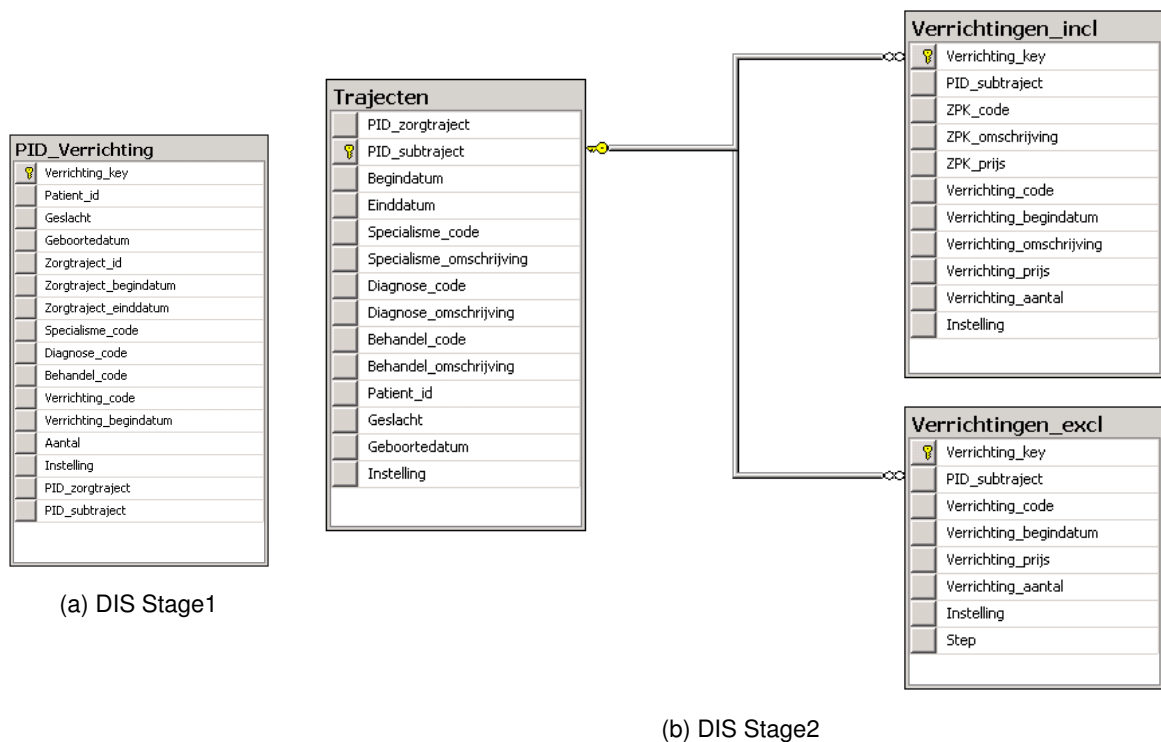


Figure A.3: Database Staging models

A.3 DIS_Stage2

Tablename	description
Carepaths	One description row for each Carepath
Activities-incl	Activities that belong to a specific Activity-class
Activities-excl	Activities that do not belong to a any Activity-class

Second cleaning stage: unjoining PID-Activities from Stage1 into ProcessInstances (Carepaths) and AuditTrailEntries (Activities)

- Select unique ProcessID's, also add descriptions for Specialism, Diagnosis and Treatment.
- Select unique Activities with linking ID's, add Activity-class and description
- Select unique Activities with linking ID's without Activity-classes separately.

A.4 DIS_Data

Tablename	description
Outputs	Overview of parameters per EXPORT
Carepaths	Collection of selected carepaths, either by Carepath level or DBC level
Activities	Activities
Prices	Overview of prices per Carepath: based on either Activities, Activity Classes and Grouped per class per day



Collection analysis sets: every analysis sample set is copied to this database with a unique ID.

- Import parameters (Specialism/Diagnosis/Treatment and PID-level)
- Two exports are available, due to different datatype requirements per tool:
 - R** Transactional data for activities per Carepath, plus a small overview of Carepath Prices (for cluster validation).
 - ProM** – ProcessInstances based on either complete Carepaths or unique DBC's,
 - AuditTrailEntries of every unique Activity (from incl)
 - AuditTrailEntries with one Activity Class per day

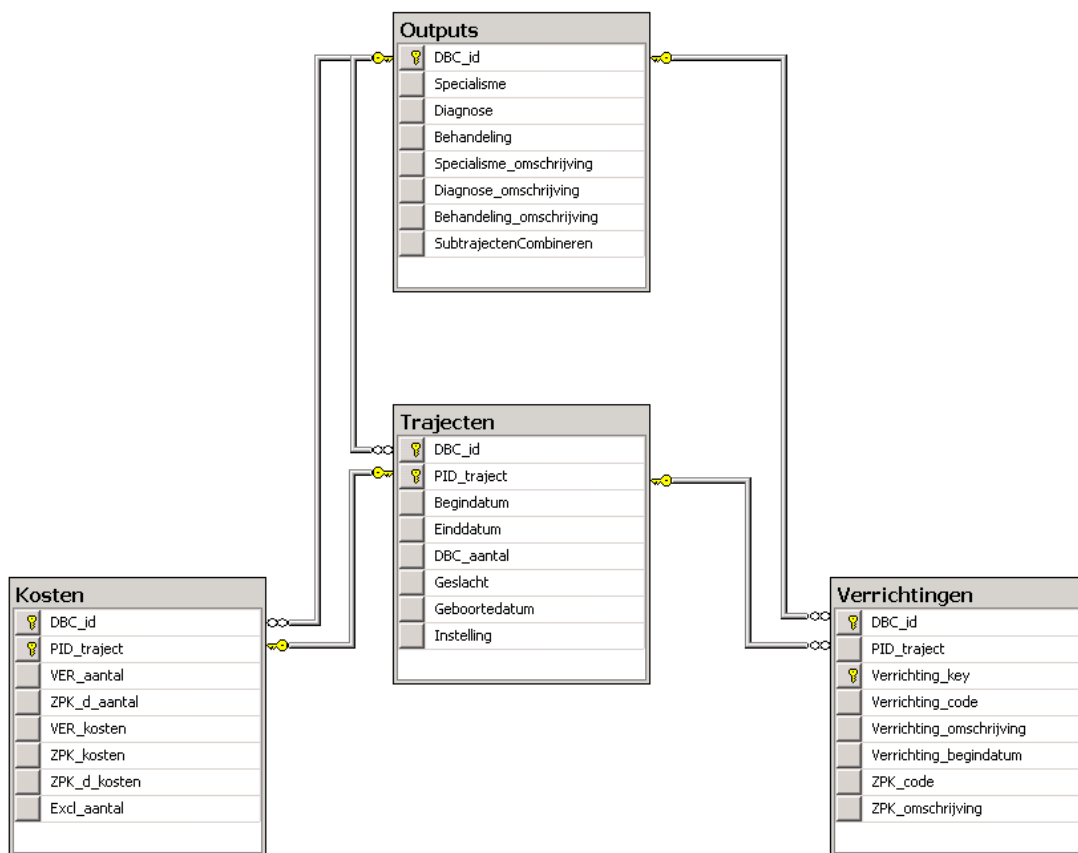


Figure A.4: Database models

A.5 R

For reading in R we use two tables as input:

Activities similar to the AuditTrailEntries table above.

Instances or Patiens or Paths, per Patient/Path-ID a list of different costs.

Before we are able to use these tables as input we first need to transform the Activities table.

String each activity at ZPK-level is represented by a character (or n -characters), such that each Patient's path is represented by a string.

Count each patient's path is represented by a vector of counts per activity.

Norm same as above, but then the values are scaled per activity over all patients between $0 \dots 1$.

A.6 ProM

ProM takes either MXML or XES formatted files as input. To convert our input sample datasets, we use ProM Import, a easy tool that can take an MS Access Database and convert it to MXML format. But the data has to be formatted to a certain transactional type of database first:

A.6.1 Using MS Access to prepare the output datasets

Taking the ProM output tables ProcessInstances (Trajecten) and AuditTrailEntries (unique or distinct Verrichtingen), we need to transform the data to the ProM Import model.

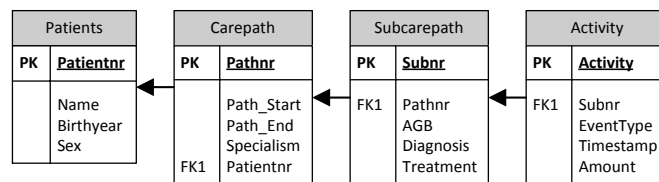


Figure A.5: Datamodel for ProM Input

For this, we use the provided MS Access template that has a macro.

A.6.2 ProM Import

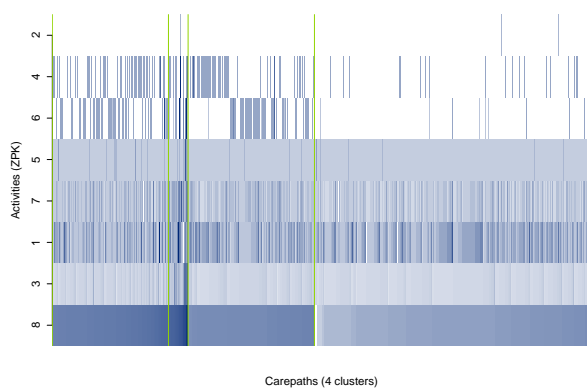
Once the MS Access File is transformed (estimated runtime per number of PI's:) we run ProM Import which exports an MXML file.



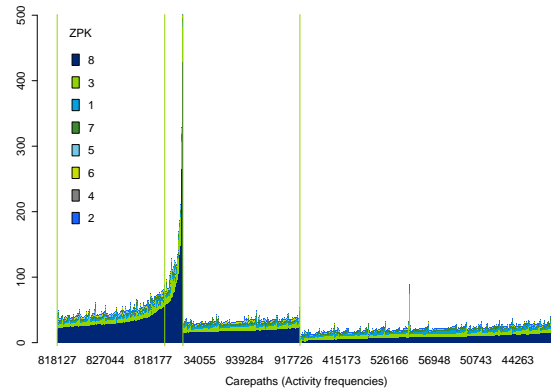
Appendix B

Modeling images

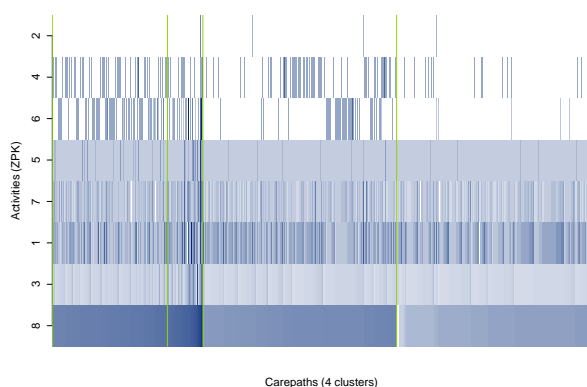
B.1 Vector clustering ZPK 1 to 8



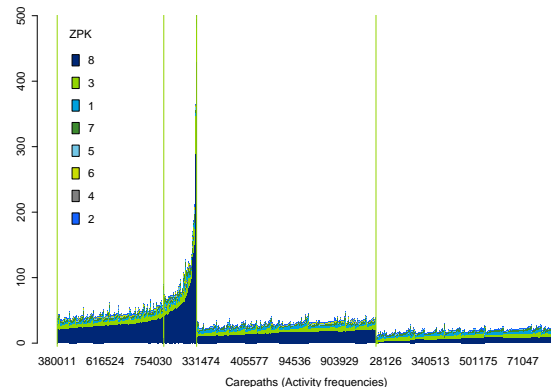
(a) Euclidean distance with `hclust`



(b) Euclidean distance with `hclust` (Barchart)



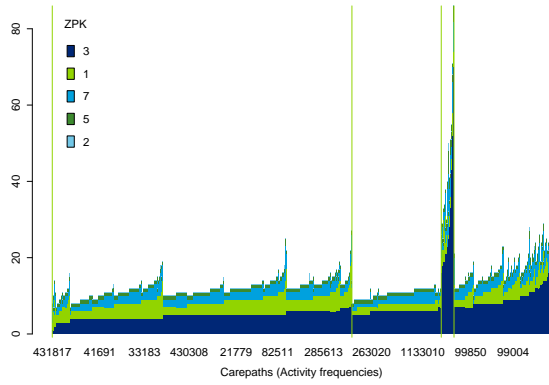
(c) Tanimoto distance with `pam`



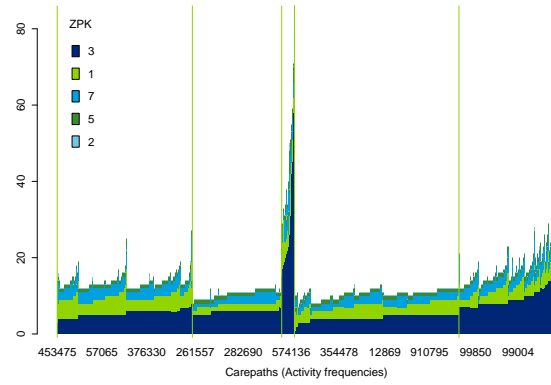
(d) Tanimoto distance with `pam` (Barchart)

Figure B.1: Clustering example for ZPK 1 to 8 (DBC 305..1701.223). We consider this a good example of *bad clustering*, as it is based on a single variable.

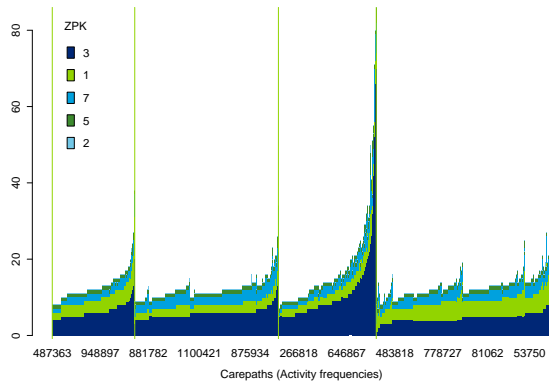
B.2 Comparing `hclust` with `pam` for different distance measures



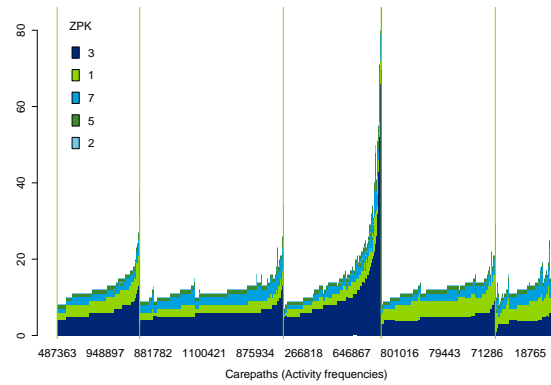
(a) Euclidean distance, 4 clusters



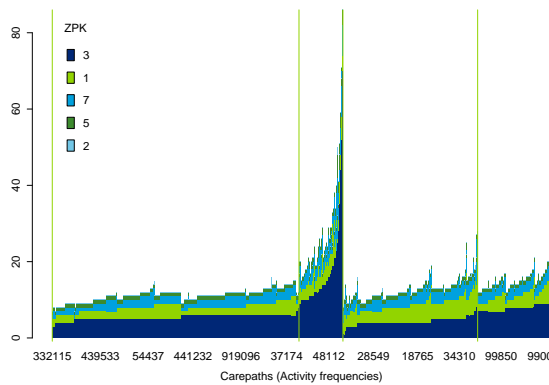
(b) Euclidean distance, 5 clusters



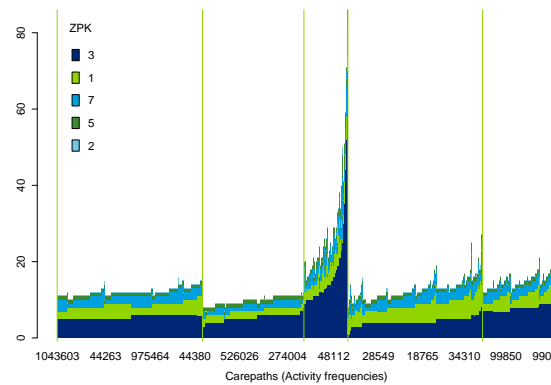
(c) Cosine distance, 4 clusters



(d) Cosine distance, 5 clusters

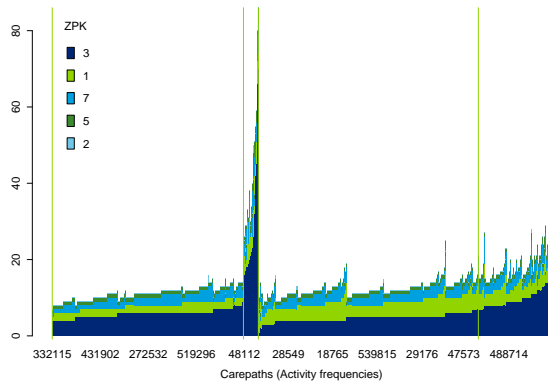


(e) Tanimoto distance, 4 clusters

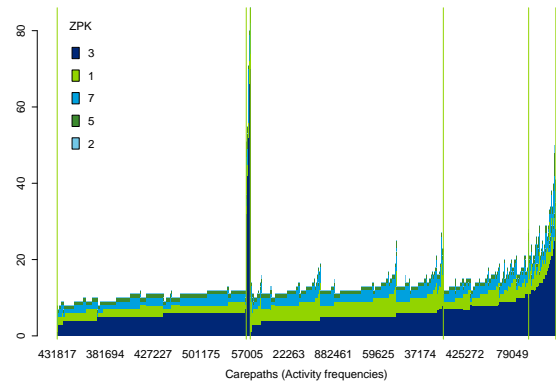


(f) Tanimoto distance, 5 clusters

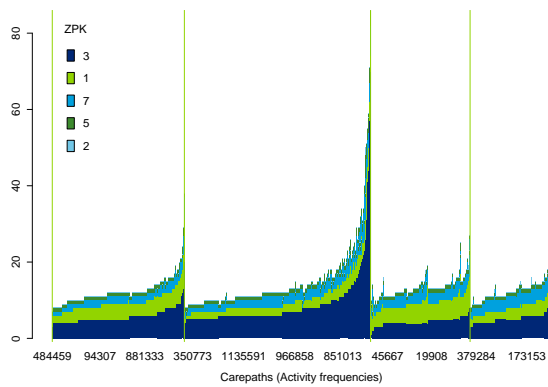
Figure B.2: Hierarchical clustering (`hclust` in R)



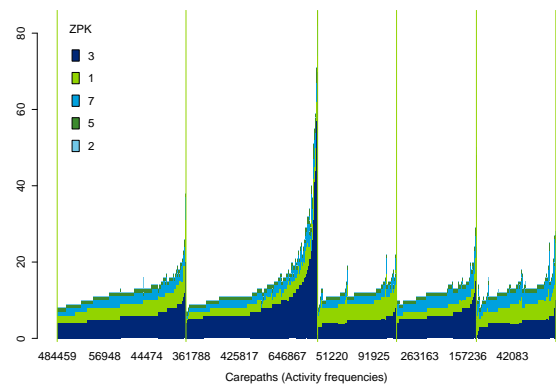
(a) Euclidean distance, 4 clusters



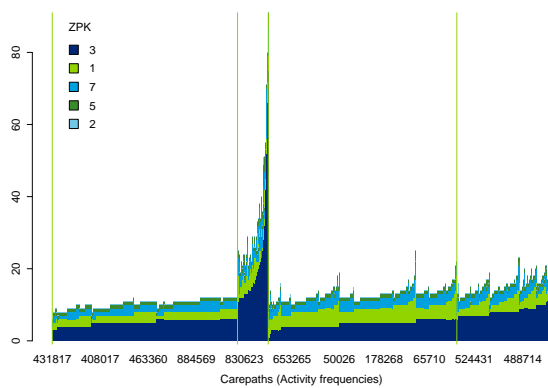
(b) Euclidean distance, 5 clusters



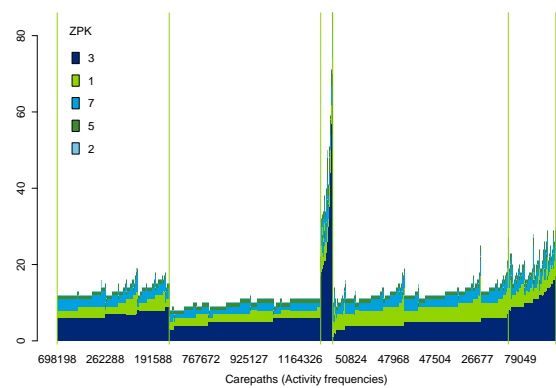
(c) Cosine distance, 4 clusters



(d) Cosine distance, 5 clusters

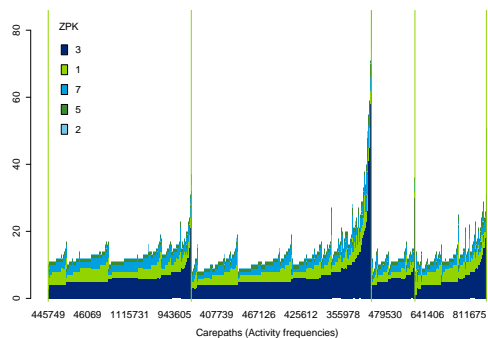


(e) Tanimoto distance, 4 clusters

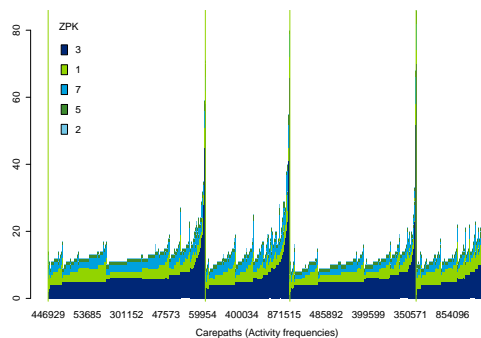


(f) Tanimoto distance, 5 clusters

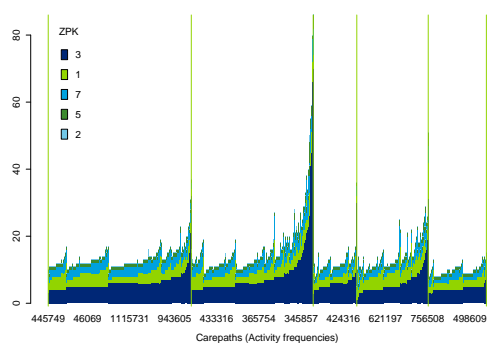
Figure B.3: Partitional clustering (`pam` in R)



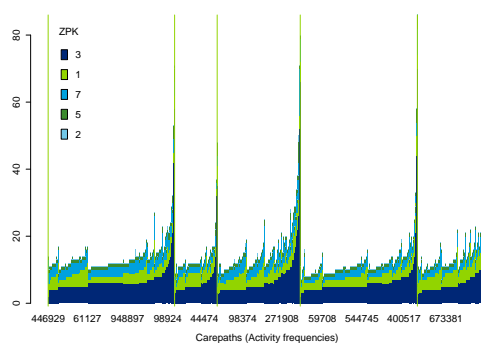
(a) Gzip compression, 4 clusters (hc1ust in R)



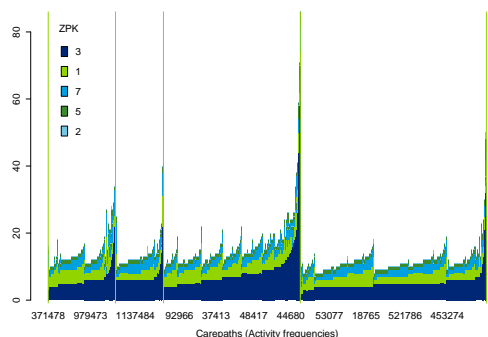
(b) Gzip compression, 4 clusters (pam in R)



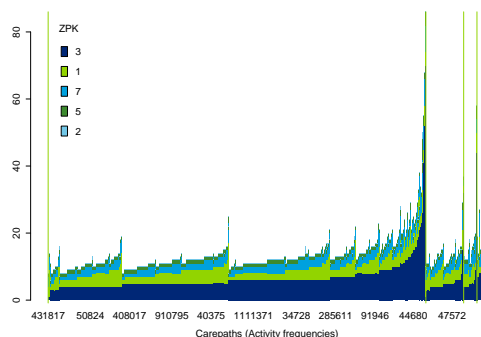
(c) Gzip compression, 5 clusters (hc1ust in R)



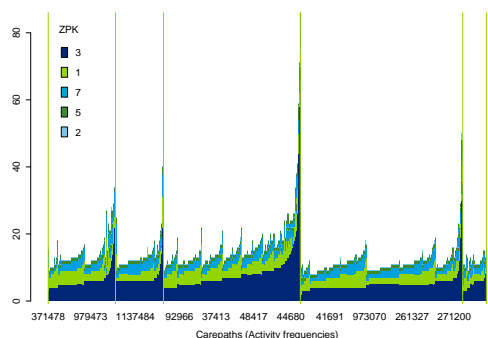
(d) Gzip compression, 5 clusters (pam in R)



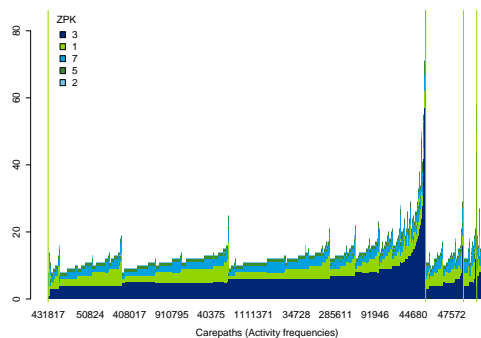
(e) Bzip2 compression, 4 clusters (hc1ust in R)



(f) Bzip2 compression, 4 clusters (pam in R)



(g) Bzip2 compression, 5 clusters (hc1ust in R)



(h) Bzip2 compression, 5 clusters (pam in R) - same as Figure B.4f

Figure B.4: Clustering Kolmogorov Complexity (Compression)



Appendix C

CP-Tables from `rpart`

Table C.1: `rpart` error

(a) Case study 1			(b) Case study 2			(c) Case study 3		
CP	nsplit	xerror	CP	nsplit	xerror	CP	nsplit	xerror
0.3840	0	1.046	0.3450	0	1.000	0.2156	0	1.014
0.0908	2	0.232	0.2231	1	0.655	0.2086	1	0.780
0.0019	3	0.141	0.1303	2	0.432	0.1181	3	0.390
0.0015	5	0.143	0.0576	3	0.302	0.0298	4	0.258
0.0008	8	0.140	0.0568	4	0.256	0.0157	6	0.193
0.0000	13	0.134	0.0526	5	0.199	0.0141	8	0.185
			0.0409	6	0.135	0.0105	9	0.158
			0.0175	7	0.094	0.0098	12	0.137
			0.0150	9	0.059	0.0000	78	0.051
			0.0067	10	0.044			
			0.0000	32	0.013			

Appendix D

Arthrosis (hip) - surgical/clinical with joint prosthesis

D.1 Activity frequency histograms

Activity frequency distributions per cluster: the clusters are given from left to right, top to bottom.

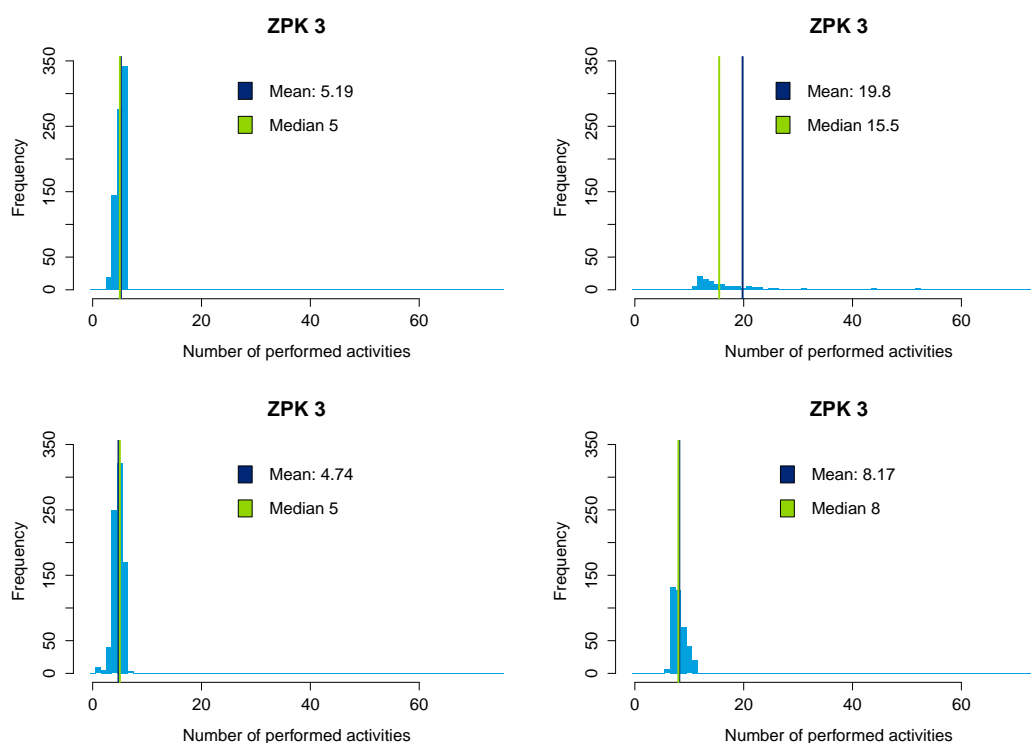
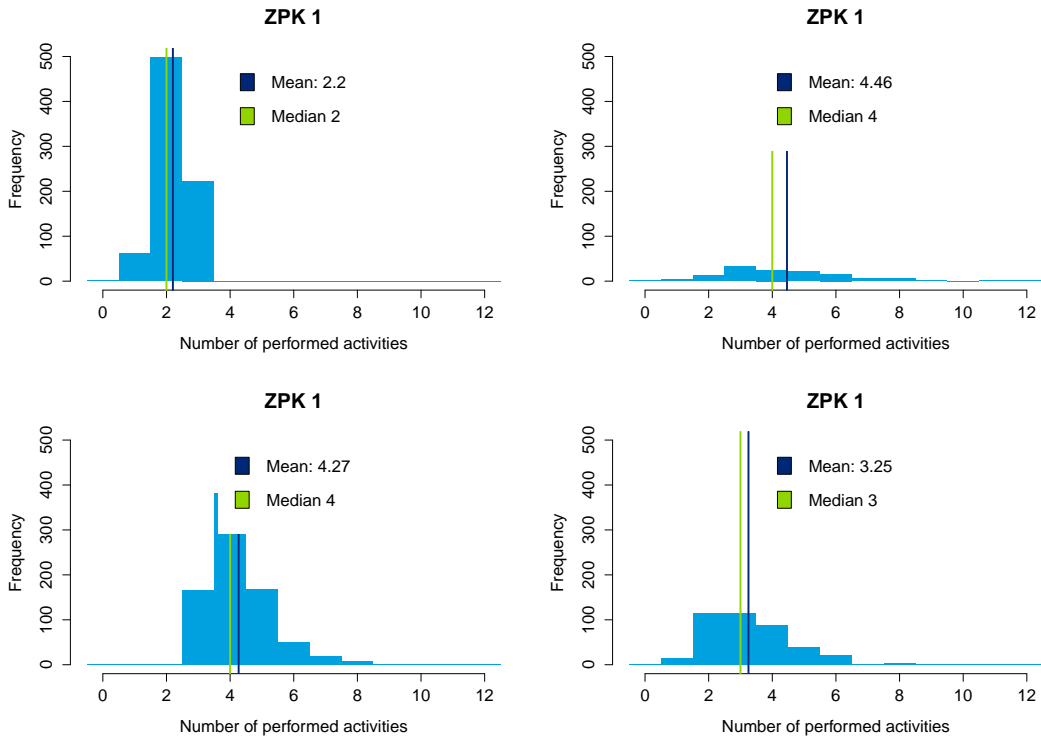
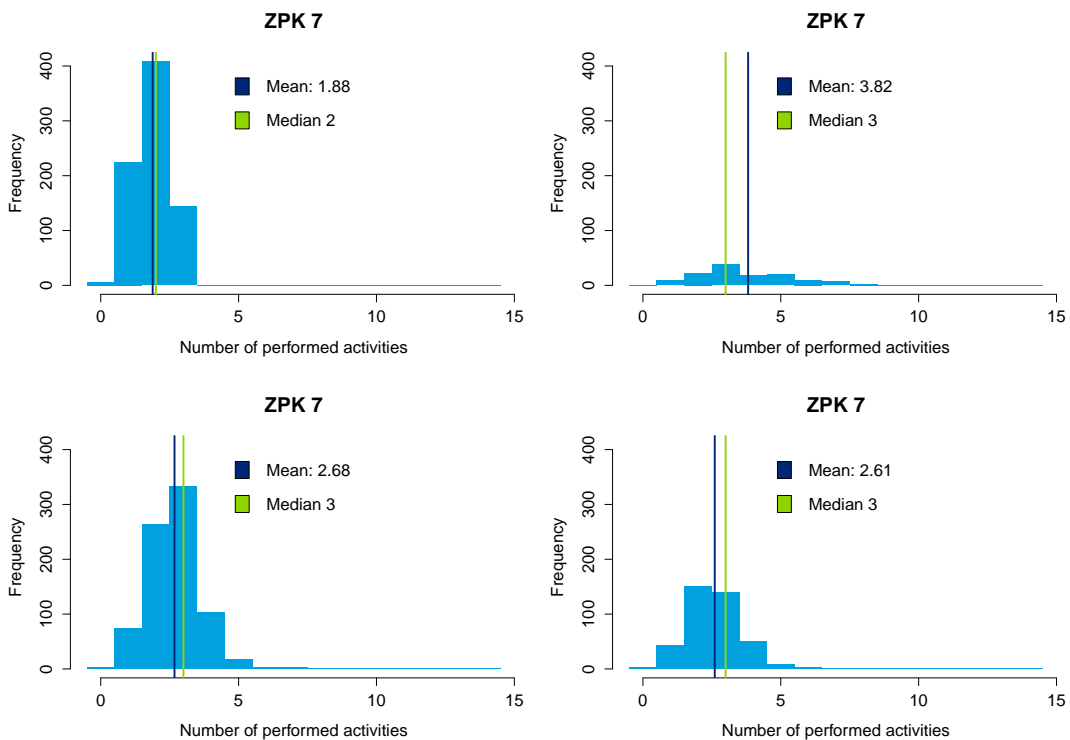


Figure D.1: The number of nursingdays shows the largest difference between individual clusters. Cluster 2 requires the longest hospitalization, with some extreme values (60+ days), whereas clusters 1 and 3 require about 5 days.

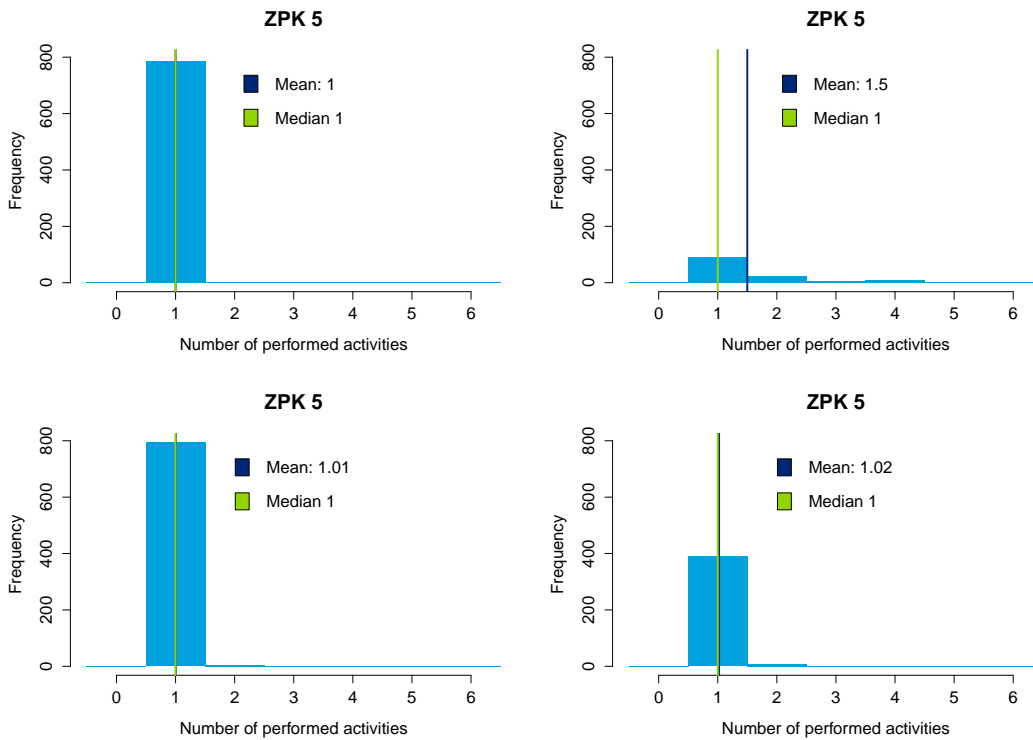


(a) Outpatient department

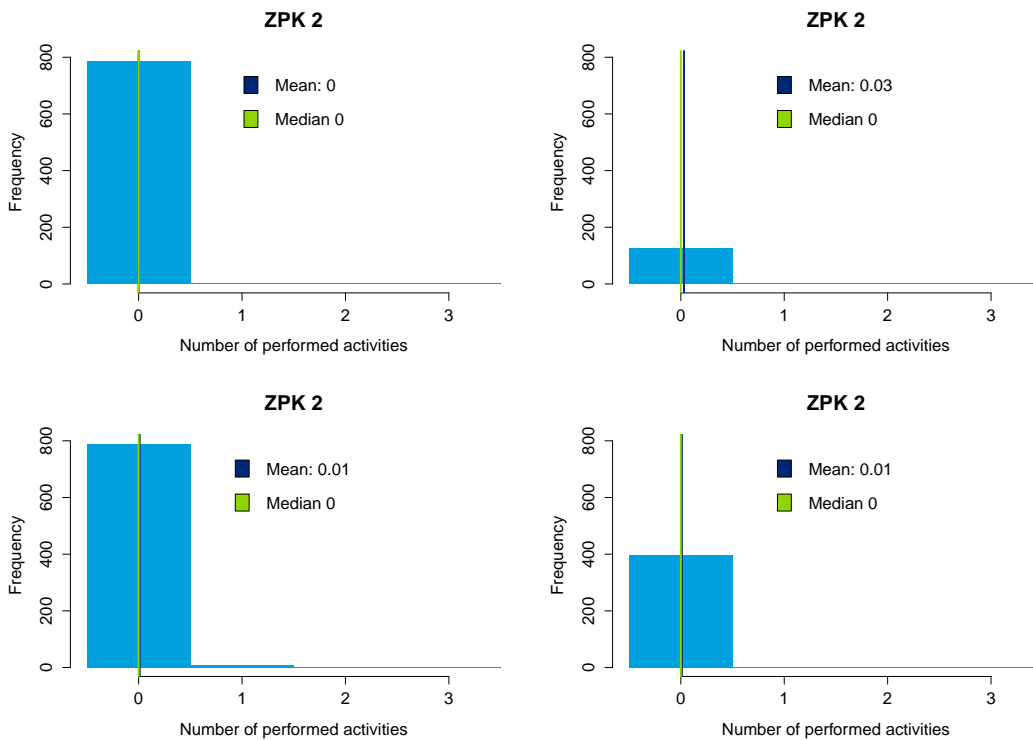


(b) Diagnostic activities

Figure D.2: Cluster 1 has the fewest activities for both ZPK's, whereas cluster 2 has the highest frequencies for both these activities. Cluster 3 and 4 are quite similar, although the latter is smaller.



(a) Surgery



(b) Other therapeutic activities

Figure D.3: For these two activities, there is not much difference in frequency per cluster. The only difference is that most double surgeries belong to cluster 2, which is also the cluster that requires the longest hospitalization (Figure D.1)

D.2 Trace Alignments

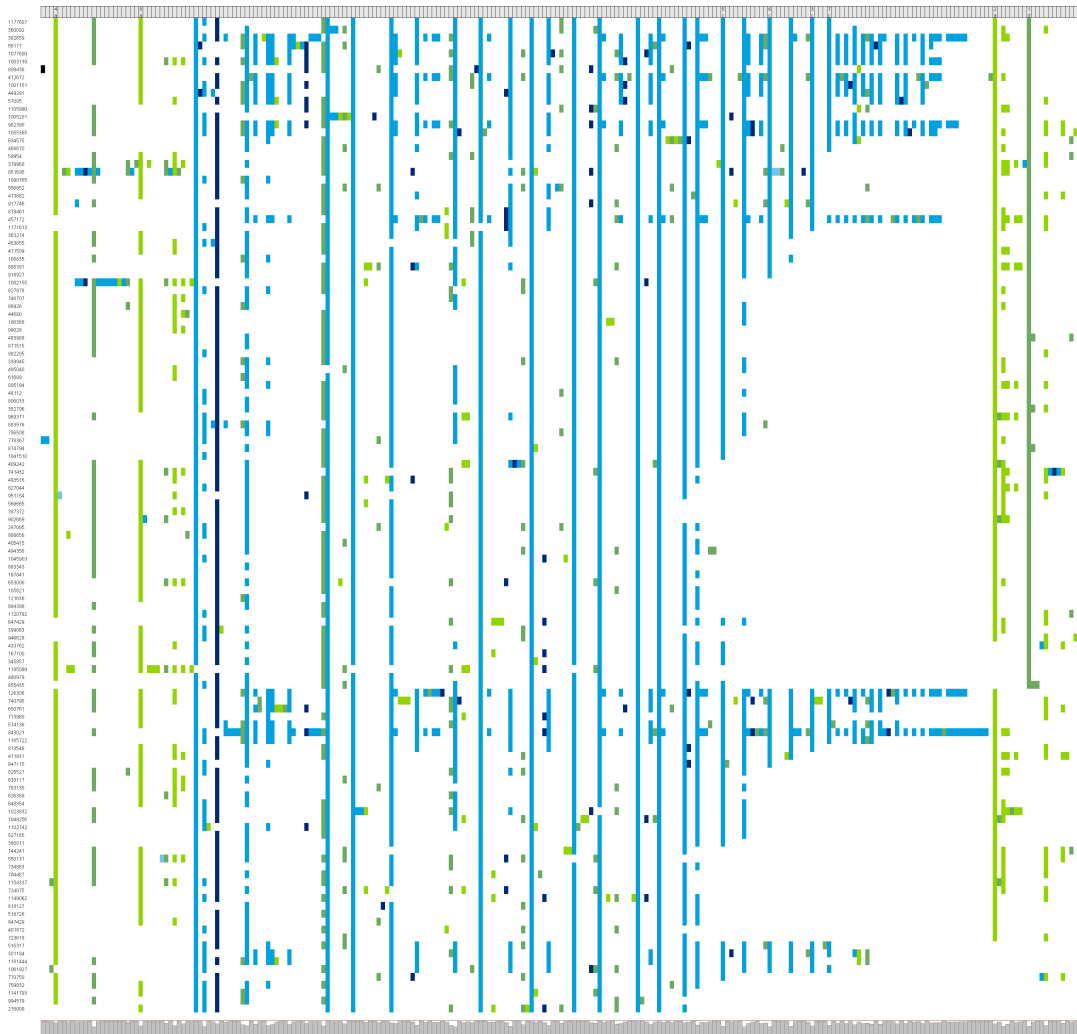
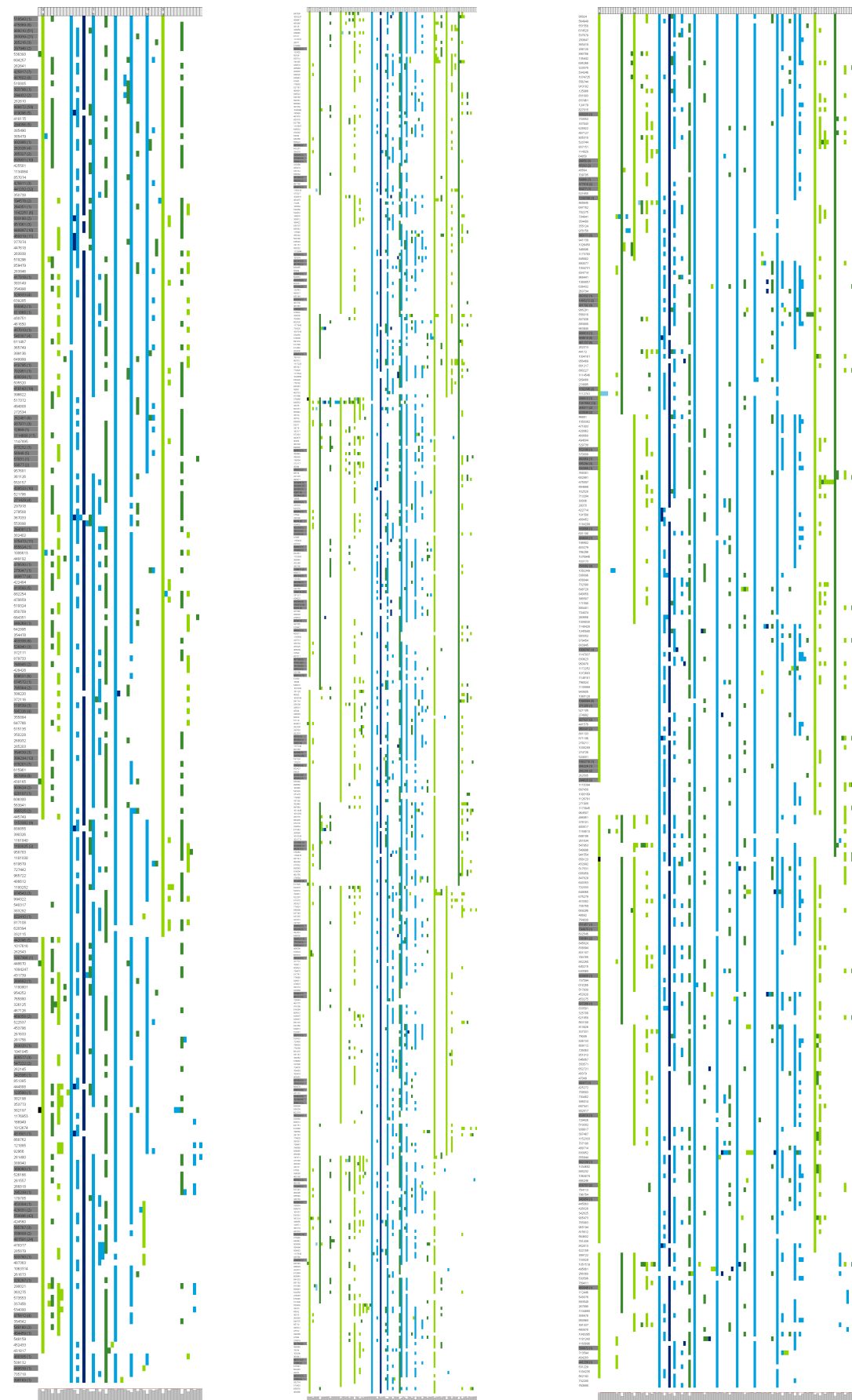


Figure D.4: Trace Alignment for cluster 2



(a) Cluster 1

(b) Cluster 3

(c) Cluster 4



Appendix E

Arthrosis (knee) - surgical/clinical with joint prosthesis

E.1 Activity frequency histograms

Activity frequency distributions per cluster: the clusters are given from left to right, top to bottom.

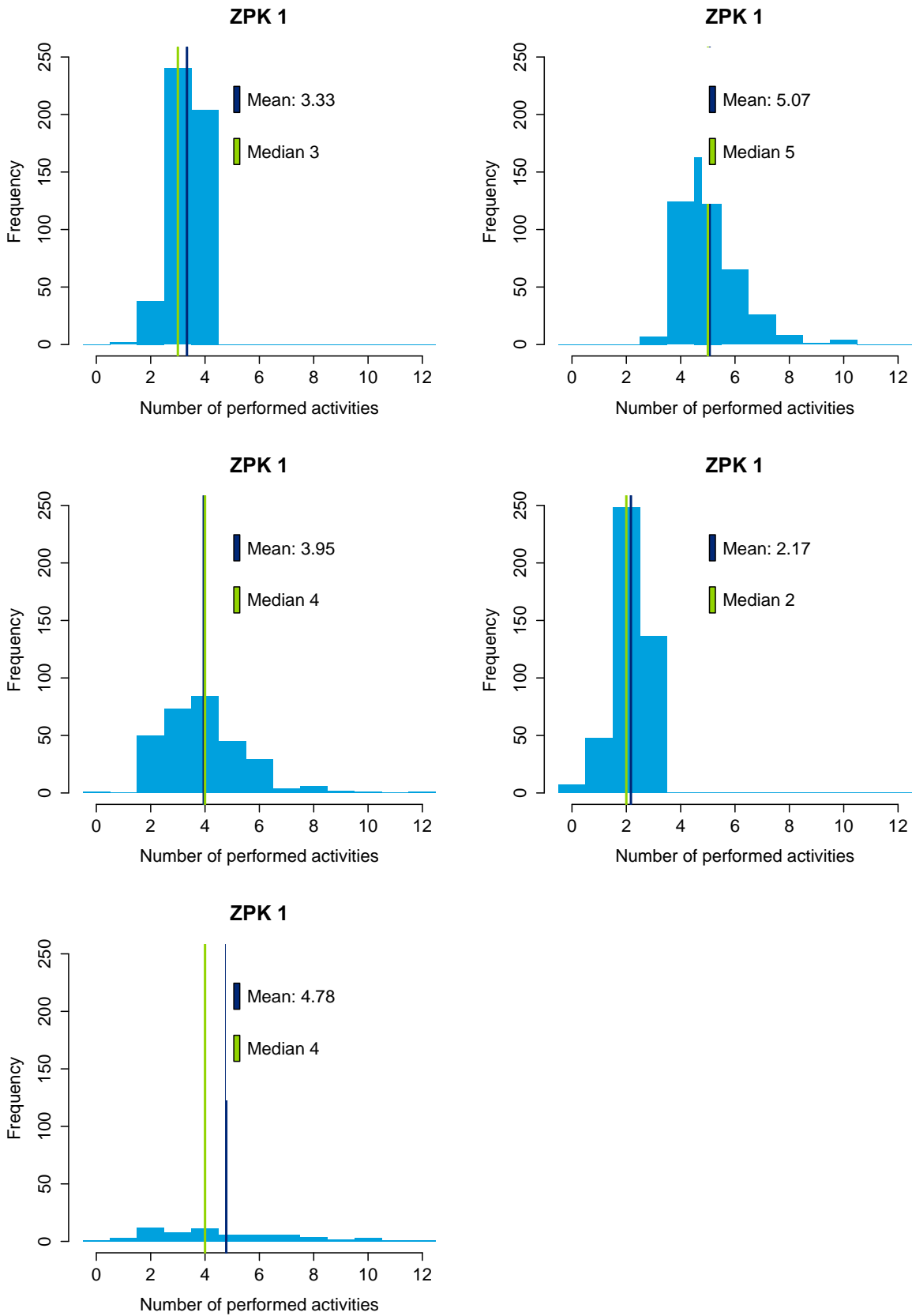


Figure E.1: Outpatient department

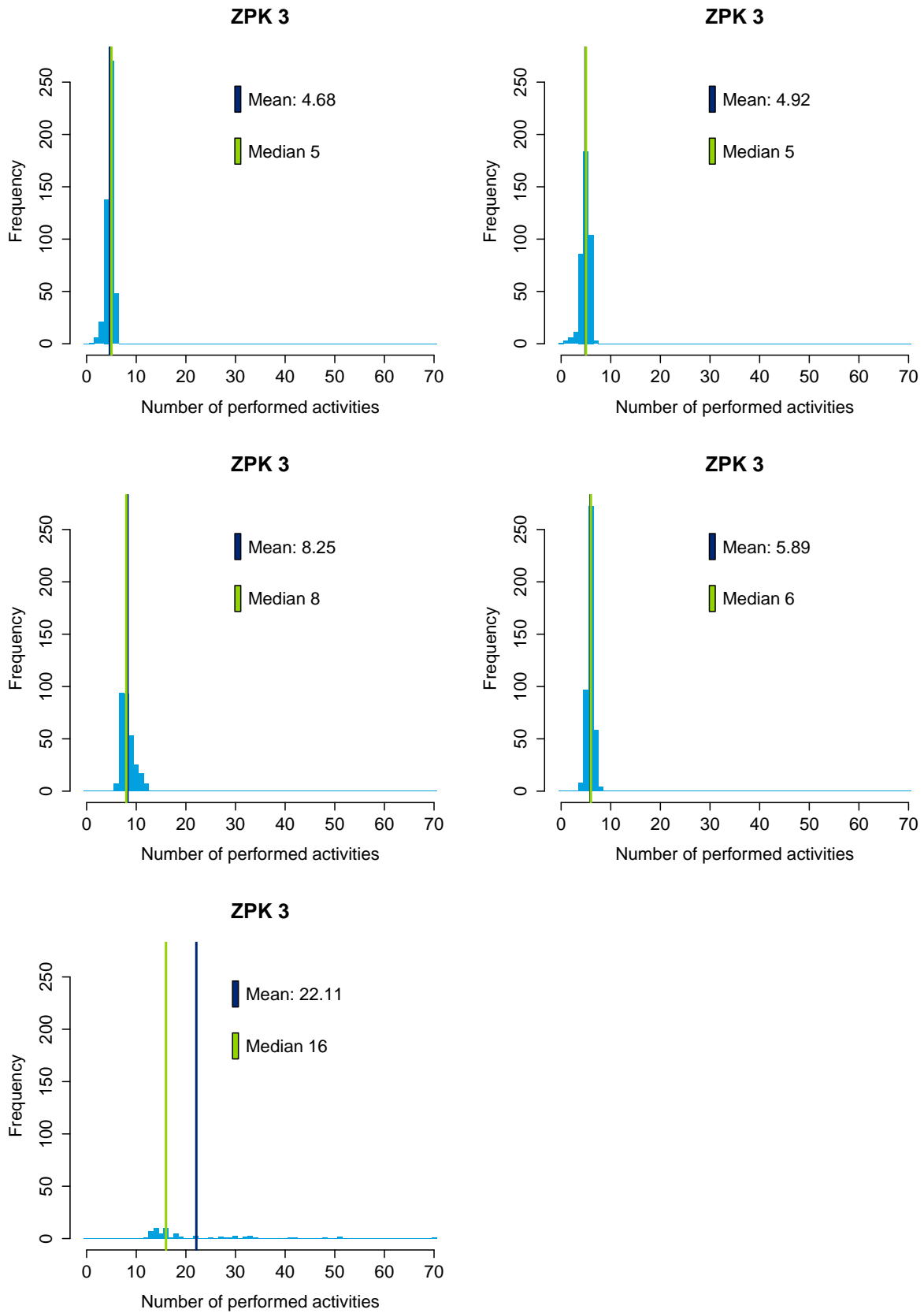


Figure E.2: Nursingdays

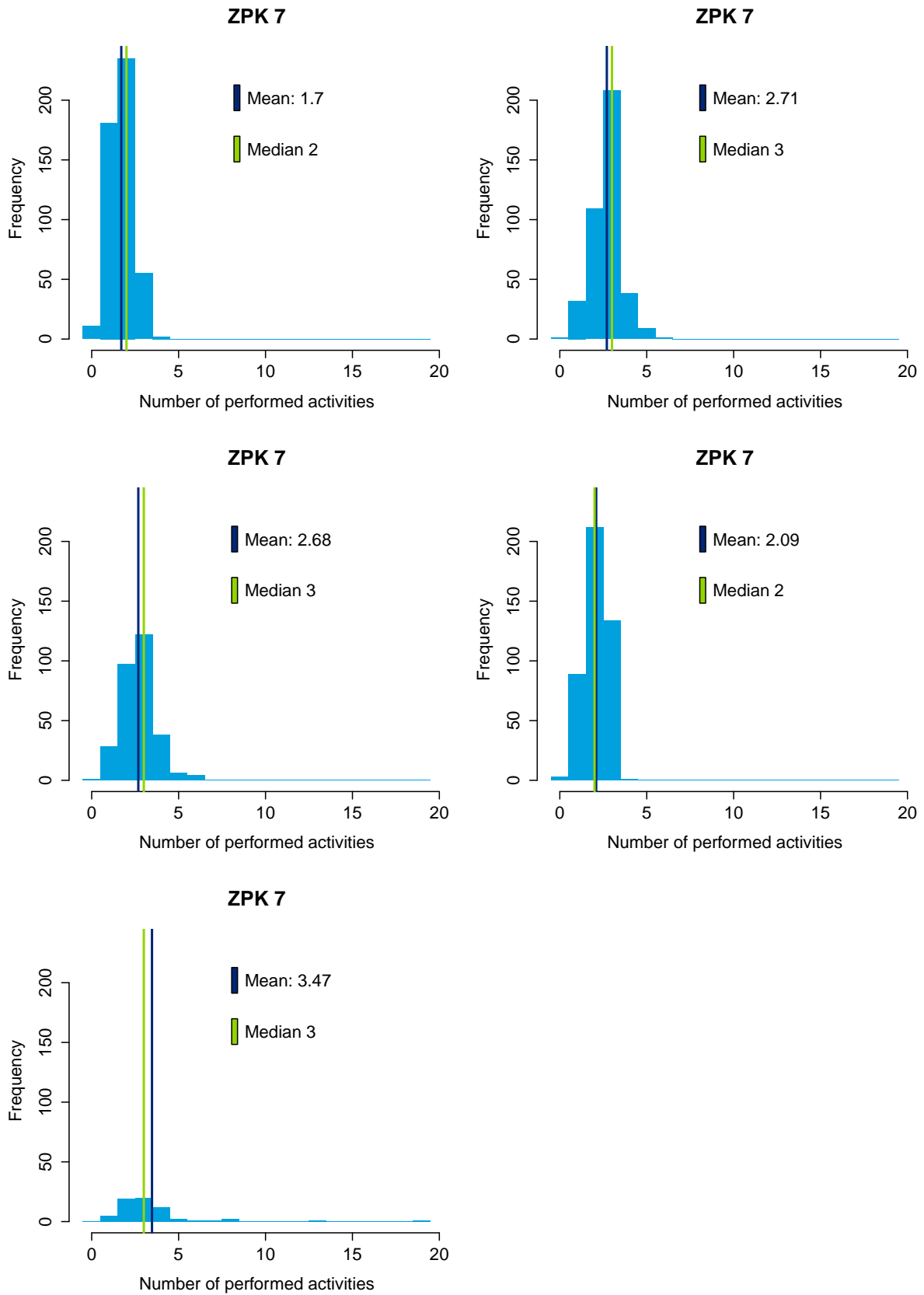


Figure E.3: Diagnostic activities

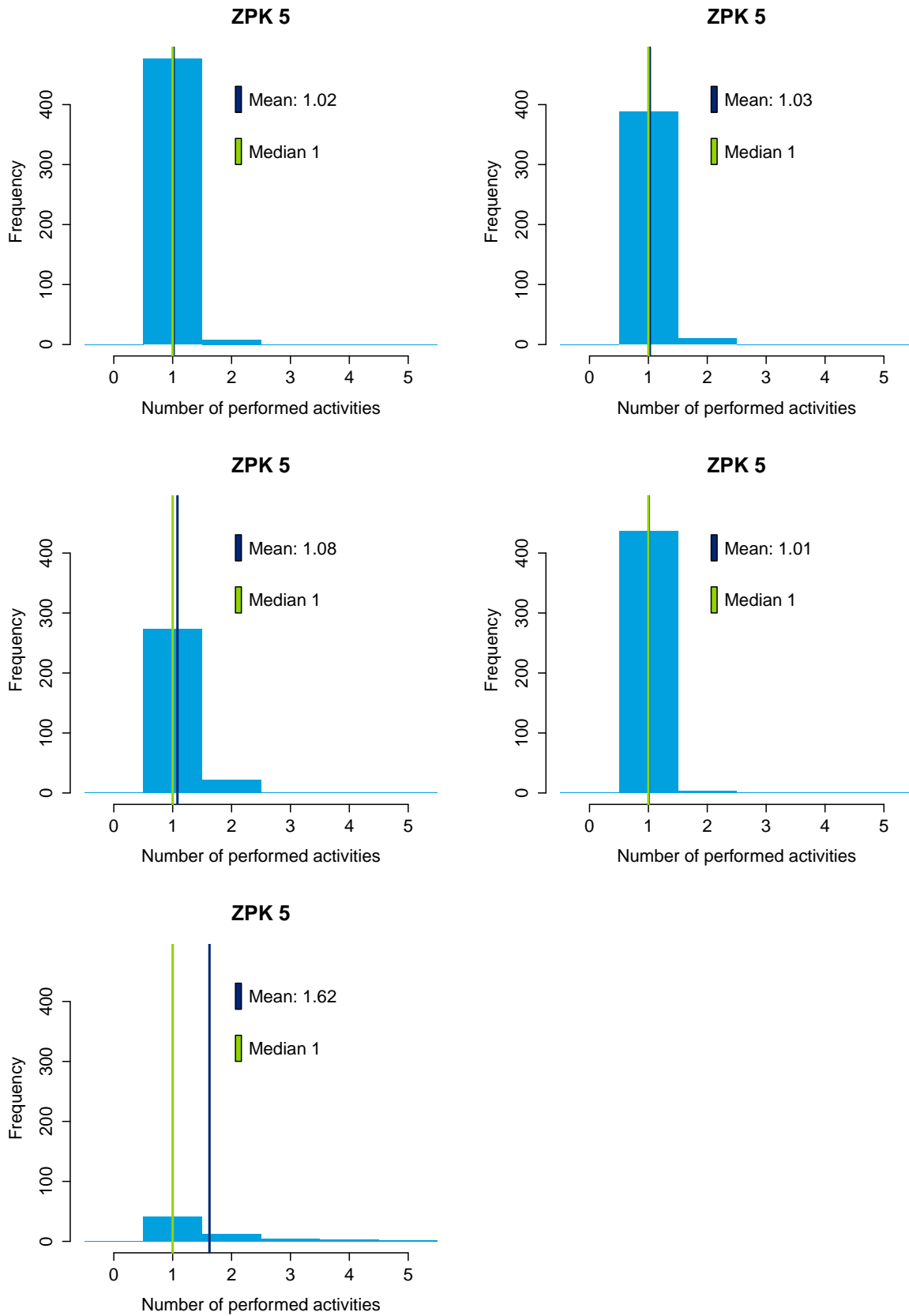


Figure E.4: Surgery

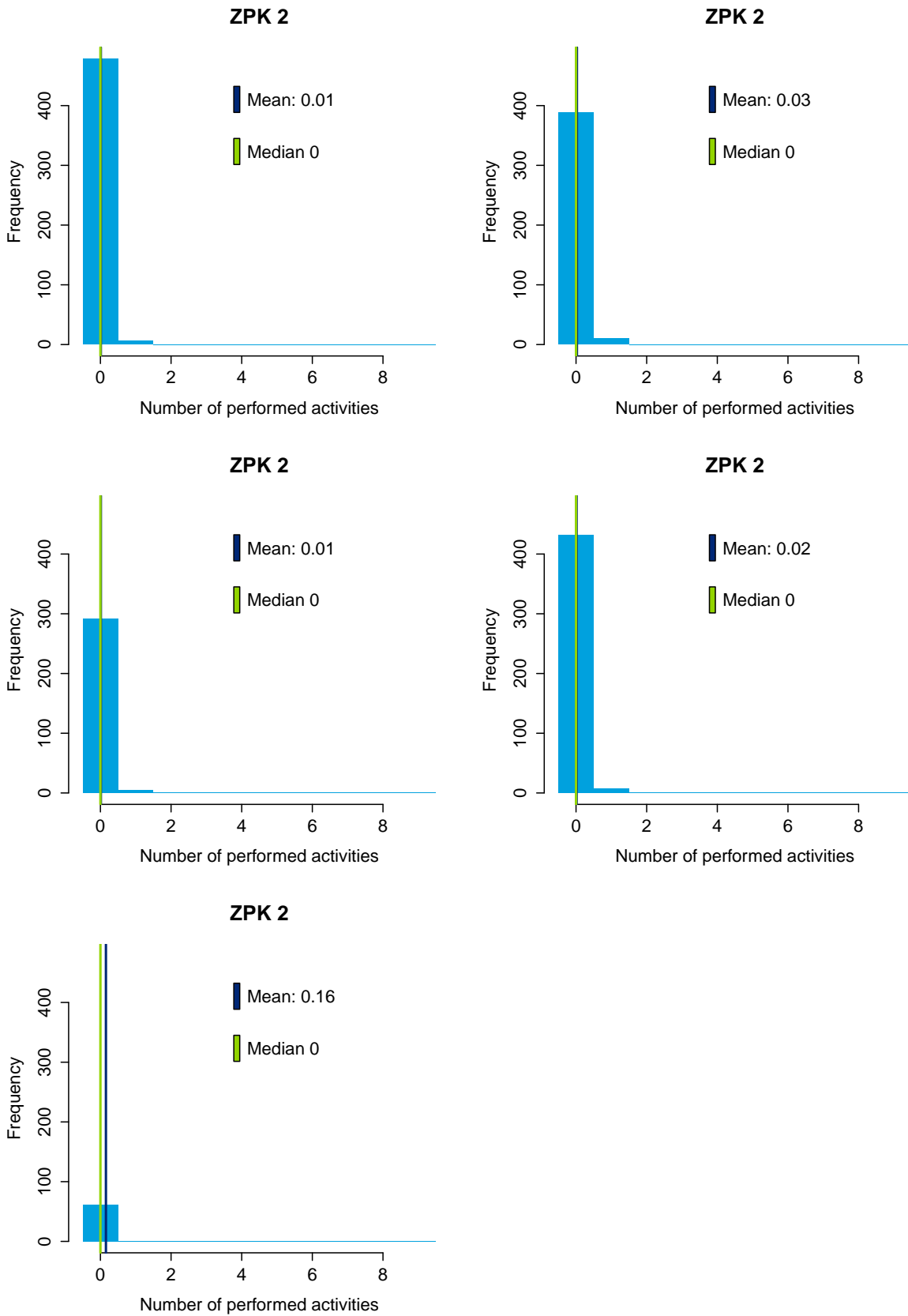


Figure E.5

E.2 Trace Alignments

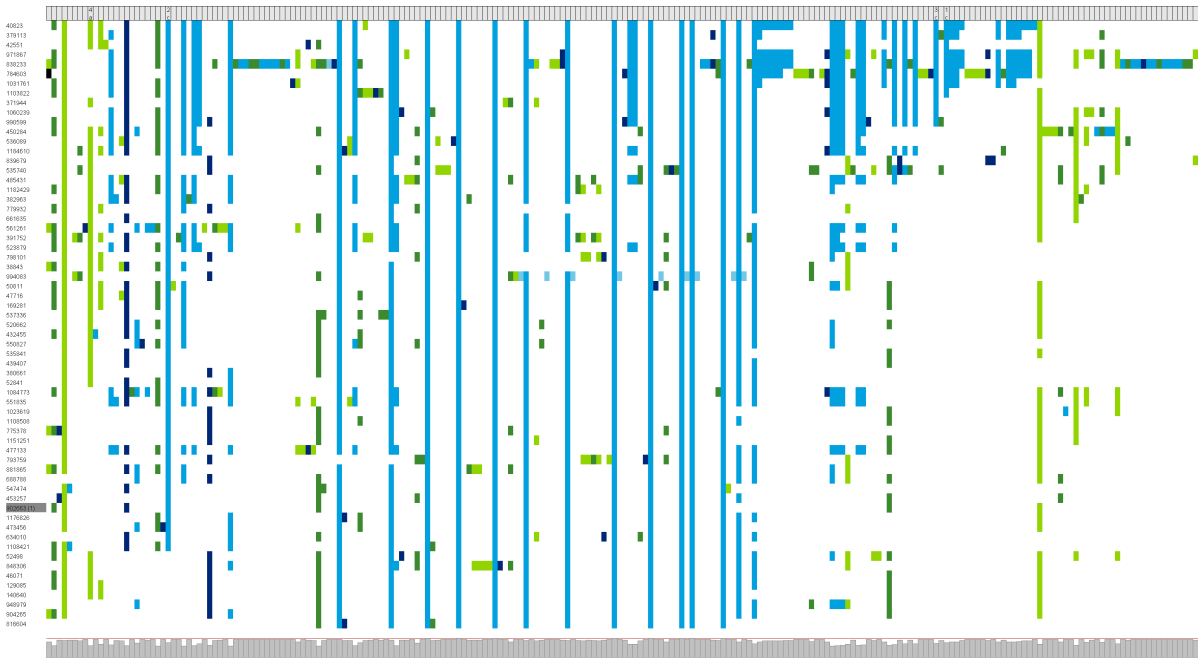
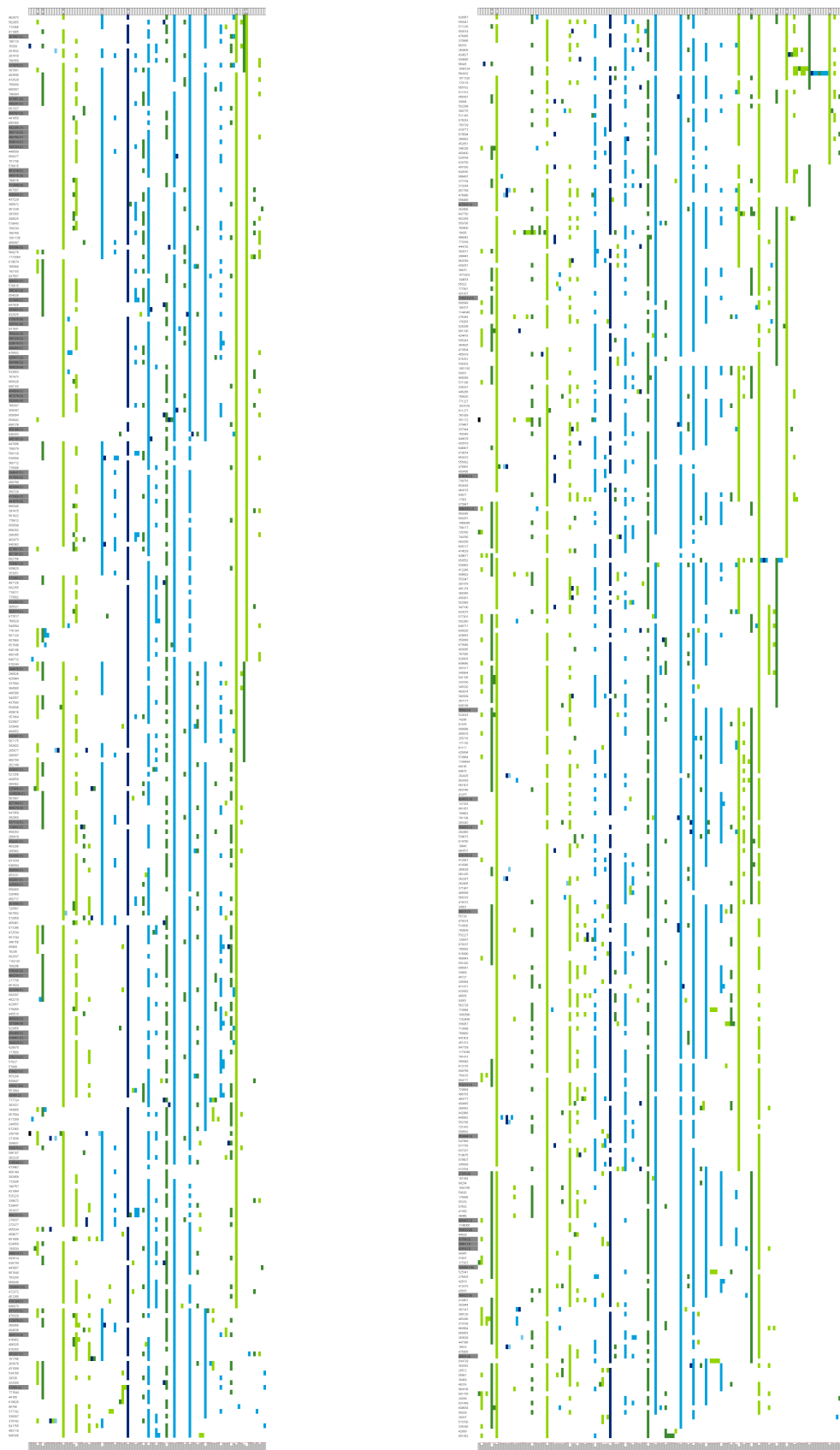


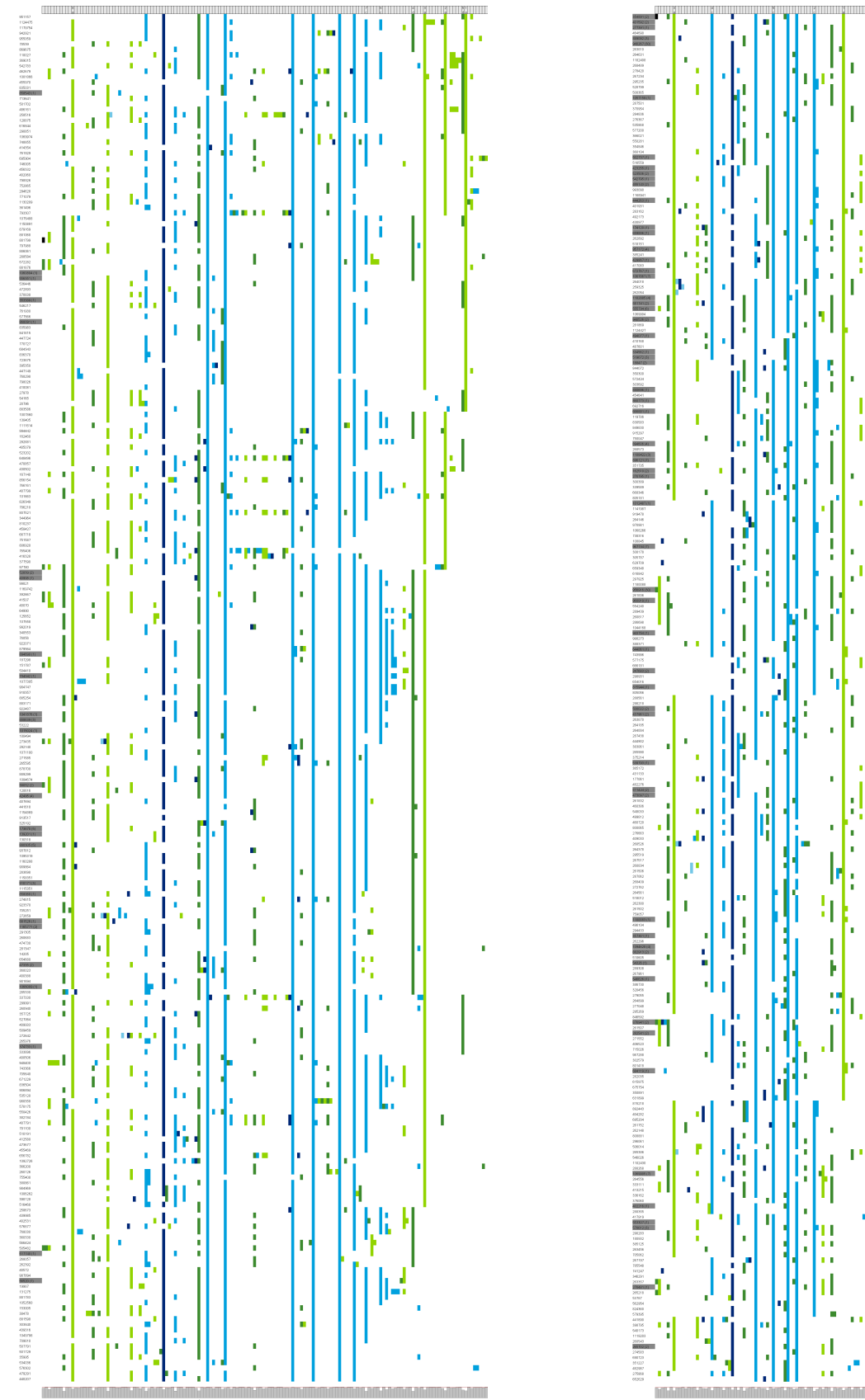
Figure E.6: Trace Alignment for cluster 5



(a) Cluster 1

(b) Cluster 2

Figure E.7: Trace Alignment for clusters 1 and 2



(a) Cluster 3

(b) Cluster 4

Figure E.8: Trace Alignment for clusters 3 and 4

Appendix F

Malignant breast neoplasm - surgical/clinical

F.1 Activity frequency histograms

Activity frequency distributions per cluster: the clusters are given from left to right, top to bottom.

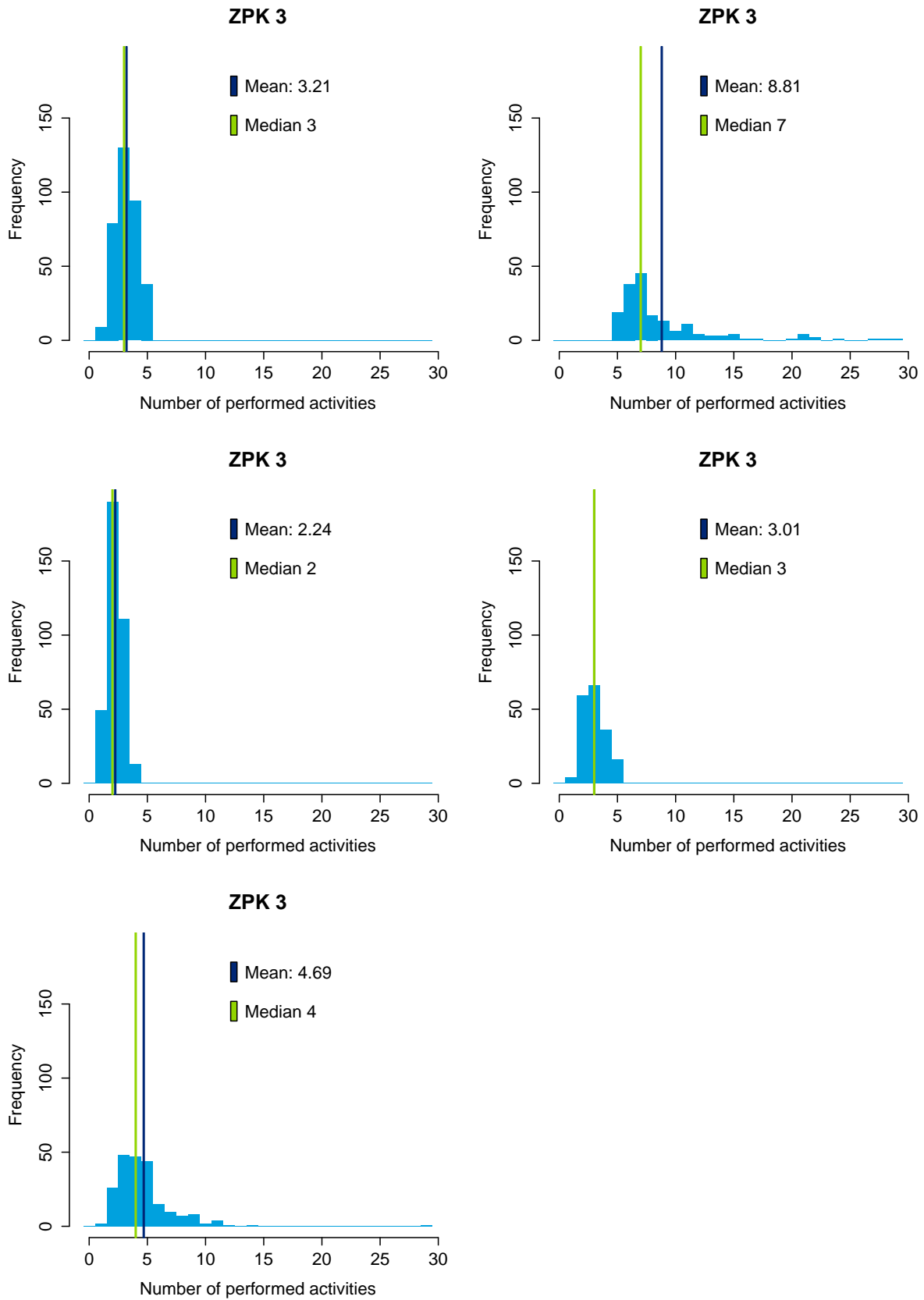


Figure F.1: Nursingdays

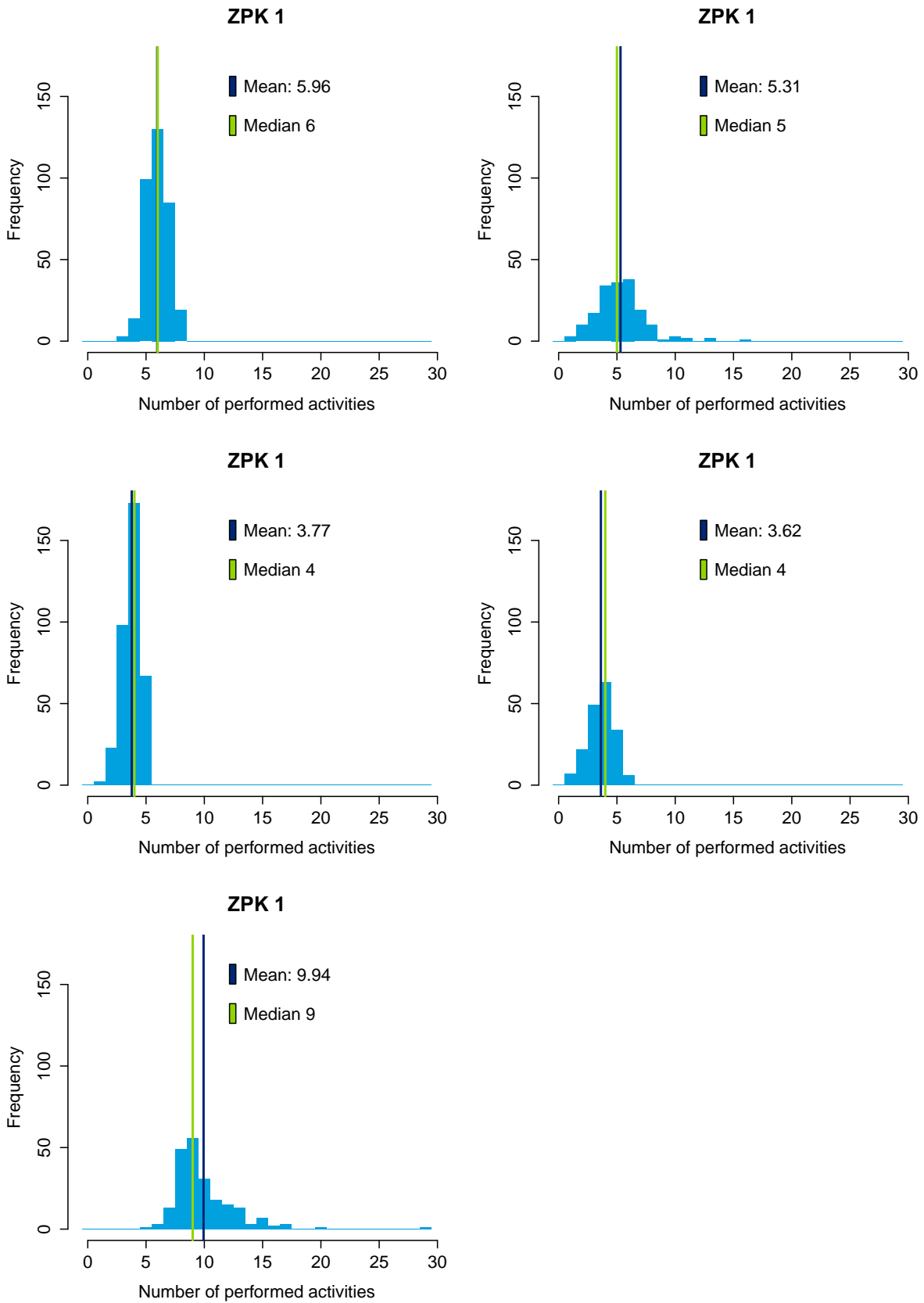


Figure F.2: Outpatient department

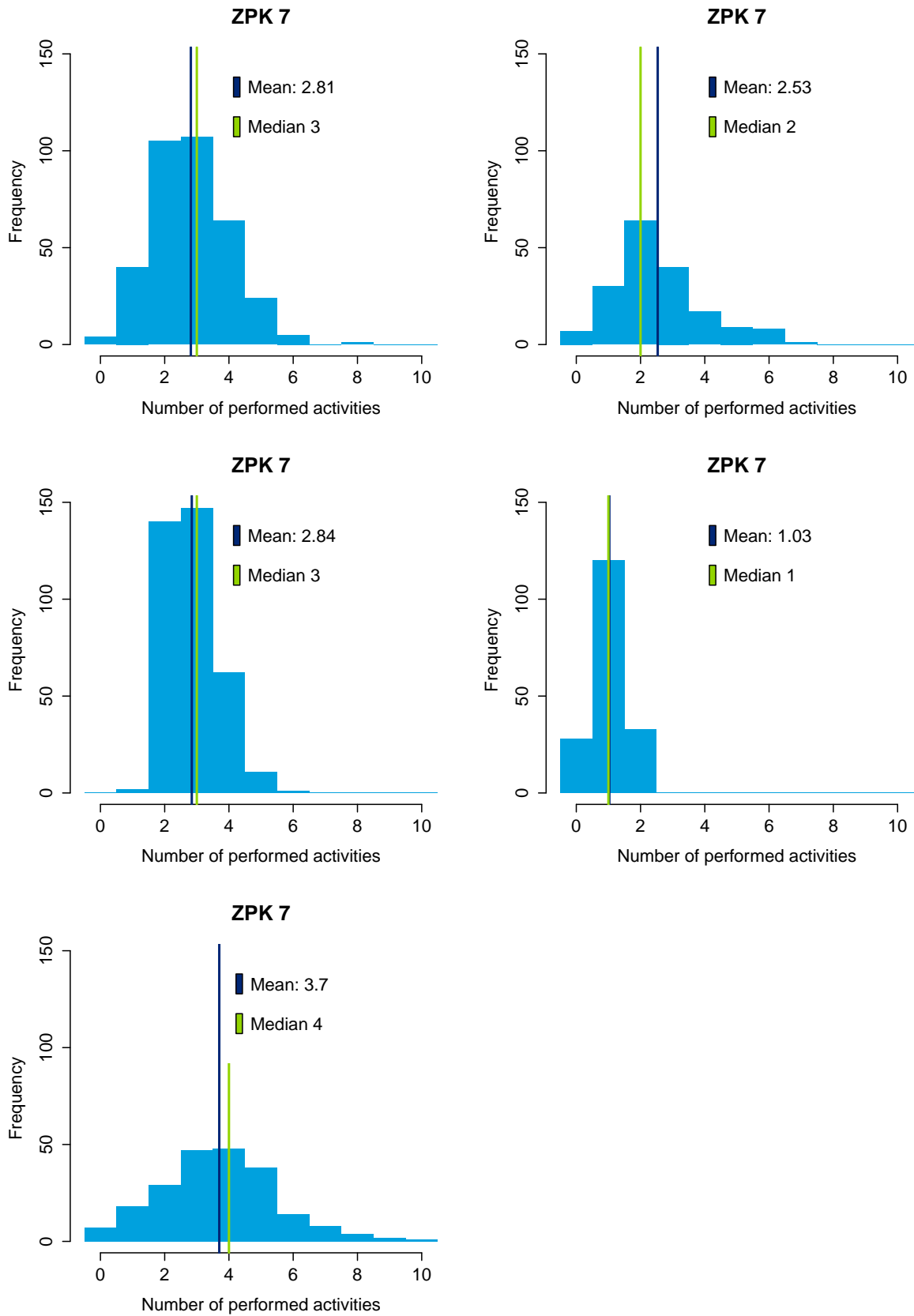


Figure F.3: Medical imaging

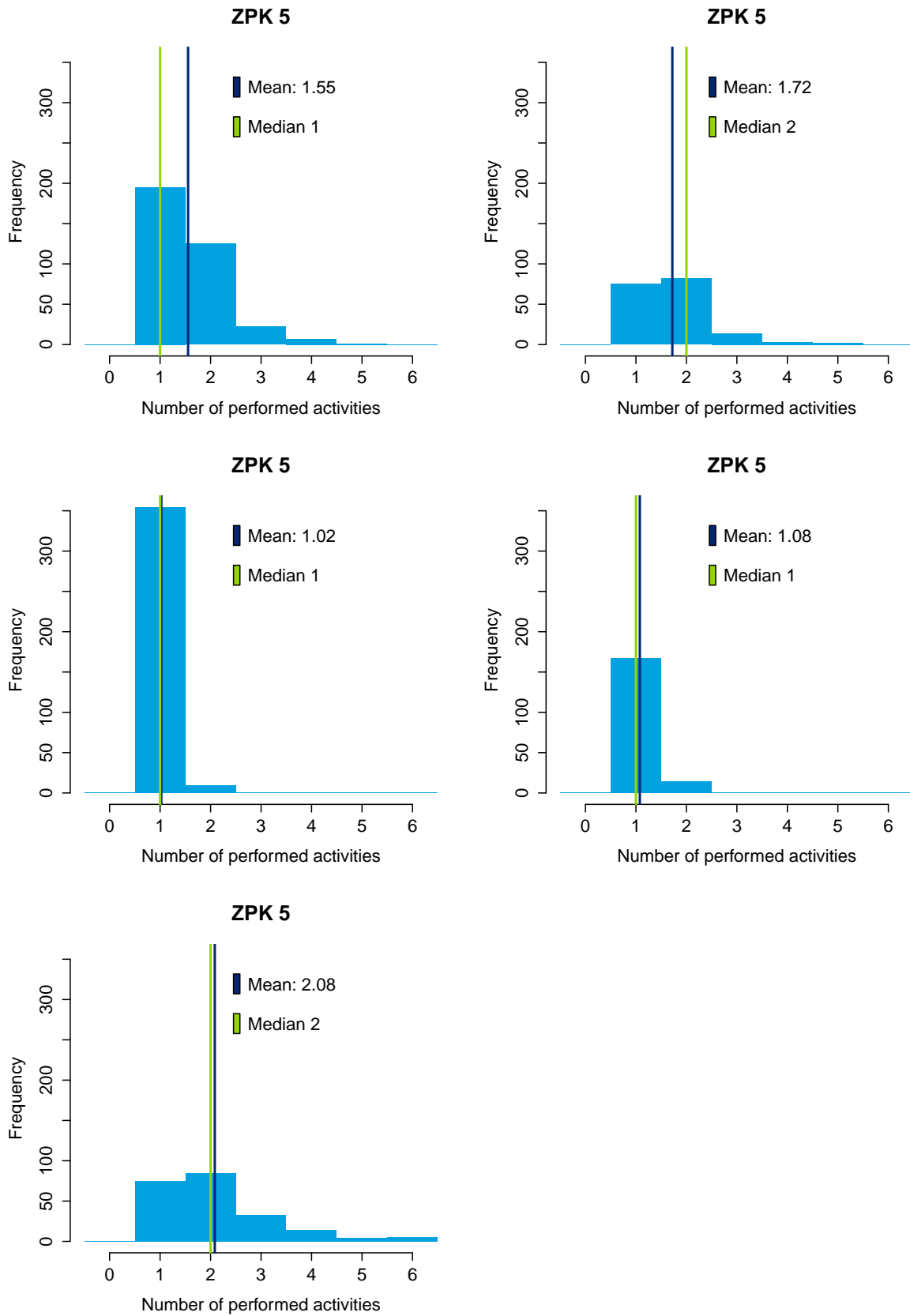


Figure F.4: Surgery

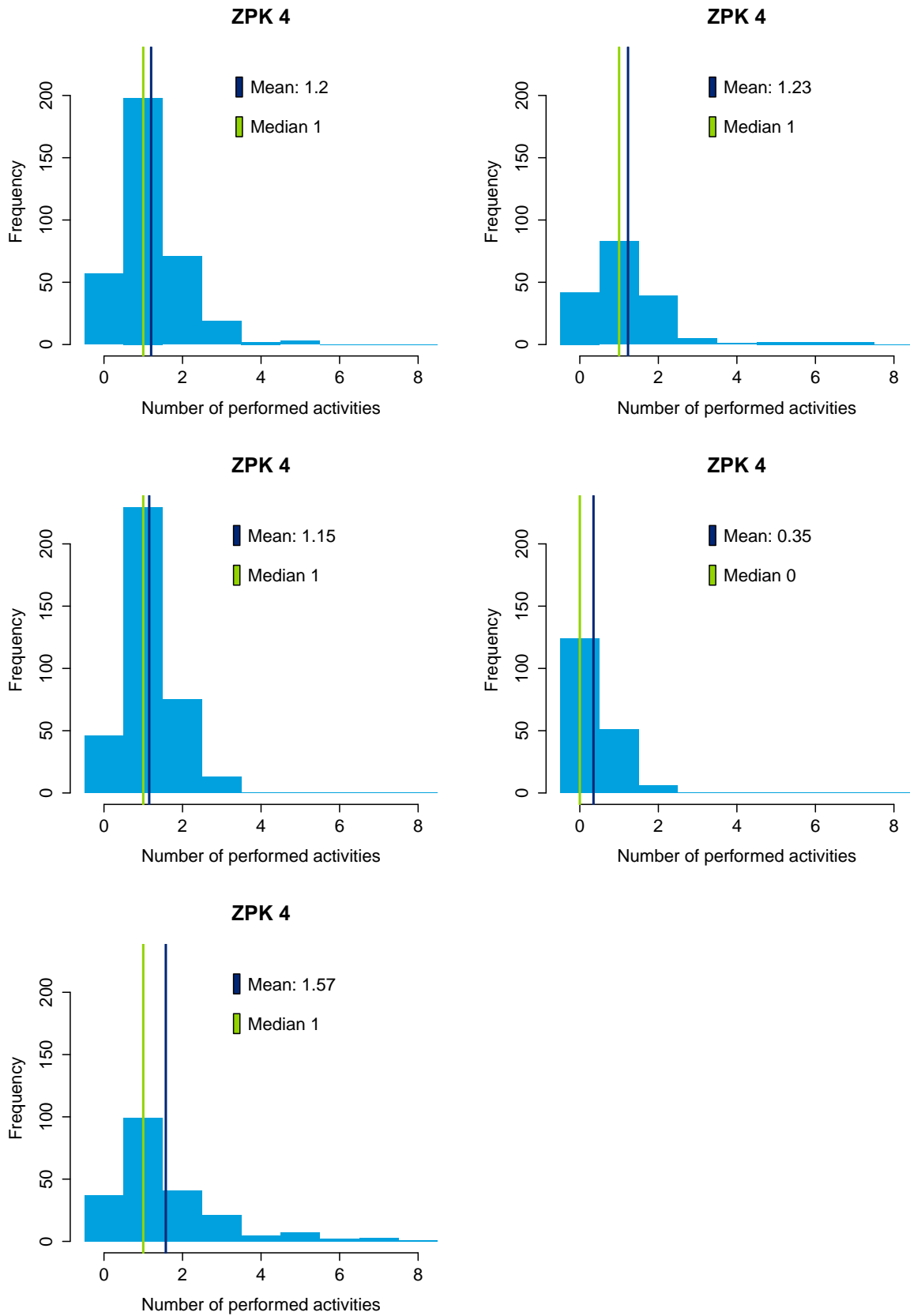


Figure F.5: Diagnostic activities

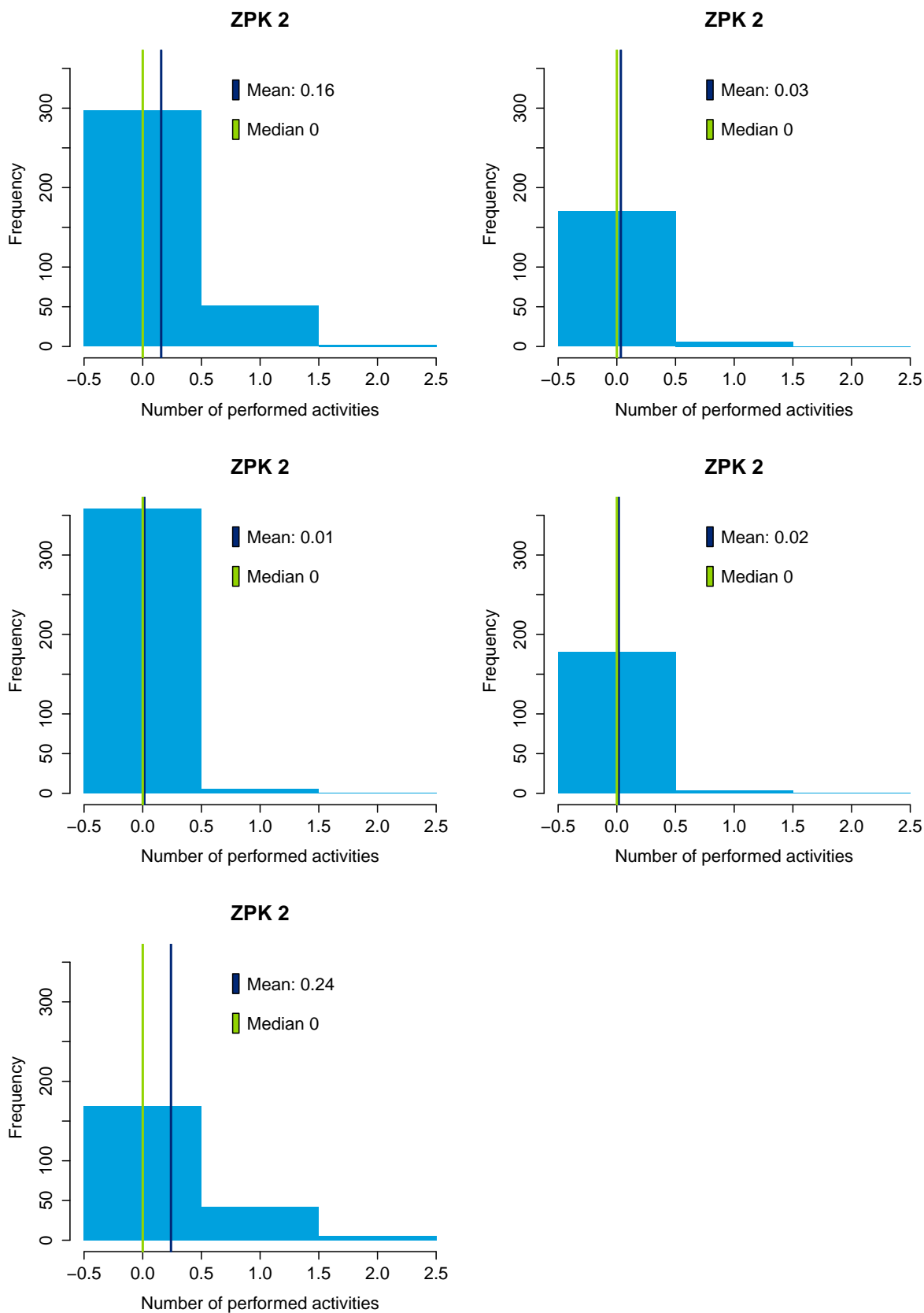


Figure F.6: Daycare

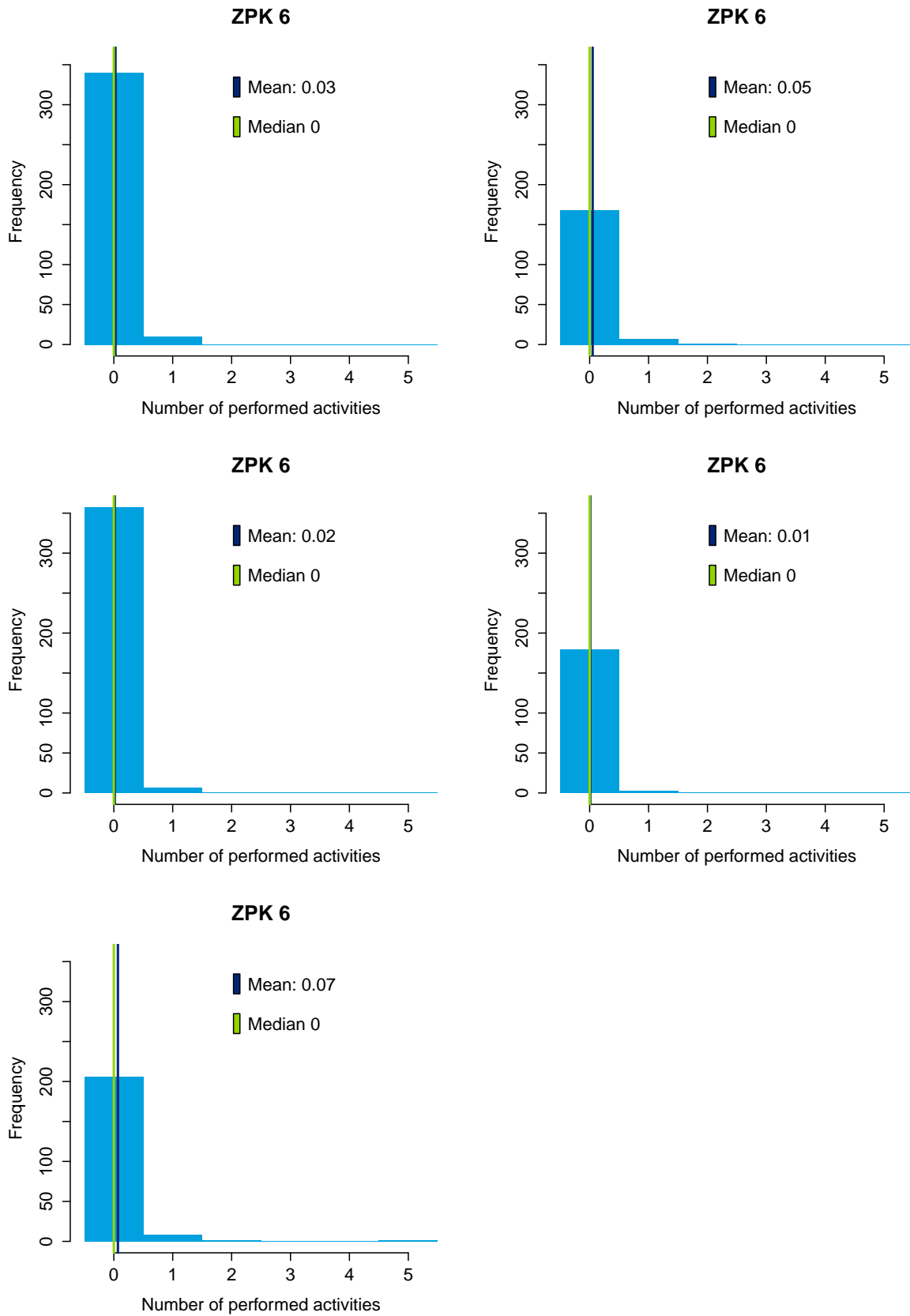


Figure F.7: Other therapeutic activities

F.2 Trace Alignments

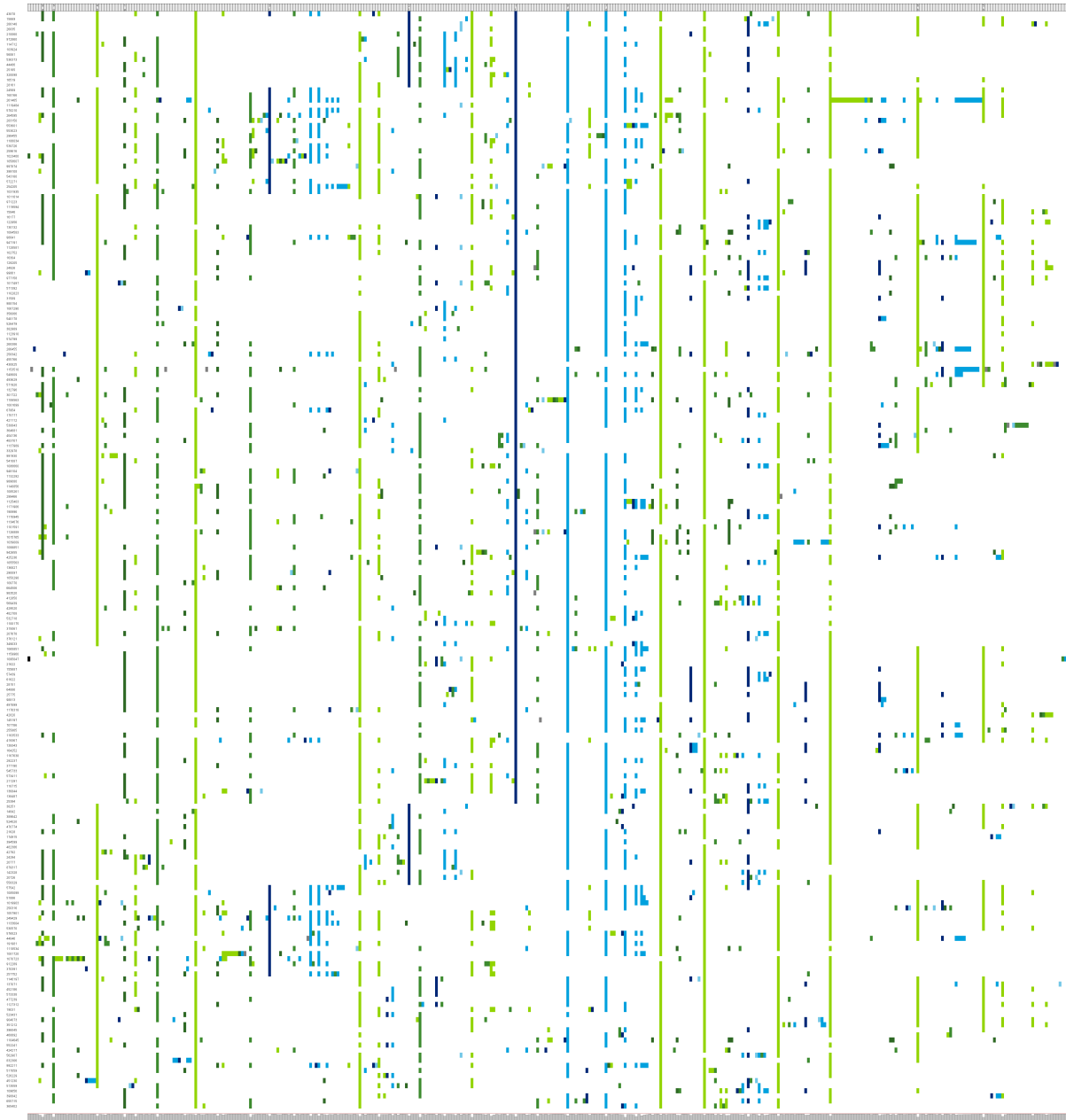
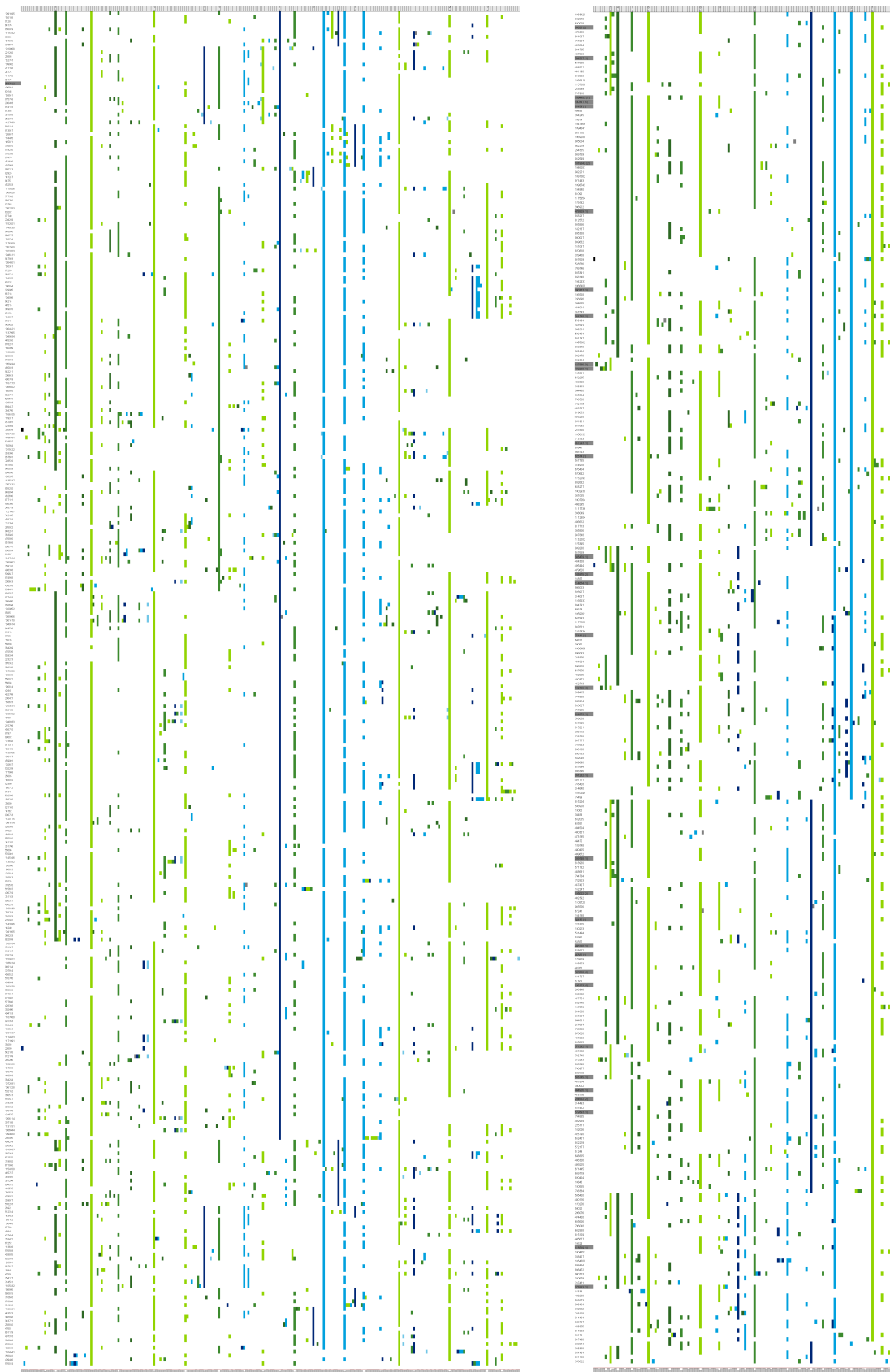


Figure F.8: Trace Alignment for cluster 5



(a) Cluster 1

(b) Cluster 3

Figure F.9: Trace Alignment for clusters 1 and 3

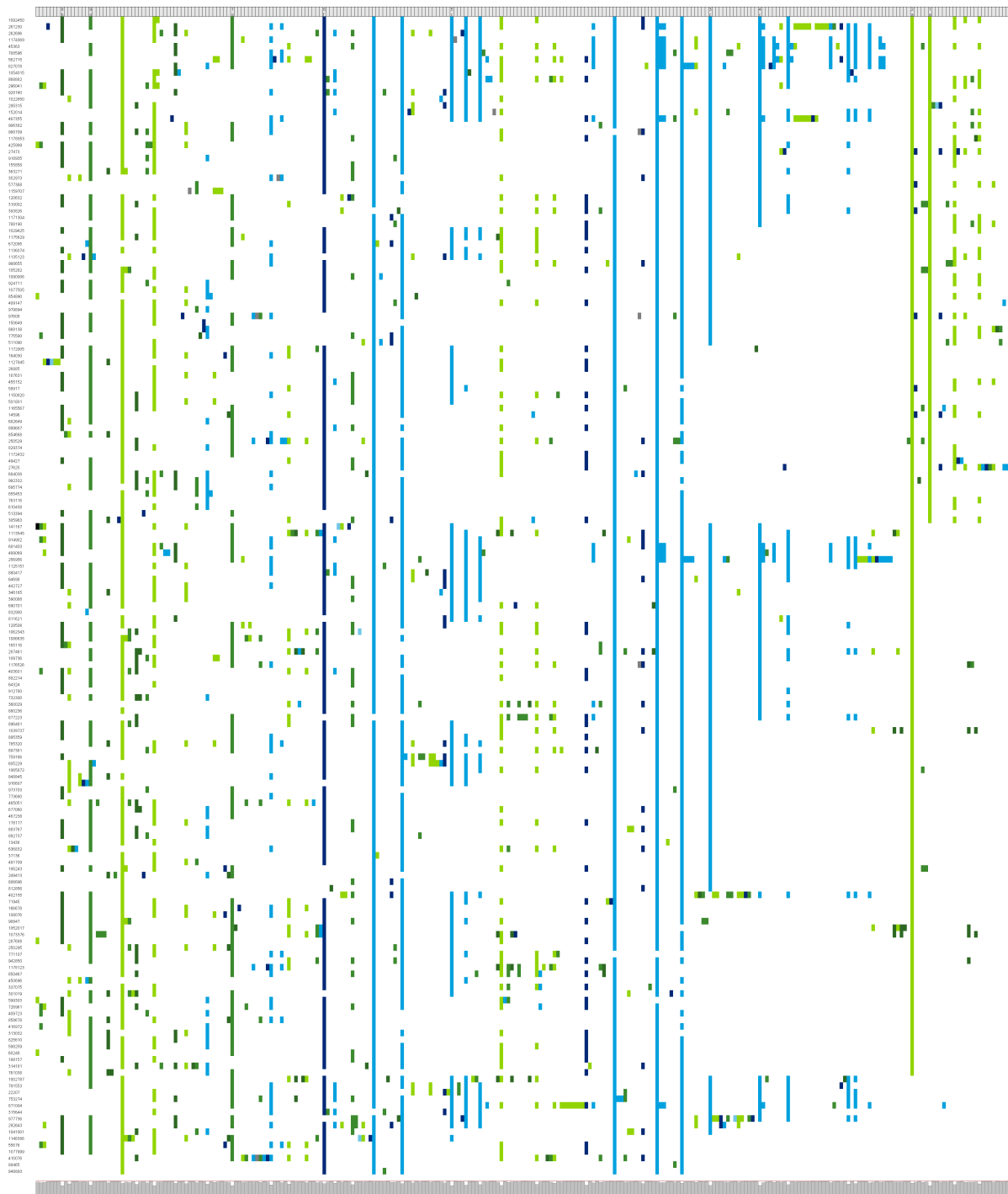


Figure F.10: Trace Alignment for cluster 2

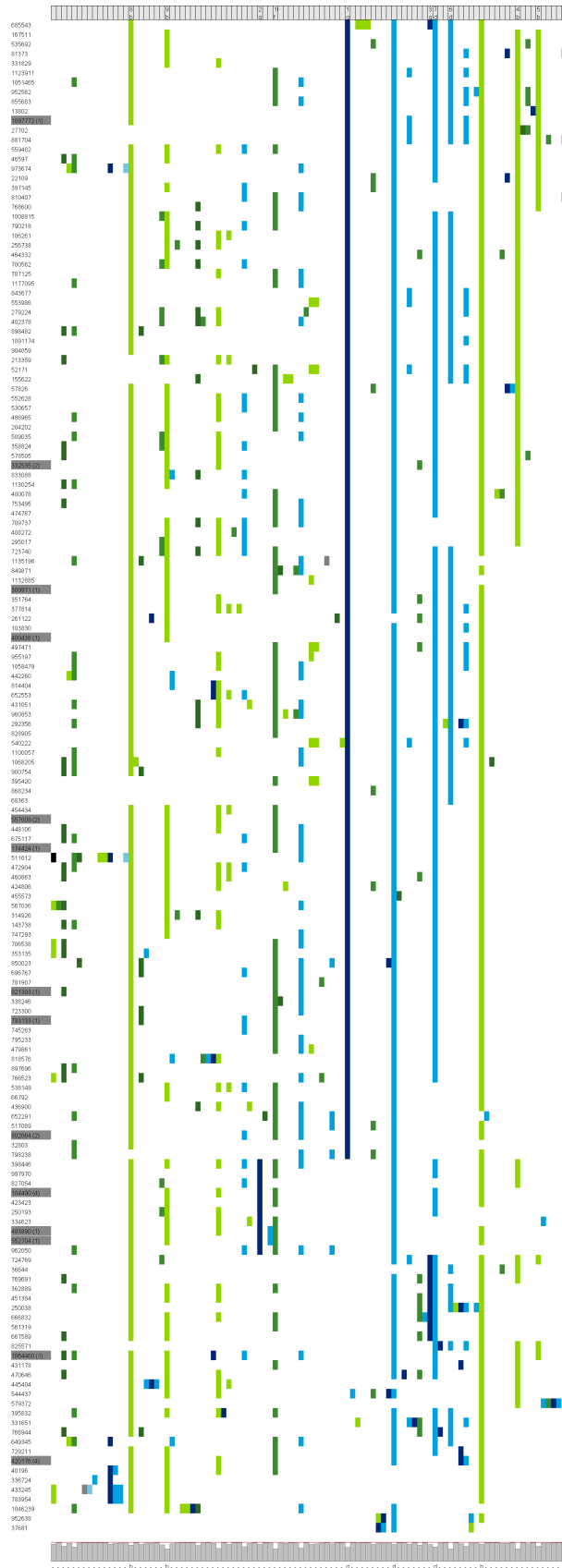


Figure F.11: Trace Alignment for cluster 4

Appendix G

ZPK-code overview

Below is a complete overview of the available ZPK-codes in the Dutch DBC and DOT systems.

Table G.1: ZPK code overview

ZPK	Definition	(Dutch)
1	Outpatient department / ER	Polikliniek- en eerste hulpbezoek
2	Daycare	Dagverpleging
3	Clinic	Kliniek
4	Diagnostic activities	Diagnostische activiteiten
5	Surgical activities	Operatieve verrichtingen
6	Other therapeutic activities	Overige therapeutische activiteiten
7	Medical imaging	Beeldvormende diagnostiek
8	Clinical chemistry and hematology	Klinische chemie en haematologie
9	Microbiology and parasitology	Microbiologie en parasitologie
10	Pathologie	Pathologie
11	Other laboratory operations	Overige laboratoriumverrichtingen
12	(Para)medical functions	(Para)Medische en ondersteunende functies
13	Prosthetic implants	Bijzondere kunst- en hulpmiddelen
14	Rehabilitation	Revalidatie
15	Blood products	Bloedprodukten
17	Long Asthma Centers	Longastmacentra
18	Other ER activities	IC zorgactiviteiten niet zijnde ic-behandeldag
19	ER treatment	IC-behandeldag
20	Expensive drugs/medicine	Dure geneesmiddelen
21	Other drugs/medicine	Weesgeneesmiddelen
22	Clotting factors	Stollingsfactoren
89	Other activities	Overige zorgactiviteiten t.b.v. afleiding
99	Not included in careprofile	Niet in profiel meegenomen



