

**MASTER**

**A data driven approach to evaluate guidelines for non-melanoma skin cancers (NMSCs)**

Kleinloog, T.C.P.

*Award date:*  
2015

[Link to publication](#)

**Disclaimer**

This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

**General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain

Eindhoven/Amsterdam, November 2015

# **A data driven approach to evaluate guidelines for non-melanoma skin can- cers (NMSCs)**

by  
T.C.P. (Tim) Kleinloog

BSc Industrial Engineering and Management Science – TU/e 2013  
student identity number 0676632

in partial fulfilment of the requirements for the degree

**Master of Science  
in Innovation Management**

**Supervisors:**

prof. dr. ir. U. (Uzay) Kaymak	(TU/e)
dr. P.M.E. (Pieter) van Gorp	(TU/e)
G.C. (Gioia) Convent	(ChipSoft BV)
H.M.A. (Jetje) Lindo	(ChipSoft BV)
dr. G.A.M. (Gertruud) Krekels	(Bravis hospital)
M. (Milan) Tjioe	(Bravis hospital)

TU/e. School of Industrial Engineering  
Series Master Theses Innovation management

Subject headings: process and compliance analysis, clinical guidelines, HIS/EPR, NMSC

# Abstract

The goal of data driven evaluation of guideline-based clinical processes, is to gain (patient centred) insights into operational clinical processes. These insights could be used to support the clinical process and optimize the guidelines studied. Due to the increasing digitalization of clinical processes and registrations, opportunities arise to improve healthcare by taking advantage of information technology. In many sectors, data driven decision-making and decision-support are a matter of course. For the healthcare sector on the other hand, this is not an obvious approach yet. In this thesis an approach is developed to evaluate clinical processes that are scientifically supported by clinical guidelines for non-melanoma skin cancers (NMSCs) (basal cell carcinoma (BCC) and (primary cutaneous) squamous cell carcinoma (SCC)). This approach is built on a technique developed to perform process and compliance analyses on BPMN 2.0 process models. This technique shows that it is theoretically possible to check the compliance between patient behavior in event logs (Mining eXtensible Markup Language (MXML)) and clinical guidelines in BPMN 2.0 (XML process definition language (XPDL)). However, there was still a lack of understanding whether real-life (complex) processes based on clinical guidelines can be actively supported by a compliance checking technique. Our approach explores the applicability of this technique and evaluates its results. It covers data extraction, data transformation and reveals some important data challenges that should be considered before it is possible to perform compliance analyses. In this thesis, the prior will be applied during a case study in a Dutch hospital <sup>1</sup>. In order to perform this case study, first a solid understanding of the hospital's datawarehouse was required. An internship at ChipSoft <sup>2</sup> supported this study with the needed knowledge of the datawarehouse structure that is used by many hospitals in the Netherlands. With this, it was possible to explore and extract the data available in an average hospital. Meanwhile the hospital's dermatology and pathology departments provided the required resources to perform a case study. During the application of this approach, data preparation proved to be a challenging and time consuming task. This is mainly due to fact that the data is originally registered for different purposes. Nevertheless, it was possible to perform a compliance analyses on process models that reflected the original paper based clinical guidelines. Finally the process diagnostics derived from the compliance analysis are evaluated and concluding remarks are provided.

---

<sup>1</sup>The Dutch hospital that took part in our case study: Bravis hospital <https://www.bravisziekenhuis.nl/>

<sup>2</sup>ChipSoft is one of the biggest supplier for health information systems and electronic patient records in the Netherlands, for more information: <http://chipsoft.nl/>





# Executive summary

## Problem definition

Decision support is receiving more and more attention in healthcare institutions. Keeping healthcare affordable and at the same time respond to the demand for patient centred care is a big challenge for hospitals. They are expected to deliver responsible high quality care given on an effective, efficient and patient-centred way that is adjusted to the realistic needs of the patient. Clinical guidelines support clinicians in how they are expected to cope with complex (patient specific) choices within clinical processes. These guidelines could streamline treatment processes, although clinicians still have to individually estimate patient risks. The information systems group of the Industrial Engineering and Innovation Science department of TU/e is investigating a technique to evaluate models in the business process model and notation (BPMN) language against event logs in a healthcare environment. Today, it is still uncertain whether real-life (complex) processes based on clinical guidelines can be actively supported by such an approach. This forms the problem definition of this research project.

## Research objectives and questions

Literature shows that the application of decision support techniques during treatment choices are very limited. For such techniques it is often unknown whether clinical registrations are complete and unambiguous enough to be used. The main research objective of this thesis is:

*Develop and test an approach in order to evaluate patients' adherence to clinical guidelines for NMSCs*

The techniques used in this thesis requires data to be of certain formats. To be able to perform the compliance analysis, this research is dependent on the data. Therefore it is studied where shortcomings in the data arise, how to deal with those shortcomings and what their influence will be on the eventual analysis.

### Research questions

In this study the following main research question is proposed:

*How can an approach be developed in order to be able to evaluate patients' adherence to clinical guidelines for NMSCs?*

This main research question is divided into the following sub-questions:

1. *How can paper-based clinical guidelines for NMSCs be translated into BPMN process models, in order to make them easy to interpret and suitable for compliance analysis?*
2. *Is the data that is currently registered electronically in an hospital suitable for compliance analysis?*

- 
3. *How to interpret and evaluate the process diagnostics that follow from the compliance analysis?*

## Methodology

To satisfy the research objective and answer the research questions, the following steps are taken:

1. Conduct a literature study on process modelling, process compliance and healthcare (clinical guidelines, dermatology, NMSCs).
2. Specify clinical guidelines in terms of BPMN 2.0 process models.
3. Study, extract and transform patient- and treatment process log data from the datawarehouse behind the HIS and EPR.
4. Describe the data challenges to be able to perform a compliance analysis.
5. Perform a compliance analysis and evaluate the results.

## Clinical case study & Results

In this study, an approach to evaluate clinical guidelines with historical real-life clinical data that was available at that time was explored. To be able to evaluate a clinical process with data, first solid domain knowledge had to be gained in order to understand the processes, guidelines and data. This helped to model clinical guidelines and transform data while taken into account their original meaning in a specific context. This study was performed during a case study in a Dutch hospital while being in close contact with the supplier for the hospital's hospital information system (HIS) and electronic patient record (EPR) (ChipSoft). When taking the investigated process as an example for an average guideline based process in a Dutch hospital, it could provide information about the feasibility of a broader application in the (near) future. The investigated guidelines are modeled and adjusted for the activities in the data. In the analysis, it was not chosen to aggregate the activities in the models for the BCC process. For the SCC process this could not be avoided, because the models became too complex to be usable in the software tool to check compliance. In general, our vision is to limit the use of aggregation with limited clinical knowledge. Interpretation of every individual clinical activity is needed in order to aggregate them. The results show compliance metrics of the execution of activities in the data-adjusted knowledge-based process models that represent the clinical guidelines for the two diagnoses (BCC and SCC). Algorithm 1 optimizes the routing of real-life patient behavior (captured in an event log) through the process models in figures A.7, A.8, A.12 and A.13. The tables 6.1, 6.2 (for the BCC patient population), 6.4 and 6.3 (for the SCC patient population) give information about the patients' processes in comparison to the guidelines. During a synchronous or compliant move, the patient behavior (as registered in the event log) is allowed by the guidelines (as registered in the BPMN process model) and expected by the algorithm to happen. When the algorithm identifies a move on model, an activity from the calculated path through the process model is missing in the log. If we show the results in percentages, For the BCC case: 42% of all events that are expected to happen according to the guidelines are skipped during diagnosis and 96% of all events that are expected to happen according to the guidelines are skipped during treatment/follow-up. For the SCC case: 47% of all events that are expected to happen according to the guidelines are skipped during diagnosis and 44% of all events that are expected to happen according to the guidelines are skipped during treatment/follow-up. Due to the issues in the data and immaturity of the software tool to check compliance in BPMN process models, these results should be interpreted carefully.

## Conclusion & Recommendations

The more clinical data is analyzed, the more it will become possible to support clinicians in their daily practice. The approach developed in this study shows there is much potential in the use of

---

information technology to evaluate guideline based clinical processes. Unfortunately the results derived are not reliable to be used in practice. This is due to current issues in the dataset that make it necessary to examine every case manually, and immaturity of the used software tool. The goal of this study was to “develop and test an approach in order to evaluate patients’ adherence to clinical guidelines for NMSCs”. At the start of this study it was uncertain whether the available data would enable us to present usable compliance results that could be used in practice. Of course, providing the clinical professionals with usable results has been a motivation during this study. After the case study was completed, it could be concluded that the method was applicable in practice. Unfortunately the data currently available combined with immaturity of compliance checking software was not able to give reliable results for the complex situation investigated. Future research should focus on using this technique on less complicated processes and try to relate real behavior to treatment outcomes. In this study it was not valuable to relate questionable compliance results with treatment outcomes. It is important to keep applying data and process mining techniques in a real-life healthcare environment. Benefits that arise include: medical professionals that become more aware of the possibilities of data. On the other side could researchers that develop new algorithms for healthcare applications experience its complexity.



# Preface

This master thesis is the result of my graduation project which completes my Master of Science degree in Innovation Management at the Eindhoven University of Technology. The project was performed during an internship at ChipSoft BV. A substantial part of this project has been realized during a case study at the Bravis hospital in Bergen op Zoom and Roosendaal. During this master thesis project I was very lucky that I was able to work with inspiring people and institutions. This did not only result in a valuable learning experience, but also opened doors that normally remain closed for most people.

The master phase of my study has been dominated by this thesis and therefore I wanted it to be a success in which I was able to apply the knowledge I gained throughout my study.

During the first year of my master degree, I had the opportunity to choose Uzay Kaymak as my first supervisor. Uzay is professor of information systems in healthcare at the Information Systems research group in the School of Industrial Engineering (TU/e). I would like to thank Uzay Kaymak for his support during this study. His challenging questions pushed me to think carefully about decisions that I made.

At the beginning of my second year, I contacted ChipSoft for a position as research intern and got in touch with Gioia Convent. Gioia is team manager of ChipSoft's datawarehouse department. Gioia gave me the opportunity to perform my study within the company. I would like to thank Gioia Convent for her confidence in me. I was able to work independent, access all required resources, and she was always there when necessary. On a daily basis, I am very thankful to Jetje Lindo. Jetje was a software consultant for ChipSoft and always made time for questions during my internship. Just like all the other colleagues at the department, she brought humour and joy during the construction of this thesis. ChipSoft adopted me as a real employee.

Furthermore, I would like to thank the clinical professionals Gertruud Krekels and Milan Tjioe for their support. They were essential in performing the clinical case study in this thesis. Because of their belief in the utility of this study, they gave us access to all required resources from the hospital. They put a lot of effort in making this study a success. Without them it would not have been possible to perform this study with real-life patient data.

Additionally, I am thankful to Pieter van Gorp (my second supervisor) and Hui Yan (researcher of important previous work) for their contributions. They could always answer my questions timely.

Last but certainly not least, I would like to thank my family, girlfriend and friends for always being there for me. Although it has been a busy time in which I had to go to Amsterdam on a daily basis, they kept supporting me. I cherish the valuable moments in which you made me think about something else for a change.

Tim Kleinloog  
November, 2015

# List of Abbreviations

7PMG	seven process modeling guidelines.
BCC	basal cell carcinoma.
BPMN	business process model and notation.
CPM	critical path method.
CRISP-DM	Cross Industry Standard Process for Data Mining.
CSV	comma-separated values.
EBM	evidence based medicine.
EPR	electronic patient record.
FTE	full time equivalent.
HIS	hospital information system.
IOM	Institute of Medicine.
IT	information technology.
MXML	Mining eXtensible Markup Language.
NMSC	non-melanoma skin cancer.
PERT	program evaluation and review technique.
RCT	randomized controlled trail.
RQ	research question.
SCC	(primary cutaneous) squamous cell carcinoma.
TAM	technology acceptance model.
TU/e	University of Technology Eindhoven.
UV	ultraviolet.
WHO	World Health Organization.
XES	eXtensible Event Stream.
XPDL	XML process definition language.

# Contents

Contents	x
List of Figures	xiii
List of Tables	xv
<b>1 Introduction</b>	<b>1</b>
1.1 Problem description	1
1.2 Research objective and questions	2
1.3 Research scope	2
1.4 Research methodology	3
1.5 Outline	3
<b>2 Preliminaries</b>	<b>5</b>
2.1 Business process compliance analysis	5
2.1.1 Business process modeling in BPMN 2.0	5
2.1.2 Event logs	6
2.1.3 Alignment	8
2.1.4 Deviation detection in BPMN 2.0	9
2.2 Clinical guidelines & -pathways	13
2.3 An approach to evaluate clinical guidelines for NMSCs	14
2.3.1 Programming in R	14
<b>3 Clinical guidelines for NMSCs</b>	<b>15</b>
3.1 Non-melanoma skin cancer	15
3.2 Modeling clinical guidelines	16
3.2.1 Clinical guidelines for basal cell carcinoma	16
3.2.2 Clinical guidelines for squamous cell carcinoma	19
3.2.3 Process model design choices	25
<b>4 Case study</b>	<b>27</b>
4.1 Bravis hospital	27
4.2 Data sources	28
4.2.1 PALGA	28
4.2.2 ChipSoft: HiX - Datawarehouse	28
4.3 Data extraction	29
4.4 Data pre-processing	29
4.4.1 Log preparation	30
4.4.2 Dataset preparation	30
4.5 Patient population in datasets	31



<b>5</b>	<b>Challenges during data collection</b>	<b>37</b>
5.1	Log and Data preparation . . . . .	38
5.1.1	Discovering recurrent patients within patient populations . . . . .	40
5.2	Aligning events in the modeled guidelines with events in the log . . . . .	45
5.2.1	Filtering events in the log . . . . .	46
5.2.2	Activities in the guidelines vs. activities in the log . . . . .	47
5.2.3	Data adjusted process models . . . . .	56
<b>6</b>	<b>Case study results &amp; Evaluation</b>	<b>57</b>
6.1	Compliance analysis . . . . .	57
6.1.1	Compliance analysis results BCC . . . . .	58
6.1.2	Compliance analysis results SCC . . . . .	61
6.2	Evaluation of compliance results . . . . .	63
6.3	Evaluation of the approach to evaluate clinical guidelines for NMSCs . . . . .	65
6.4	Evaluation of BPMN compliance analysis technology . . . . .	67
<b>7</b>	<b>Conclusion &amp; Recommendations</b>	<b>69</b>
	<b>Bibliography</b>	<b>71</b>
	<b>Appendix</b>	<b>75</b>
<b>A</b>	<b>BPMN 2.0 process models</b>	<b>75</b>
A.1	Initial process models . . . . .	75
A.2	Discovered process models . . . . .	78
A.3	Data adjusted process models . . . . .	83
<b>B</b>	<b>Influential patient characteristics for NMSCs</b>	<b>91</b>
<b>C</b>	<b>Data mining</b>	<b>92</b>
C.1	Data modeling in R . . . . .	93
C.1.1	Predictive model . . . . .	94
<b>D</b>	<b>SQL query HiX-datawarehouse</b>	<b>96</b>
<b>E</b>	<b>R script to aggregate activities for SCC</b>	<b>98</b>

# List of Figures

2.1	Real- vs. modeled behavior [13]	6
2.2	A simple state space	10
2.3	From state space to action space	11
2.4	Software tool for deviation detection in BPMN 2.0 (user interface)	12
2.5	Approach to evaluate clinical guidelines	14
3.1	Disease management system for chronic skin cancer [22]	16
3.2	Process model for overall BCC therapy process (high abstraction level)	17
3.3	Biopsy process for BCC	17
3.4	Treatment process for BCC	18
3.5	Follow-up process for BCC	18
3.6	Process model for overall SCC therapy process (high abstraction level)	19
3.7	Initial physical examination process for SCC	20
3.8	Physical examination for patients suspected of SCC	20
3.9	Biopsy process for SCC	20
3.10	Additional diagnostics process for SCC	22
3.11	Treatment process for SCC	24
3.12	Follow-up process for SCC	24
4.1	Histogram for the pathological cases diagnoses with BCC as primary diagnosis	31
4.2	Histogram for the age of patients diagnosed with BCC	32
4.3	Histogram for the gender of patients diagnosed with BCC	33
4.4	Histogram for the pathological cases diagnoses with SCC as primary diagnosis	34
4.5	Histogram for the age of patients diagnosed with SCC	35
4.6	Histogram for the gender of patients diagnosed with SCC	36
5.1	Recurrent patients (BCC)	44
5.2	Recurrent patients (SCC)	44
5.3	Detailed discovered process using process discovery for BCC case	45
5.4	Detailed discovered process using process discovery for SCC case	45
5.5	Simplified discovered process using process discovery BCC case	47
5.6	Simplified discovered process using process discovery SCC case	48
6.1	Compliance tool output	58
A.1	Initial process model for diagnostics BCC	76
A.2	Initial process model for diagnostics BCC	77
A.3	Discovered process model for diagnostics BCC (aggregated activities)	79
A.4	Discovered process model for diagnostics BCC (filter 50%, aggregated activities)	80
A.5	Discovered process model for diagnostics SCC (aggregated activities)	81
A.6	Discovered process model for diagnostics SCC (filter 50%, aggregated activities)	82
A.7	Data adjusted process model for diagnostics BCC	84

*LIST OF FIGURES*

---

A.8	Data adjusted process model for treatment and follow-up BCC . . . . .	85
A.9	Data adjusted process model for diagnostics SCC . . . . .	86
A.10	Data adjusted process model for additional diagnostics SCC . . . . .	87
A.11	Data adjusted process model for treatment and follow-up SCC . . . . .	88
A.12	Data adjusted process model for diagnostics (incl. additional diagnostics and aggregated events) SCC . . . . .	89
A.13	Data adjusted process model for diagnostics treatment and follow-up (excl. additional diagnostics and aggregated events) SCC . . . . .	90

# List of Tables

2.1	BPMN 2.0 notation . . . . .	7
3.1	TNM classification . . . . .	21
3.2	Treatment options for SCC . . . . .	23
3.3	Seven process modeling guidelines . . . . .	26
4.1	Key figures Bravis hospital, from 31/12/2014 - 24/9/2015 . . . . .	28
4.2	Columns extracted from the HiX - datawarehouse . . . . .	29
4.3	Most important columns in PALGA extraction . . . . .	29
4.4	Log structure . . . . .	30
4.5	Dataset structure . . . . .	31
5.1	Spectrum of clinical data sources . . . . .	37
5.2	Initial dataset - PALGA extraction . . . . .	38
5.3	Dataset - PALGA extraction (after transformation) . . . . .	39
5.4	Initial dataset - HiX extraction . . . . .	40
5.5	Dataset - HiX extraction (after transformation) . . . . .	41
5.6	Terms used to registers tumor locations . . . . .	41
5.7	Events in model vs. events in log for guidelines BCC (Diagnostics) . . . . .	49
5.8	Events model vs. events in log for guidelines BCC (Treatment) . . . . .	50
5.9	Events in model vs. events in log for guidelines BCC (Follow-up) . . . . .	51
5.10	Events in model vs. events in log for guidelines SCC (Diagnostics) . . . . .	52
5.11	Events in model vs. events in log for guidelines SCC (Treatment) . . . . .	55
5.12	Events in model vs. events in log for guidelines SCC (Follow-up) . . . . .	56
6.1	Results of activities in the model for guidelines BCC (diagnostics) . . . . .	59
6.2	Results of activities in the model for guidelines BCC (Treatment + follow-up) . . . . .	60
6.3	Results of activities in the model for guidelines SCC (excl. treatment) . . . . .	62
6.4	Results of activities in the model for guidelines SCC (excl. additional diagnostics) . . . . .	63
B.1	Important patient variables for SCC (according to guidelines) . . . . .	91



# Chapter 1

## Introduction

Today's world is constantly changing. One of the major innovations healthcare is facing today, is digitization of their processes [18]. Additionally, these innovations create new opportunities on itself. Opportunities to look at alternatives from the traditional ways things are done. Traditional ways patients are treated for example (e.g., e-health and data driven decision support).

“Innovations always sound good in retrospect, after theyve worked, and in isolation, when all the surrounding barriers to change dont have to be taken into account. Arguably, the main roadblock to innovation in health care is not the limits of human imagination and creativity; it is how a complex system has grown up in which most players have incentives for keeping their piece intact while hoping to seize a piece from someone else” [28].

Innovation management focuses on how to analyze, design and manage new product processes in technology-driven firms.

### 1.1 Problem description

Keeping healthcare affordable and at the same time respond to the demand for patient centred care is a big challenge for healthcare institutions nowadays. Standardizing clinical processes within hospitals seems a way to reduce costs (e.g., by clinical pathways). However, designing a (prescriptive) standardized clinical process for all patients is often not possible. Healthcare institutions in the Netherlands are even obligated by law <sup>1</sup> to deliver responsible high quality care given on an effective, efficient and patient-centred way that is adjusted to the realistic needs of the patient.

Clinical guidelines support clinicians in how they are expected to cope with complex (patient specific) choices within clinical processes. These guidelines are based on evidence based medicine (EBM) [27]. Clinical guidelines require consensus among medical experts with domain specific knowledge. The guidelines are there to support clinicians at interpreting existing evidence, but this does not mean that they do not have to individually estimate patients risks [37]. What stands out in the previous sentences are: ”consensus” and ”estimate patients’ risks”. This indicates that decisions are often not based on unambiguous evidence.

The use of decision support by clinicians in order to cope with these uncertainties is very limited. Another difficulty concerns registrations. Reliable decisions support systems require registrations during a treatment process to be complete, reliable and unambiguous. Unfortunately this is not always observable in current practice. Information technology (IT) seems to have the potential to contribute to challenges healthcare institutions are facing [37]. Complex healthcare processes do need optimal process support that requires cooperation of different organizational units and medical disciplines. Clinicians argue that clinical information systems in hospitals internally seem to have limited effect due little changes in efficiency and security [53]. Especially slow and difficult systems seem to be the problem here.

---

<sup>1</sup>Artikel 2: “Kwaliteitswet zorginstellingen” (Quality law healthcare institutions)

The information systems group of the Industrial Engineering and Innovation Science department of University of Technology Eindhoven (TU/e) is investigating techniques to evaluate process models in BPMN 2.0 against event logs in a healthcare environment. Today, it is still uncertain whether complex abstract processes based on clinical guidelines can be actively supported by such an approach. This forms the problem definition of this thesis study.

## 1.2 Research objective and questions

This thesis focusses on the fact that data driven decision support during treatment choices is rarely used in practice. For this purpose a technique is used that is recently developed at the TU/e and is able to check compliance between a process model and event log. The problem considered arises from a lack of understanding whether in practice, clinical registrations are complete and unambiguous enough to be useful in such a technique. Therefore, the main research objective of this thesis is:

*Develop and test an approach in order to evaluate patients' adherence to clinical guidelines for NMSCs*

The techniques used in this study requires data to be in certain formats. To be able to perform the compliance analysis, this research is dependent on the data. Therefore it is studied where shortcoming in the data arise, how we will be able to deal with those shortcomings and what their influence will be on the eventual analysis.

This study will consist of three main parts. The first part will focus on the study of the processes that are based on the guidelines for NMSCs, the second part will focus on the data that is available in the different data sources and the third part will focus on the compliance analysis and its results.

In this study the following main research question is proposed:

*How can an approach be developed in order to be able to evaluate patients' adherence to clinical guidelines for NMSCs?*

This main research question is divided into the following sub-questions:

1. *How can paper-based clinical guidelines for NMSCs be translated into BPMN process models, in order to make them easy to interpret and suitable for compliance analysis?*
2. *Is the data that is currently registered electronically in an hospital suitable for compliance analysis?*
3. *How to interpret and evaluate the process diagnostics that follow from the compliance analysis?*

## 1.3 Research scope

The goal of this study is to apply a theoretical compliance checking approach and study the challenges faced during this attempt. This approach strives to at least give useful process diagnostics about the events followed in the guidelines for NMSCs. As already stated in the introduction of this chapter, the focus will be on checking compliance of BPMN 2.0 process models. This study should complement to the development of the technique in [62]. Moreover only the guidelines for NMSCs are considered. The processes in this clinical context are expected to be challenging. In order to meet the goals in this thesis, data from different data sources is required. The data in this study originates from a public Dutch hospital: the Bravis hospital. To get a good understanding of the patients, it is attempted to discover when patients return in our data for the same disease. The main focus however will be on the steps that will lead to analyzing patient's compliance to guidelines for NMSCs.

## 1.4 Research methodology

To satisfy the research objective and answer the research questions, the following steps are taken:

1. Conduct a literature study on process modeling, process compliance and the healthcare domain (clinical guidelines, dermatology, skin cancer).
2. Specify clinical guidelines in terms of a process modeling language.
3. Study, extract and transform patient- and treatment process log data from the datawarehouse behind the HIS and EPR.
4. Describe the data challenges to be able to perform a compliance analysis.
5. Perform a compliance analysis and evaluate the results.

The development of this aforementioned approach is an iterative process. The available data will eventually determine in what way the approach will be applied.

## 1.5 Outline

The remainder of this thesis is organized as follows. In the next chapter (Chapter 2), the preliminary concepts that are used throughout this thesis are introduced. The different sections in this chapter contain literature overviews that cover the areas: business process modeling, clinical guidelines/pathways and compliance analysis of BPMN process models. Chapter 3 introduces NMSCs and shows how to translate the paper-based guidelines for the diseases in BPMN 2.0 process models. The case study is introduced in chapter 4. It contains more detailed information about the specific clinical context in which the approach is applied and how it can be applied to the methodology with the available resources. Because the process form data extraction until compliance analysis is not straightforward, chapter 5 is dedicated to the challenges faced during preparation of the data. Chapter 6 shows and evaluates the compliance results. The evaluation of the performed approach is given in Chapter 7. Finally, chapter 8 concludes this thesis and describes recommendations for future work.





# Chapter 2

## Preliminaries

This chapter describes the preliminary concepts used throughout this thesis and concludes with an approach to combine these preliminary concepts in order to be able to evaluate patients compliance to clinical guidelines. Section 2.1 provides an overview of business process compliance analysis as applied in this thesis. Section 2.2 provides an introduction to clinical guidelines and -pathways.

### 2.1 Business process compliance analysis

This study uses deviation detection between clinical guidelines in BPMN 2.0 and clinical event logs. The algorithm of this technique is introduced in [62] and uses the notion of alignment. In this section, the concepts that are used are discussed: BPMN 2.0 (subsection 2.1.1), (clinical) event logs (subsection 2.1.2), alignment (subsection 2.1.3), the algorithm to check compliance between process models in BPMN 2.0 and (clinical) event logs (subsection 2.1.4).

#### 2.1.1 Business process modeling in BPMN 2.0

A useful aid to explain business processes and the complexity of problems that it faces, is by using diagrams. Humans are very good at understanding diagrams accompanied by explanatory texts, which are called business process models when used to discuss business processes [17]. Process modeling makes a problem more at hand, but does not decrease the complexity of describing it. In this thesis, process models are used to give a graphical representation of a documented existing clinical process. The goal of business process modeling, is to “determine a representation of organizational processes with the goal of analyzing and studying various aspects of the real-world business process such as the activities, the products and the actors related to the process execution [34]”.

Process modeling is used in practice for several reasons [2]. First of all, visualization by using a process model increases the insight in the process. The visualization can be used to discuss with stakeholders. When process models are stored they can serve as documentation for instructions and certification purposes. Process models can also be actively used in analyses: as verification to find errors in systems or procedures (e.g., deadlocks) and as performance analysis technique in simulations. Moreover, process models are used in developing systems by giving feedback to designers, as agreement between end user and management, and eventually to configure a system.

The challenge for business process modeling is to align the modeled behavior as close to reality as possible. Models therefore aim to be descriptive without being normative [3]. Potential issues with process models are [2]: the inability to show all possible scenarios. Often, only the ideal situation of reality is shown in a possible wrong abstraction level. Moreover, most process models are not able to adequately capture human behavior. Therefore, much attention is currently pointed towards the field of process mining. Process mining uses discovery algorithms to model the process

from event data in logs. Because event logs capture the real behavior in a process, it seems a very reliable method to model real behavior in a process.

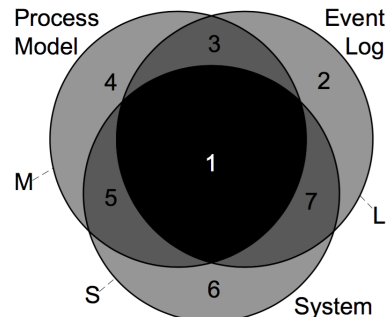
To be able to deal with the challenges above, the clinical guidelines that describe how clinical process are supposed to be executed according to consensus about clinical research (see section 2.2 for more information) are modeled. BPMN 2.0 [43] was chosen as process modeling language. BPMN 2.0 is a graphical representation for specifying business processes [43]. It is the standard notation for business process modeling [8] and has rich semantics to cover complex constructs. Only recently, BPMN 2.0 semantics are formalized to make them suitable for compliance analysis [24]. BPMN 2.0 is chosen as modeling language instead of for example Petri Nets [45], because it is easier to interpret (in general) and BPMN is seen as the industry standard for process modeling. This has as main advantage, that anyone who knows the language, is capable of understanding the process without further explanation.

The BPMN language uses explicit gateways to model control-flow logic [2]. These gateways are represented as diamonds. A diamond with an “X” sign represents a XOR-split or -join. A XOR-split allows a choice between subsequent activities. A XOR-join on the other hand only needs input from one prior activity. A diamond with a “+” sign represent an AND-split or join. An AND-split forces all subsequent activities to be activated. An AND-join on the other hand needs all prior activities as input. The sequence flow logic can be explained by “tokens”. Tokens are moved in a process model according to the BPMN rules for the model. When a process starts, the start event creates a token. As a process evolves, token movements occur independent from each other. When a token arrives at an activity, this activity receives the token and performs the task. After the task is completed, the token is released to the outgoing sequence flow. The modeler of a BPMN process has to define the conditions in a way that always exactly one of the condition is true [8]. BPMN does not prescribe how conditions are defined or checked. Especially in more complex process models, a token may travel several times through loops and decision gateways. The process finishes when the token arrives at the end event. Table 2.1 gives a summary of frequently used symbols in the BPMN 2.0 modeling language.

### 2.1.2 Event logs

















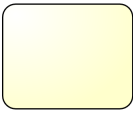

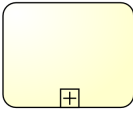



Software logs that contain information about activities of a process in a system are called event logs (also known as transaction logs or audit trail). Event logs do at least contain data about the start and/or the completion of an activity, a time-stamp and the performer [3]. It is important to notice that event logs are captured over a fixed time period. The events within the log are recordings of real behavior, but do not capture all possibilities of real behavior in a process (figure 2.1).

Figure 2.1: Real- vs. modeled behavior [13]: An aim of process modeling is to capture as much real behavior within a process as possible. Event logs do not record all possible occurrences of real behavior within a certain process. The Venn diagram in this figure shows seven areas. These areas can be described as following: (1) *Modeled and observed system behavior*:  $L \cap M \cap S$  (2) *Unmodeled exceptions*:  $(L/M)/S$  (3) *Modeled and observed exceptions*:  $(L \cap M)/S$  (4) *Modeled but unobserved and non-system behavior*:  $(M/S)/L$  (5) *Modeled but unobserved system behavior*:  $(M \cap S)/L$  (6) *Unmodeled and unobserved system behavior*:  $(S/L)/M$  (7) *Unmodeled but observed system behavior*:  $(S \cap L)/M$




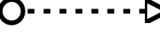

To understand event logs in a clinical setting, a definition based on [31, 32, 33] is used. A clinical event log contains clinical events ( $e$ ). Clinical events are assumed to have three properties: a patient identifier ( $pid$ ), an activity type ( $a$ ) and a time-stamp ( $t$ ) of the clinical event. Let  $PID$  be the patient identifier domain,  $A$ , a set of clinical activities and  $T$ , the time domain. A clinical

Table 2.1: The BPMN 2.0 notation consists of events, activities, gateways and flow objects to indicate relations. Sometimes also data objects, artefacts and swimlanes are used. Start events start or trigger the flow of a process. An intermediate event affects the flow of the process (e.g., by showing where messages are expected). End events end the (sequence) flow of a process and have a specific result (e.g., a message). Activities can be looked at as collapsed sub-processes or single tasks that are performed by an end-user and/or application. Gateways are used to model control-flow logic; next to parallel and exclusive gateways, inclusive and event-based gateways can model more complex control-flows. Flow object are used to connect the different parts of a process model. For a complete overview of the BPMN 2.0 modeling language we refer to OMG’s BPMN 2.0 manual <sup>a</sup>.

Events			
	start event		end event
	start message		intermediate throwing message
	intermediate catching message		end message
	intermediate catching error		end error
	intermediate catching escalation		intermediate throwing escalation
	intermediate catching signal		intermediate throwing signal
	intermediate catching link		intermediate throwing link
	intermediate timer		intermediate condition
Activities	Gateways		
		parallel gateway	
		data-based exclusive (XOR) gateway	
		inclusive gateway	
		event-based gateway	

<sup>a</sup><http://www.omg.org/spec/BPMN/2.0/PDF>

## BPMN 2.0 notation (continuation)

Relations	
	sequence flow
	message flow
	association

event  $e$  is represented as  $e = (pid, a, t)$ , where  $pid$  is the patient identifier of  $e$  ( $pid \in PID$ ),  $a$  is the activity type (e.g., admission) of event  $e$  ( $a \in A$ ), and  $t$  is the occurrence time of activity  $a$  ( $t \in T$ ). A clinical event is a clinical activity occurring at a particular time-stamp. A patient trace ( $\varepsilon_i$ ) consists of one or more clinical events and is represented as  $\varepsilon = \langle tid, \langle e_1, e_2, \dots, e_n \rangle \rangle$ , where  $tid$  identifies each patient trace and  $\langle e_1, e_2, \dots, e_n \rangle$  is a sequence of clinical events. A clinical event log ( $L$ ), contains a set of patient traces such that each event appears at most once in the entire log, i.e., for any  $\varepsilon_1, \varepsilon_2 \in L; \forall e_1 \in \varepsilon_1, \forall e_2 \in \varepsilon_2, e_1 \neq e_2$  or  $\varepsilon_1 = \varepsilon_2$ . Activity types ( $a$ ) (that are part of a clinical event) can happen more than once.

### 2.1.3 Alignment

Activity types  $a$  are used to explain the notion of alignment [4] between specified behavior (e.g., a process model) and observed behavior (e.g., an event log). Let  $A_L$  be a set of activities observed from an event log and  $A_M$  a set of activities specified in a model,  $a_L$  is an activity type  $a$  in an event log  $e$  ( $a_L \in A_L$ ),  $a_m$  is an activity type  $a$  specified by a process model ( $a_m \in A_M$ ). The idea is to find traces in the observer behavior that are as similar as possible to the modeled behavior. Differences between traces  $\sigma$  indicate deviations. Alignment will be illustrated in the following example. Suppose that  $\sigma_1 = \langle a, b, c, b, d \rangle$  is a sequence of activity types from  $A_L$  and  $\sigma_2 = \langle a, b, c, d \rangle$  is the corresponding sequence of activity types from  $A_M$ . Possible alignments  $\gamma$  between  $\sigma_1$  and  $\sigma_2$  are:

$$\begin{aligned} \gamma_1 &= \begin{array}{|c|c|c|c|c|} \hline a & b & c & b & d \\ \hline a & b & c & \perp & d \\ \hline \end{array} \\ \gamma_2 &= \begin{array}{|c|c|c|c|c|c|} \hline a & b & c & \perp & b & d \\ \hline a & b & c & d & \perp & \perp \\ \hline \end{array} \\ \gamma_3 &= \begin{array}{|c|c|c|c|c|} \hline a & b & c & b & d \\ \hline a & b & c & d & \perp \\ \hline \end{array} \end{aligned}$$

Each column is a pair  $(x, y)$  where  $x \in A_L \cup \perp$  and  $y \in A_M \cup \perp$ . When relating log type attributes  $a_L$  in the log trace  $\sigma_L$  to specification type attributes  $a_M$  in the model trace  $\sigma_M$ , different control-flow alignment moves are possible.

1. A synchronous move happens when  $x = y$
2. Non-synchronous moves can occur in different forms:
  - (a) a move on log ( $L$ ):  $x \in A_L$  and  $y = \perp$  (e.g.,  $(b, \perp)$  in  $\gamma_1$ )
  - (b) a move on model ( $M$ ):  $x = \perp$  and  $y \in A_M$  (e.g.,  $(\perp, d)$  in  $\gamma_2$ )
  - (c) a move on both:  $x \neq y$  (e.g.,  $(b, d)$  in  $\gamma_2$ )

The notion of alignment will return in the algorithm that will be introduced in subsection 2.1.4. This technique returns an optimal alignment for (patient) traces through a specified model such that no other alignment has fewer non-synchronous moves. In the field of process mining,

control-flow conformance checking techniques [4, 5, 46] use the notion of alignment in order to find optimal control-flow alignments in process models based on Petri Nets [45]. Often the main objective of alignment in process mining is to assess the goodness of a discovered process model (replay). For more information about the use of alignment in process mining we refer to [2].

### 2.1.4 Deviation detection in BPMN 2.0

There was not yet a backward compliance analysis technique that was able to check for individual non-compliance in BPMN 2.0. Recently, a first contribution is made to the development of a technique that is able to do a so called non-compliance check [62]. Its algorithm is shown below (algorithm 1). This new technique uses an adapted A\* algorithm [16] to reduce the exploration space in finding the best possible alignment. The A\* algorithm has been developed to find the shortest path between a source node and a target node. Two other properties that are considered in finding target alignments are: 1. the quality of ongoing alignments  $g(n)$ , which is the sum of all deviations and 2. the heuristic estimate of potential costs to become a target alignment  $h(n)$ . The average number of remaining traces in the log and action space are used as heuristic costs. From all candidates to explore, first the ones with least deviations are selected. From those, the candidates with the least heuristic costs are selected.

---

**Algorithm 1:** Best alignment search with A\* based algorithm

---

```

1: procedure ReplayTechnique ( $\sigma, G_{ac}, out : T_{tg}$ )  $\triangleright$ 
    $\sigma$ : a trace in log;  $G_{ac}$ : an action space from BPMN 2.0 process model;
    $T_{tg}$ : a set of final steps of alignment.
2:   //Initialize
3:   A set of ongoing steps:  $I = \phi$ 
4:   A set of selected steps for exploration:  $S_c = \phi$ 
5:   Create a new step  $n$  and add  $n$  into  $I$ 
6:   while  $I$  is not empty do
7:     Add the steps in  $I$  with lowest  $gScore$  and lowest  $hScore$  into  $S_c$ 
8:     for each  $n$  in  $I$  do
9:        $flag \leftarrow false$ 
10:      if  $n$  is a targeting alignment then
11:         $flag \leftarrow true$ 
12:        Add  $n$  to  $T_{tg}$ 
13:      end if
14:    end for
15:    if  $flag$  is true then
16:      break;  $\triangleright$  End while
17:    end if
18:    for each  $n$  in  $S$  do
19:       $e \leftarrow n$ .Entry's next entry in  $\sigma$ 
20:       $A_s \leftarrow n$ 's successive action set in  $G_{ac}$ 
21:      if  $e$ .Name  $\in A_s$ .Name than
22:        Create step  $n_1$  with  $n_1.gScore = n.gScore$ 
23:        Calculate  $hScore$  for  $n_1$ 
24:      else
25:        Create step  $n_2$  with  $n_2.gScore = n.gScore + 1$ 
26:        Calculate  $hScore$  for  $n_2$ 
27:      end if
28:      Add  $n_1$  and  $n_2$  into  $I$ 
29:    end for
30:  end while
31: end procedure

```

---

In order to get a complete understanding of the algorithm, a few definitions have to be clarified.

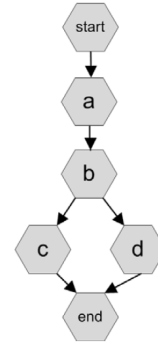
A BPMN 2.0 model is defined as a tuple  $(F_o, D_t, C_n, L_p, R_a)$ , where:

1.  $F_o$  is a set of flow objects, which are the main graphical elements to define the behavior of a business process.
  - $F_{ac}$  is a finite set of activities in BPMN, with  $F_{ac} \subseteq F_o$
  - $F_{ev}$  is a finite set of events in BPMN, with  $F_{ev} \subseteq F_o$
  - $F_{gw}$  is a finite set of gateways in BPMN, with  $F_{gw} \subseteq F_o$
2.  $D_t$  is a set of data, having data objects, data input, data output and data stores.
3.  $C_n$  is a set of connecting objects, which are used to connect the flow objects to each other or other information. There are four kind of connecting objects: sequence flows, message flows, associations and data associations.
4.  $L_p$  is a set of swimlanes, which are used to group the primary modeling element. They have pools and lanes.
5.  $R_a$  is a set of artefacts which are used to provide additional information about the process. Currently, there are two standardized artefacts: group and text annotation.

Given the BPMN 2.0 model, a state space is constructed. Figure 2.2 shows a simple state space. A generated state space can be defined as: a transition system  $P = (M, M_I, M_F, A_C, T)$  over a set of actions  $A_C$  with markings  $M$ , initial markings  $M_I \subset M$ , final markings  $M_F \subset M$ , and transitions  $T \subset M \times A_C \times M$ .

- $A_C$  is a tuple  $(O_f, T_{type}, M_{pre}, M_{post})$ , where  $O_f \subseteq F_o$  is a flow object;  $T_{type} = Start, Complete$  indicates the action type;  $M_{pre}$  is a set of markings before  $A_C$  and  $M_{post}$  is a set of markings after  $A_C$ .

Figure 2.2: A simple state space: this example state space shows that for any execution of a model  $M$ , first the events  $a$  and  $b$  occur subsequently. After which the events  $c$  and  $d$  can occur, but both should not happen together in one model execution. State spaces can be generated automatically for combinations of advanced constructs in BPMN 2.0, such as inclusive OR-joins. These advanced constructs are necessary to describe complex synchronization rules for clinical events.



After the state space has been generated, it can be transformed into a format whereby only actions and sequential relations between these actions are kept. An action space is a directed graph  $G = (V, E)$  where:

- $V$  is a finite set of action nodes, with  $V \subseteq A_C$
- $E$  is a finite set of ordered pairs of  $V$ , called arcs, directed edges, or arrows

An example for the construction of an action space from a state space is shown in figure 2.3

Deviation detection in BPMN 2.0 is applied in order to identify the paths that patients take through the complex synchronization rules for clinical events as specified by clinical guidelines. Additionally, deviation detection gives us information about patients' treatment adherence to clinical guidelines.

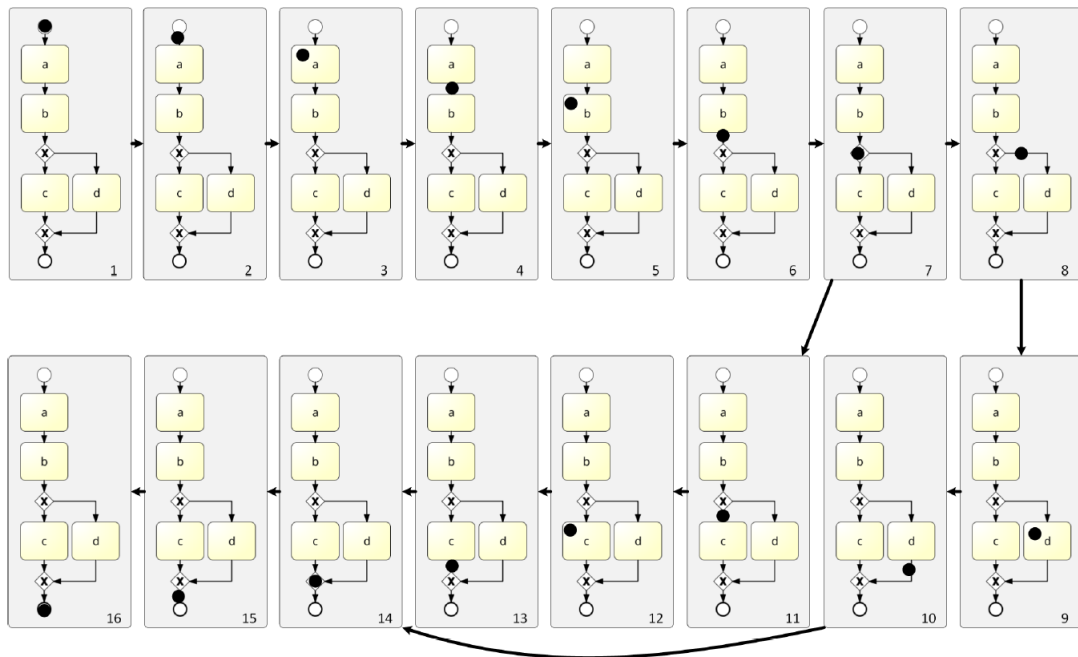


Figure 2.3: From state space to action space

The technique described above has been implemented in a software tool <sup>1</sup>. A screenshot of its user interface is shown in figure 2.4. As input, this software tool needs a process model and event log. The file format for the process model should be of the XPDL. A simple sequence pattern in XPDL for the activities “A” and “B” is shown below:

```

1 <WorkflowProcess Id="Sequence">
2     <processHeader DurationUnit="Y"/>
3     <Activities>
4         <Activity Id="A">
5             ...
6         </Activity>
7         <Activity Id="B">
8             ...
9         </Activity>
10    </Activities>
11    <Transitions>
12        <Transtion Id="AB" From="A" To="B"/>
13    </Transitions>
14 </WorkflowProcess>

```

An evaluation of all patterns in the XPDL language can be found in [1]. Software tools such as Signavio <sup>2</sup> and TIBCO <sup>3</sup> can be used to generate XPDL files from graphical process models.

<sup>1</sup><https://github.com/TUe-IS/BPMN20GrGen>

<sup>2</sup><http://academic.signavio.com>

<sup>3</sup><http://www.tibco.nl>



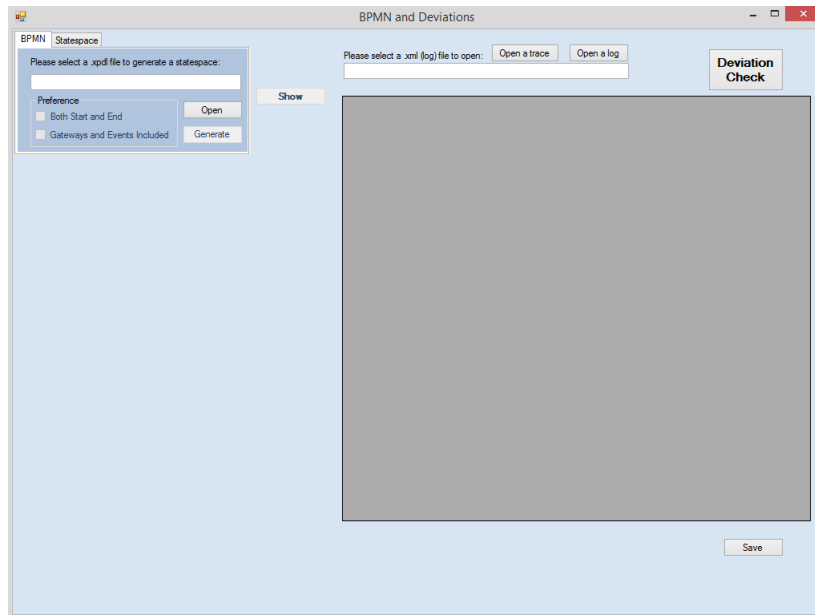


Figure 2.4: Software tool for deviation detection in BPMN 2.0 (user interface)

Data streams are stored in event logs (as shown earlier for a clinical environment). Event logs can have numerous formats. Two common formats to describe observed behavior in an event log are XPDL and eXtensible Event Stream (XES). XES is currently adopted by the IEEE Task Force on Process Mining as the standard format for logging events<sup>4</sup> and can be seen as the successor of MXML. Both formats enable researchers to benefit from each other’s ideas and implementations with little effort and improves applicability of for example process mining in business environments. Unfortunately only the XPDL format is currently supported by the aforementioned software tool. Therefore we will use the MXML language in this study. The technique in [62] as well as the ProM<sup>5</sup> framework both are able to use event logs in the MXML format.

An example of a trace in a log with the the process instances “A” and “B” in the MXML language is shown below:

```

1  <LogFile>
2      <ProcessInstance>
3          <AuditTrailEntry>
4              <WorkflowModelElement>A</WorkflowModelElement>
5              <EventType>start</EventType>
6              <Timestamp>yyyy-MM-ddThh:mm:ss</Timestamp>
7          </AuditTrailEntry>
8          <AuditTrainEntry>
9              <WorkflowModelElement>B</WorkflowModelElement>
10             <EventType>start</EventType>
11             <Timestamp>yyyy-MM-ddThh:mm:ss</Timestamp>
12         </AuditTrailEntry>
13     </ProcessInstance>
14 </LogFile>

```

<sup>4</sup><http://www.processmining.org/logs/start>

<sup>5</sup>ProM is an extensible framework that supports a wide variety of process mining techniques, <http://www.processmining.org>, <http://www.promtools.org>

## 2.2 Clinical guidelines & -pathways

Clinical guidelines support clinicians in how they are expected to cope with complex (patient specific) choices within clinical processes. These guidelines are based on EBM [27]. EBM aims to optimize decision-making by emphasizing evidence from medical research [19, 49]. The degree of empirical support is dependent on the epistemological strength. Only research from the strongest types (e.g., meta-analyses, randomized control trails and systematic reviews) are recommended by EBM. Until recently, medical decisions were highly subjective, often called “clinical judgement”. Although they were highly patient centred, the clinician merged evidence (if any) with personal beliefs. Currently, EBM is applied in programs that are designing clinical guidelines and programs that teach EBM [25]. Clinical guidelines require consensus among medical experts with domain specific knowledge. Moreover, clinical guidelines emphasize the use of evidence in clinical practices, which could lead to contradictions with experience-based practices. Therefore most effective clinical decision making still asks for subjective judgements, rather than just formal knowledge.

The development of clinical guidelines emerges from a need to make better informed healthcare choices that will lead to improved health outcomes and quality of care. The Institute of Medicine (IOM) specified this process in an interesting infographic (<http://resources.iom.edu/widgets/systematic-review/infographic.html>). The vast volume of evidence from clinical research makes it difficult for clinicians to stay abreast. Thoroughly developed guidelines from systematic reviews have the power to translate the complexity of scientific research into recommendations for clinical practice.

**Definition clinical guidelines** The IOM<sup>6</sup> defines clinical (practice) guidelines as: “statements that include recommendations intended to optimize patient care that are informed by systematic review of evidence and an assessment of the benefit and harms of alternative care options. Each guideline recommendation should include an explanation of the reasoning behind the recommendation, a rating of the level of confidence or certainty about the underlying evidence, and a rating of the strength of the clinical recommendation.”

Clinical pathways (also known as care pathways, critical pathways, care maps or integrated pathways) can be seen as the next step in optimized clinical decision support. They emerged from planning methods developed in industry (e.g., critical path method (CPM) and program evaluation and review technique (PERT)) [50]. These methods were used to improve planning in complex processes. Clinical guidelines are applied in clinical pathways. Next to emphasizing evidence, clinical pathways aim to improve process quality and resource usage. This requires consensus among healthcare professionals as well as cooperation. Clinical pathways can be seen as a healthcare application of process management thinking.

**Definition clinical pathways** The definition of clinical pathways developed over the years due to changes in clinical practice. Still the exact definition of clinical pathways is subject of discussion [57]. In [55] clinical pathways are defined as: “a complex intervention for the mutual decision making and organization of care processes for a well-defined group of patients during a well-defined period. Defining characteristics of care pathways include: 1. An explicit statement of the goals and key elements of care based on evidence, best practice, and patients expectations and their characteristics; 2. the facilitation of the communication among the team members and with patients and families; 3. the coordination of the care process by coordinating the roles and sequencing the activities of the multidisciplinary care team, patients and their relatives; 4. the documentation, monitoring, and evaluation of variances and outcomes; and 5. the identification of the appropriate resources. The aim of a care pathway is to enhance the quality of care across the continuum by improving risk-adjusted patient outcomes, promoting patient safety, increasing patient satisfaction, and optimizing the use of resources.” This definition is adopted by the European Pathway Association.

---

<sup>6</sup><http://www.iom.edu>

## 2.3 An approach to evaluate clinical guidelines for NMSCs

This chapter started with an explanation of business process compliance analysis (section 2.1). In order to perform a compliance analysis on clinical guidelines, first these guidelines have to be translated into process models for which BPMN 2.0 is used as explained in subsection 2.1.4. In order to model clinical guidelines it is necessary to understand the domain investigated (dermatology, NMSCs, see section 3.1).

After the domain/business is thoroughly understood, guidelines can be modeled. Meanwhile, data should be understood and extracted. The data should be used for two main purposes: 1. identifying and understanding the patient population diagnosed with NMSCs studied, 2. preparing event logs for the activities corresponding to the activities in the guidelines for NMSCs.

The preparation of event logs (as explained in subsection 2.1.2) is part of data preparation phase. The steps above have much in common with the Cross Industry Standard Process for Data Mining (CRISP-DM) [51] as commonly used for data mining problems. In appendix C more information can be found about CRISP-DM and data mining.

Eventually our approach is meant to be able to evaluate the behavior of patients in comparison to the guidelines for NMSCs built around the compliance algorithm explained in subsection 2.1.4.

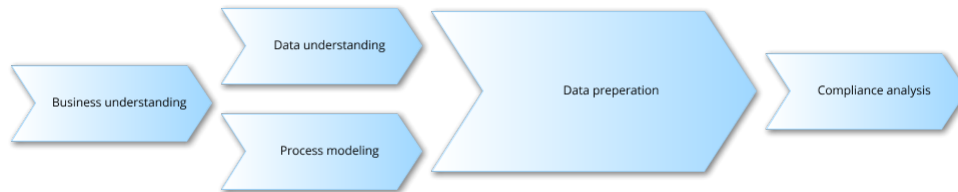


Figure 2.5: Approach to evaluate clinical guidelines: this is an abstract view of the approach applied in this thesis. The exact interpretation of the approach is depended on the clinical context it will be applied to.

### 2.3.1 Programming in R

R is used in this thesis to get a deep understanding of every step in the data understanding and preparation phases. R is a free software environment for statistical computing and graphics<sup>7</sup>. R is widely used by data analysts and academia around the world [54]. A strong capability of R is that it is extended by packages created by users. Two important packages that are used throughout this thesis are listed below:

- **dplyr**: is a consistent grammar for data manipulation.
- **ggplot2**: is a plotting system in R based on a grammar of graphics [60].

---

<sup>7</sup><http://caret.r-forge.r-project.org/>

## Chapter 3

# Clinical guidelines for NMSCs

This chapter starts with a section about NMSCs in general and why these diseases require increased attention (section 3.1). Section 3.2 describes how paper based clinical guidelines are translated into process models that make them suitable for compliance analysis. Section 3.2 is subdivided for the different parts of the guidelines.

### 3.1 Non-melanoma skin cancer

The worldwide incidence of NMSC has increased markedly during the last decades [23]. Increasing exposure to sun and the use of solariums (ultraviolet (UV) radiation) in combination with an aging population contribute to the significant increase in incidences of skin cancer [58]. NMSCs are carcinomas and represent the majority of skin cancers, comprising BCC and SCC. BCC is the most common type of skin cancer, but rarely lethal [15]. In 2015 more than 26 000 new cases of BCC are predicted in the Netherlands annually and this number is expected to keep rising [58]. SCC on the other hand can be fatal, especially for immune suppressed patients [61]. In 2015 more than 6000 new cases of SCC are predicted in the Netherlands annually. NMSCs are a burden for patients as well as for the whole healthcare system. For patients, operations are often disfiguring. For healthcare systems, the number of patients, tumors and therefore costs are enormous. In [23] is argued that NMSC should deserve increased attention from research, clinicians and politicians: “To manage the future costs and quality of care for skin cancer patients, a revised health strategy is needed. These strategies should be combined in a disease management system, a system that organizes health care for one well documented health care problem with a systematic approach, which includes prevention, education, multidisciplinary care, information technology and management.” Moreover, patients diagnosed with NMSC should be aware that it could be the start of a chronic (or noncommunicable) disease and is defined by the World Health Organization (WHO) <sup>1</sup> as a disease “of long duration and generally slow progression.” In this thesis the urge for improvements to the healthcare system for chronic NMSC is recognized. In figure 3.1 [23] a disease management system is proposed to manage this expanding healthcare problem. It recognizes the importance of increased attention to primary and secondary prevention, early detection and efficient and effective treatments.

This thesis focusses on evaluating the guidelines for BCC and SCC by investigating patients’ treatment process data. Clinical guidelines have been developed for NMSC treatment to assure the care quality and improve patient safety. On the other hand, today’s HISs contain a wealth of data. Moreover, the volume of healthcare related data is growing with the increasing use of EPRs and HISs. The treatment process data is used to construct patients’ paths through a healthcare process that will be compared to the Dutch National Guidelines for both diseases <sup>2 3</sup>.

---

<sup>1</sup><http://www.who.int>

<sup>2</sup><http://www.huidziekten.nl/richtlijnen/richtlijn-basaalcelcarcinoom-2014.pdf>

<sup>3</sup><http://www.huidziekten.nl/richtlijnen/richtlijn-plaveiselcelcarcinoom-2010.pdf>

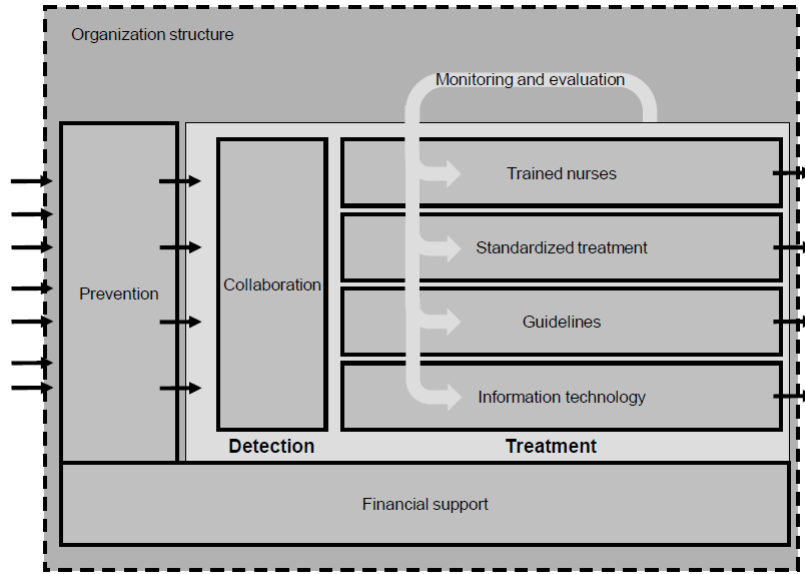


Figure 3.1: Disease management system for chronic skin cancer [23]

## 3.2 Modeling clinical guidelines

In order to make clinical guidelines suitable for compliance analysis, they should be captured within a process model. This requires a translation, because clinical guidelines are often written down in extensive paper-based documents. In subsection 3.2.1 and 3.2.2 the guidelines are summarized and design choices during translation for subsequently BCC and SCC are explained.

### 3.2.1 Clinical guidelines for basal cell carcinoma

BCC is the most common cancer. It is a slow-growing invasive malignant skin tumor that predominantly affects people with Fitzpatrick skin types I and II [52]. It is a tumor that infiltrates tissue in a three-dimensional way. Metastases are nevertheless extremely rare. BCC is classified into different histological subtypes (e.g., nodular (nBCC)). Other factors that influence the prognosis of BCC are: tumor size, tumor site, definition of clinical margins, histological features of aggression, failures of previous treatments and immunosuppression [52].

The Dutch National Guidelines for BCC are an initiative of the Dutch Institute for Dermatology and Venereology and was last revised in 2014. They present evidence-based guidance for treatment and aim to ensure that patients with BCC receive the best possible care.

The treatment process as described in the clinical guidelines \* can be divided in three main parts (sub-processes). Some of these sub-processes contain smaller parts that could be seen as individual *procedures*. The different sub-processes and procedures will be considered as following: 1. diagnosis (sub-process: figure A.1, procedure: figure 3.3) 2. treatment (sub-process: figure 3.4), 3. and follow-up (sub-process: figure 3.5). Figure 3.2 shows how these sub-processes function together in the main process.

As can be seen in figure 3.2 loops are possible in the control flow dependent on choices made at gateways.

\*The process model(s) for the clinical guidelines used in this study contain(s) design choices made by the researcher. When any doubt existed, these design choices were validated by clinical experts that use the guidelines for BCC on a daily basis.

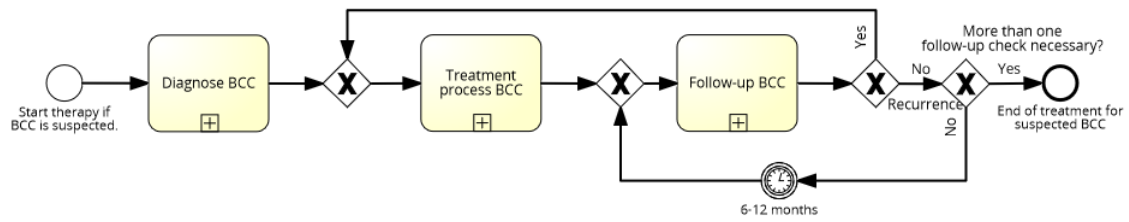


Figure 3.2: Process model for overall BCC therapy process (high abstraction level)

### Diagnosis process BCC

The diagnosis process for BCC is an extensive process (see figure A.1). For that reason, this part of the guideline will be explained by considering the complete diagnosis process and biopsy process separately. Often the guidelines do not precisely state why a certain type a action should be taken. In that case, the eventual choice depends on the subjective, often experience based, judgment of the dermatologist.

In many cases, dermatologists can make a confident diagnosis of BCC [52]. They start with an anamnesis and if there exists any doubt about the initial diagnosis, the guidelines say that a biopsy should be used to clarify. The biopsy process is shown in figure 3.3. First the dermatologist chooses the best type of biopsy for the specific patient and characteristics of the potential carcinoma. Clinicians may also choose to skip biopsy if they feel certain enough about the initial diagnosis. The guidelines do not give a unambiguous recommendation here. As can be seen in figure 3.2 two clinical specialisms (dermatology and pathology) work together in order to make reliable diagnoses for BCC. They communicate with each other by sending (digital) messages. In the pathology report, histological examination gives the histological subtype that could influence the selection of a treatment.

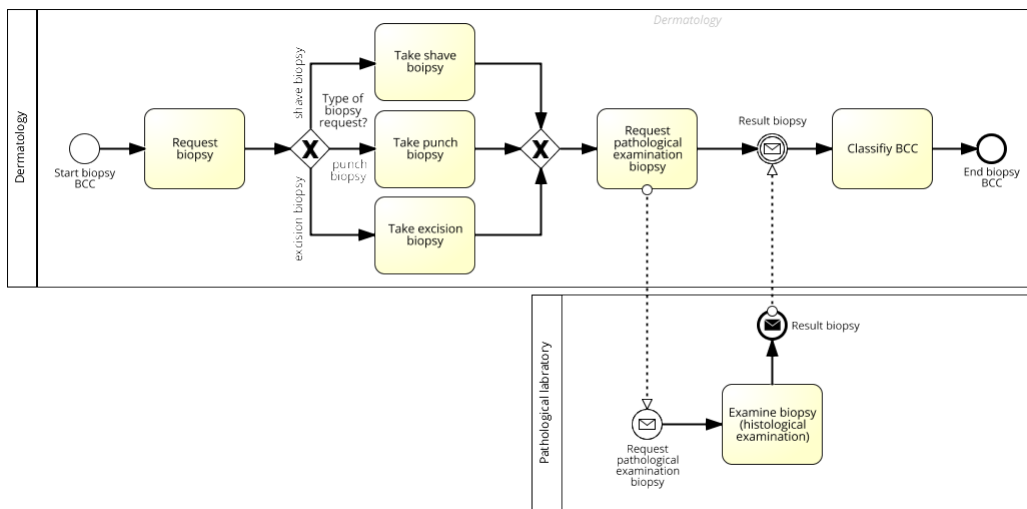


Figure 3.3: Biopsy process for BCC

### Treatment process BCC

Figure 3.4 shows the treatment process for BCC. What stands out is the large number of treatments that a clinician may choose from. The guidelines do not clearly recommend a treatment for certain subtypes of BCC. Therefore the choice for a treatment is dependent on the subjective

judgment, experience and skill set of the physician. Very important is to check whether the treatment did completely remove the tumor. An indication of incomplete removal, should lead to a situation where the patient should undergo another treatment. Histological examination of the removed tissue can be used to decide whether the tumor was completely removed. Moreover, the pathologist can confirm the diagnosis and register additional information about the tumor during the histological research.

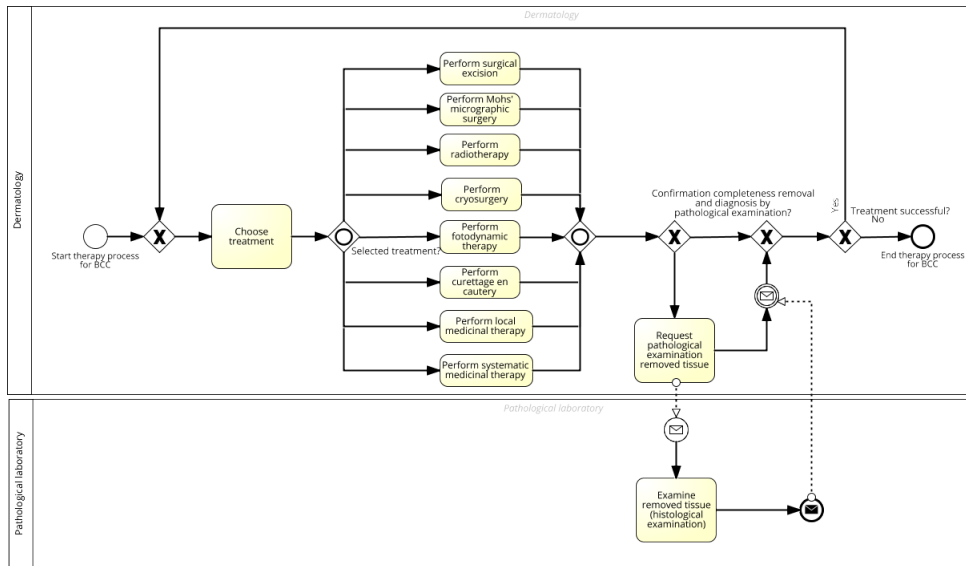


Figure 3.4: Treatment process for BCC

### Follow-up process BCC

Premalignant skin combined with skin cancer is a growing chronic disease [23]. The risk on consecutive BCCs is high. The cumulative chances on a recurrence after six months, one year and 5 years in the Netherlands are respectively: 11%, 14% and 29%. The cumulative 5 year risk worldwide is higher (36%) compared to the Netherlands (29%) [21]. After treatment of BCC, the guidelines require patients and healthcare professionals to be aware of recurrences of the disease. The guidelines instruct regular self-examination by patients and at least one clinical follow-up check. In most cases, one follow-up check after 6-12 months is considered to be enough. The focus should be on self-examination by the patient.

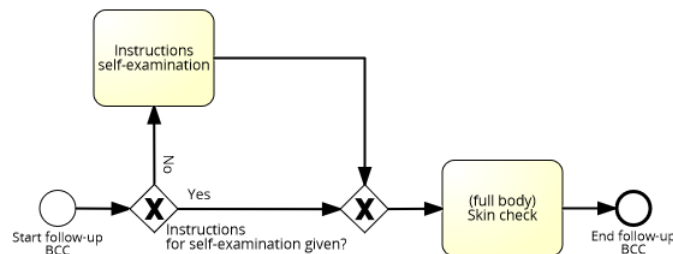


Figure 3.5: Follow-up process for BCC

### 3.2.2 Clinical guidelines for squamous cell carcinoma

SCC is the second most common skin cancer. It is a locally invasive malignant skin tumor and has potential to metastasize to other parts of the body [41]. Its occurrence is mainly due to chronic UV light exposure, which is illustrated by the fact that more than 80% of the SCCs are located in the head and neck region and the other 20% mainly in sites that are exposed to sunlight. The diagnosis for SCC is established histologically. SCC is classified into different histopathological subtypes (e.g., spindle) with various degrees of differentiation (e.g., well), tumor depths (mm), dermal invasions (Clark’s levels). It is also investigated whether the tumor is in absence of perineural, vascular or lymphatic invasions. These histological grades determine the severity of the tumor.

The Dutch Nation Guidelines for SCC are also an initiative by the Dutch Institute for Dermatology and Venereology and was last revised in 2012. They recommend evidence-based guidance for treatments and psychosocial patient support.

The treatment process as described in the clinical guidelines for SCC <sup>2</sup> can be divided in the same parts (sub-processes) as the BCC guidelines (see figure 3.6). Each of these sub-processes can again be divided in smaller parts that in some cases contain individual procedures, which will be discussed for each sub-process. 1. diagnosis (sub-process: figure A.2, procedures: figures 3.7, 3.8, 3.9, 3.10), 2. treatment (procedure: 3.11) , 3. and follow-up (procedure: 3.12). In figure 3.6 the control flow for the overall process (containing the various sub-processes and procedures) is shown.

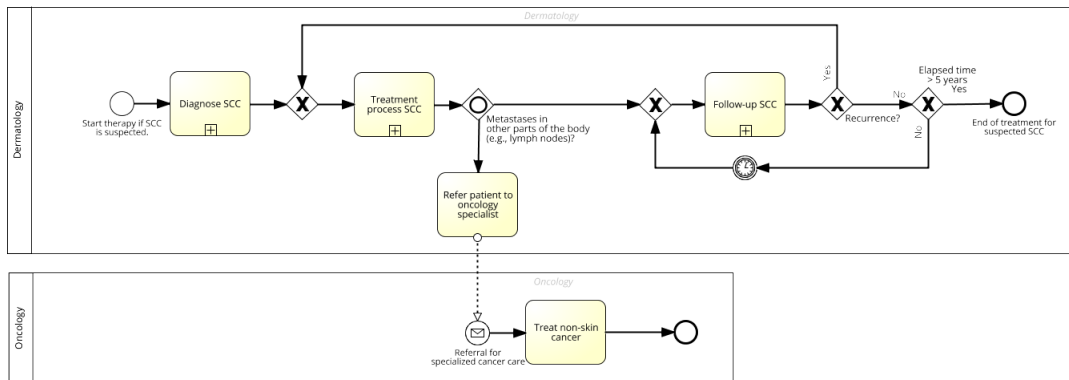


Figure 3.6: Process model for overall SCC therapy process (high abstraction level)

#### Diagnosis process SCC

The diagnosis process for SCC is an comprehensive process (see figure A.2). For that reason, the complete diagnoses process will be explained in parts. These parts are based on the different procedures within the complete diagnosis process. The procedures are “initial physical examination (figure 3.7)”, “physical examination for patients suspected of SCC (figure 3.8)”, “biopsy (figure 3.9)” and “additional diagnostics (figure 3.10)”. The overall control flow for the diagnosis (sub-)process is shown in figure 3.6.

Because of the prognostic factors of SCC, a good initial physical examination after the anamnesis is essential. This procedure is shown in figure 3.7. Location, size and induration of the potential tumor should be assessed and registered. Palpation should give more information about perineural, vascular and lymphatic invasion. The anamnesis in combination with the initial physical examination should give enough information to confirm or reject clinical suspicion for SCC.

<sup>2</sup>The process model(s) for the clinical guidelines used in this study contain(s) design choices made by the researcher. When any doubt existed, these design choices were validated by clinical experts that use the guidelines for SCC on a daily basis.



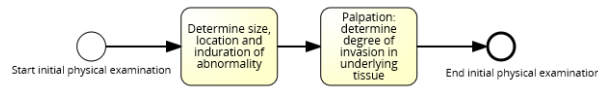


Figure 3.7: Initial physical examination process for SCC

If a patient is suspected of SCC, two separate procedures can be started. A more thorough physical examination should indicate whether there are metastases in the lymph nodes. Additionally, the entire skin should be investigated for more skin tumors.

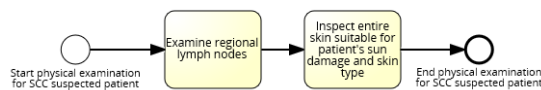


Figure 3.8: Physical examination for patients suspected of SCC

Meanwhile, a biopsy is needed to enable histological examination of the tumor. The biopsy process is shown in figure 3.9. First the dermatologist chooses whether to take a punch biopsy or an excision biopsy. This choice is based on the specific patient and tumor characteristics. A biopsy should be deep enough in order to evaluate the depth of invasive ingrowth, potential perineural growth and/or angio-invasive growth. For that reason a shave biopsy is not suitable.

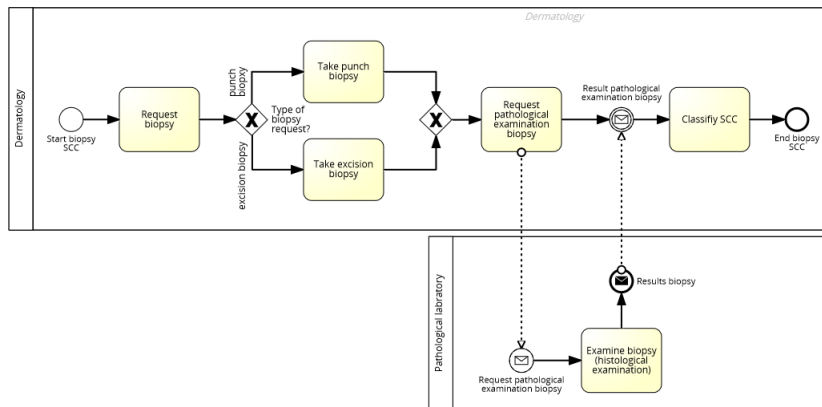


Figure 3.9: Biopsy process for SCC

Because SCC has metastatic potential, additional diagnostics are necessary. Factors which influence this potential include anatomical site, size, tumor thickness, level of invasion, rate of growth, aetiology, degree of histological differentiation and host immunosuppression [41]. Most of these factors contribute to a so called TNM classification that describe the stage of a patient's cancer. T describes the size of the tumor and whether it had invaded other tissue. N describes involved lymph nodes. M describes the presence of metastases. SCC TNM levels higher than II could be seen as high risk tumors. In table 3.1 the prognostic grouping for TNM levels is shown.

Table 3.1: TNM classification

Prognostic group	Size	Regional lymph node metastases	Distant metastases
<b>Stadium 0</b>	T in situ	N0 <sup>a</sup>	M0 <sup>b</sup>
<b>Stadium I</b> <sup>c</sup>	T1 <sup>d</sup>	N0	M0
<b>Stadium II</b>	T2 <sup>e</sup>	N0	M0
<b>Stadium III</b>	T3 <sup>f</sup>	N0	M0
	T1, T2, T3	N1 <sup>g</sup>	M0
<b>Stadium IV</b>	T1, T2, T3	N2 <sup>h</sup>	-
	T4 <sup>i</sup>	-	-
	-	N3	-
	-	-	M1 <sup>j</sup>

<sup>a</sup>N0 = no regional lymph metastases

<sup>b</sup>M0 = no distant clinical metastases

<sup>c</sup>Stadium I with two or more risk factors for SCC should be classified in stadium II

<sup>d</sup>T1 = tumor  $\leq$  2 cm

<sup>e</sup>T2 = tumor  $>$  2 cm

<sup>f</sup>T3 = dermal invasion in muscle, bone, cartilage, jaw or orbit

<sup>g</sup>N1 = one lymph node metastasis  $>$  3 cm

<sup>h</sup>N2 = one or more lymph node metastases  $>$  3 cm and  $<$  6 cm

<sup>i</sup>T4 = direct tumor ingrowth, perineural invasion in skull or axial skeleton

<sup>j</sup>M1 = distant metastases

Additional diagnostics for SCC are especially needed when a patient's cancer is classified higher or equal to TMN level II. The additional diagnostics process is shown in figure 3.10. This process could be an interplay of multiple appointments at different specialisms. The anatomical site and prognostic grouping mainly decide what path a patient takes through the additional diagnostics process. When outcomes of tests are negative, but there exists a strong suspicion of SCC, a patient can loop through the process, whereby it could be referred to a specialized cancer institute.

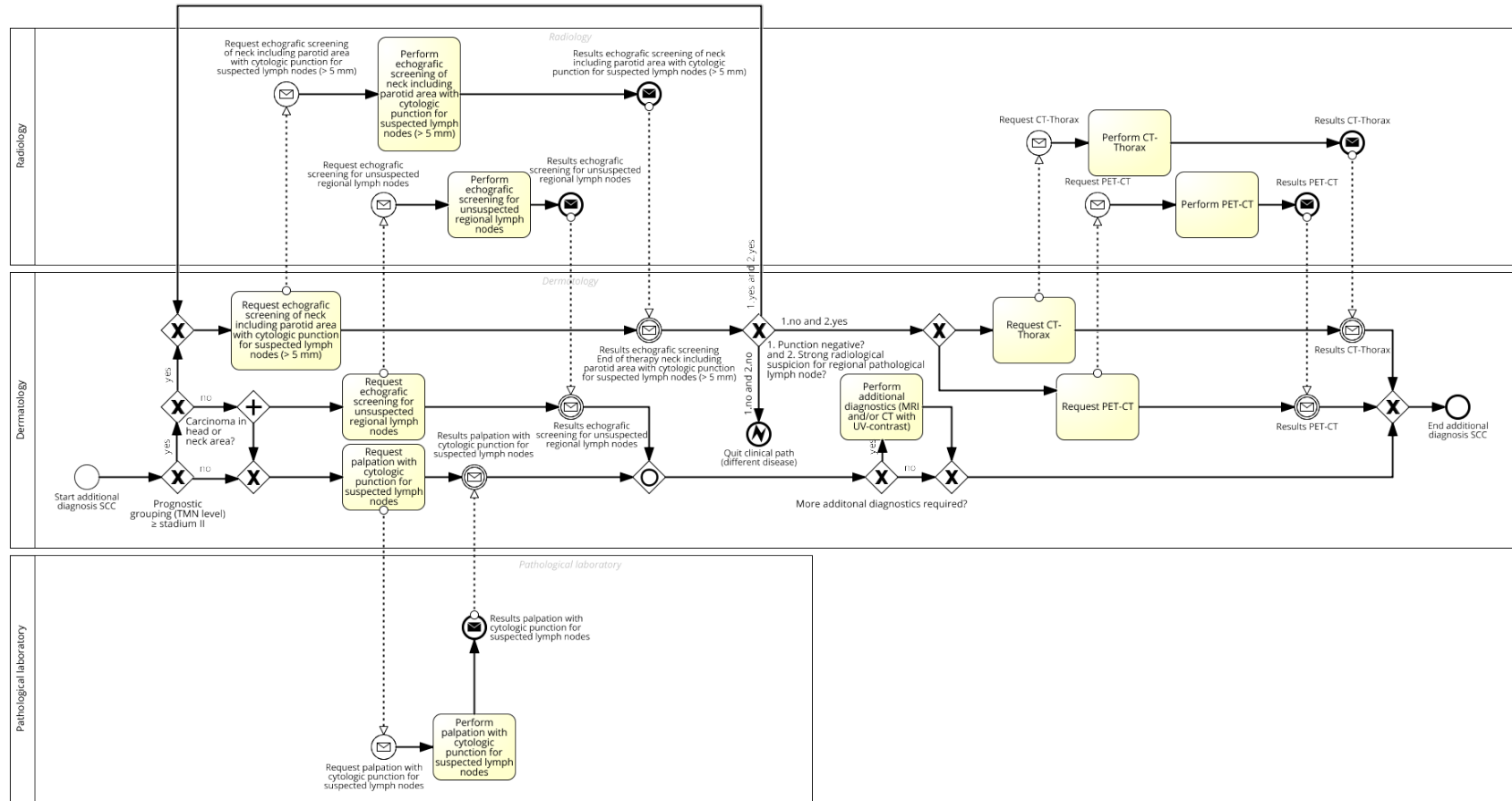


Figure 3.10: Additional diagnostics process for SCC

### Treatment process SCC

As can be seen in figure 3.11, the guidelines recommend a smaller number of treatments for SCCs than for BCCs. The goal of a SCC treatment is complete removal or destruction of the tumor and any of its metastases [41]. Histological assessment is seen as the best way to identify tumor margins and achieve the goal for SCC treatment. There is a lack of randomized controlled trials (RCTs) for the treatment of SCC. Thence, the guidelines for SCC do not clearly recommend a treatment for (subtypes of) SCC. The choice for a treatment is dependent on the subjective judgment, experience and skill set of the physician. Curettage and cautery, cryosurgery, and to a lesser degree radiotherapy are all techniques in which the outcome depends of the experience of the physician. Although the same could be said of surgical excision and Mohs micrographic surgery, these two modalities provide tissue for histological examination that allows the pathologist to assess the adequacy of treatment and for the physician to undertake further surgery if necessary [41]. These different treatment options are explained in table 3.2. Although histological confirmation of complete removal is preferable, most treatments rely on clinical judgment. An indication of inadequate removal, should lead to a repetition of the treatment.

In the British guidelines for SCC [41] some guidance is given considering treatment decision-making. Treatment options including considerations are shown in table 3.2.

Table 3.2: Treatment options for SCC

Treatment	Indications	Contraindications	Notes
<b>Surgical excision</b>	All resectable tumors	Where surgical morbidity is likely to be unreasonably high	Generally treatment of choice for SCC High risk tumors that need wide margins or histological margin control
<b>Mohs' micrographic surgery</b>	High risk tumors and recurrent tumors	Where surgical morbidity is likely to be unreasonably high	Treatment of choice for high risk tumors
<b>Radiotherapy</b>	Non-resectable tumors	Where tumor margins are ill-defined	
<b>Curettage and cautery</b>	Small, well-defined low-risk tumors	High risk tumors	Curettage may be useful prior to surgical excision
<b>Cryosurgery</b>	Small, well-defined low-risk tumors	High risk tumors, recurrent tumors	Only suitable for experienced practitioners

### Follow-up process SCC

To improve the chances of survival of patients with recurrent disease, early detection is essential. A recurrent tumor may arise due to failure to treat the whole tumor or from local metastases.

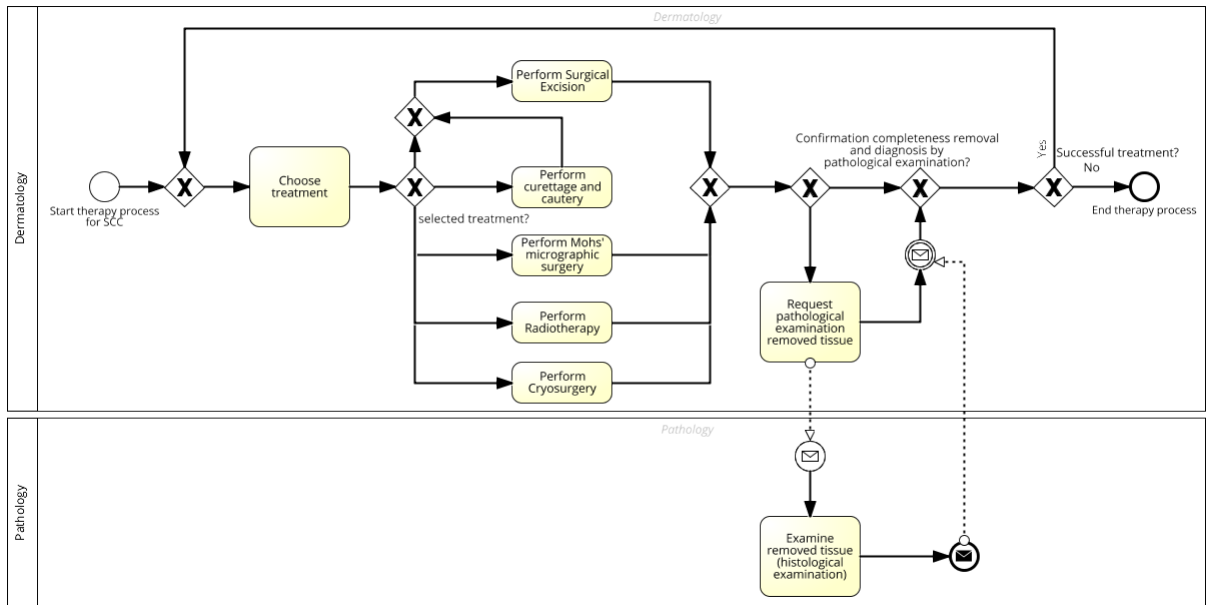


Figure 3.11: Treatment process for SCC

75% of local recurrences and metastases are detected within 2 years and 95% within 5 years. [48]. This indicates that it is reasonable to keep patients who have had a high risk SCC to be kept under close observation.

The guidelines require patients and healthcare professionals to be aware of recurrences of the disease. The guidelines instruct regular self-examination by patients and clinical follow-up checks for a period of 5 years. The suggested frequency of follow-up appointments is shown below:

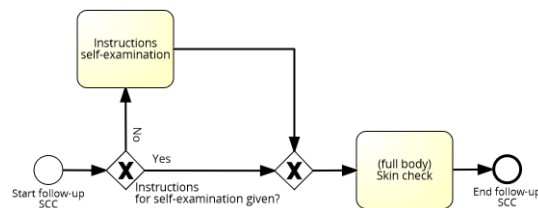


Figure 3.12: Follow-up process for SCC

Suggested follow-up schema for high risk SCCs (TNM level II or higher):

- 1 appointment per 3 months (year 1 after treatment)
- 1 appointment per 4 months (year 2 after treatment)
- 1 appointment per 6 months (year 3 after treatment)
- 1 appointment per 12 months (year 4 & 5 after treatment)

Suggested follow-up schema for low risk SCCs:

- 1 appointment per 6 months (year 1 & 2 after treatment)
- 1 appointment per 12 months (year 3, 4 & 5 after treatment)

### 3.2.3 Process model design choices

Applying a traditional business process methodology to healthcare raises several challenges. In [47] it is stated that its “processes span many disciplines, most involving complex sets of clinical activities. There is great variability from institution to institution depending on the clientele, the range of services offered and the technological infrastructure. Unlike a manufacturing production line, every patient has his or her own unique pathway through the system, which in most cases, cannot be entirely foreseen. Healthcare institutions are also subject to constant changes, for example, new clinical procedures, departmental reorganizations and new standards.” This study aims to capture the clinical process as specified by the clinical guidelines within a process model. The guidelines allow patients to have unique pathways through the system. They aim to prescribe behavior in stead of capturing all behavior possible, which makes manual modeling easier. In order to make a translation from textual documents to process models, design choices are unavoidable. From experience, one of the most important considerations during modeling is the level of detail considered (i.e., granularity). The granularity indicates how detailed a process is represented. It is difficult to choose an appropriate granularity, meanwhile there is no scientific evidence for it [12]. In order to determine the appropriate granularity, the approach proposed by [12] is used. This approach starts with a simple model while step-by-step adding more details until the model meets all criteria. Because the process model will be used for compliance checking against event data, its granularity is restricted by the data’s level of detail. Disparities between their levels of detail, would lead to inconsistencies during compliance checking.

Starting from an abstract model while step-by-step increasing the level of detail, is an example of a top-down modeling approach [57]. The requirements as specified in the guidelines were the starting point for this study. For ambiguous descriptions in the guidelines, interviews with medical professionals that use the guidelines on a daily basis were held.

In order to optimize the model structure and understandability while minimizing the error probability, the Seven process modeling guideliness (7PMGs) [40] were used as guidance. These guidelines are summarized in table 3.3.

Table 3.3: Seven process modeling guidelines

---

Seven process modeling guidelines
<b>Guideline 1: Use as few elements in the model as possible</b> Larger models are harder to understand and the likelihood of errors becomes larger when models are larger. The advice is to keep the model as small as possible.
<b>Guideline 2: Minimize the routing paths per element</b> The more input and output arcs an element contains, the harder it becomes to understand a model.
<b>Guideline 3: Use one start and one end event</b> The use of multiple start and end events increased the probability of errors. Additionally, models satisfying this requirement are easier to understand and can be used to run analyses.
<b>Guideline 4: Model as structured as possible</b> Every split connector should have a matching join connector. Models that do not satisfy this requirement have a higher probability of errors and are more difficult to understand.
<b>Guideline 5: Avoid OR routing elements</b> Using only AND- and XOR-elements decreases the chance on errors and increases the chance that a certain system can deal with the model.
<b>Guideline 6: Use verb-object activity labels</b> The verb-object style is significantly less ambiguous and more useful than other styles (e.g., action-noun labels).
<b>Guideline 7: Decompose the model if it has more than 50 elements</b> This guideline relates to the first guideline. The risk of errors is more than twice at large in models with more than 50 elements. Sub processes can be added and replaced with one activity (collapsed sub-process).

---

# Chapter 4

## Case study

The method and techniques described in the previous chapters are applied during a case study at the dermatology department of the Bravis Hospital. This chapter describes the steps needed to perform the methodology for the specific clinical context in order to find answers to the research questions (RQs) described in chapter 1.

Below the methodology is shown as introduced in chapter 1. In this chapter, these steps are discussed in more detail with respect to the case study.

1. *Conduct a literature study on process modeling, process compliance and the healthcare domain (clinical guidelines, dermatology, non-melanoma skin cancer).*

The main concepts from literature that are used in this thesis are discussed in chapter 2. Moreover, non-melanoma skin cancer is introduced in the previous chapter. In this chapter the specific context in which the methodology is applied will be discussed.

2. *Specify clinical guidelines in terms of a process modeling language.*

This was the main topic of the previous chapter (chapter 3). It was shown how the clinical guidelines for BCC and SCC were translated in terms of the BPMN 2.0 process modeling language. Why BPMN 2.0 is chosen was discussed in subsection 2.1.1. Models were build according to the 7PMGs [40].

3. *Study, extract and transform patient- and treatment process log data from the datawarehouse behind the HIS and EPR.*

This chapter introduces the resources used, what their role is in this study and their origin. Additionally the approach of extracting and transforming the data derived from the resources is explained.

4. *Describe the data challenges to be able to perform a compliance analysis.*

Because the data preparation of the data in this study was everything but straightforward, it is decided to dedicate a separate chapter (chapter 5) to the data challenges in order to be able to evaluate the clinical guidelines for NMSCs and patients diagnosed with this disease.

5. Perform a compliance analysis and evaluate the results.

As the data study and preparation is crucial in this approach, this chapter starts to explore the data that is available during the case study.

### 4.1 Bravis hospital

The Bravis hospital offers a broad spectrum of specialistic care in the western region of Noord-Brabant (Netherlands). Its care region covers parts of Zeeland (until Tholen) and North-Western Belgium (until Essen) as well.

Below some statistics for the Bravis hospital are listed:



Table 4.1: Key figures Bravis hospital, from 31/12/2014 - 24/9/2015

Indicators	numbers <sup>a</sup>
Number of specialists	287
Number of employees (full time equivalent (FTE))	1,953
Number of clinical admissions	29,809
Number of nursing days	133,744
Number of outpatient treatments	26,169
Number of outpatient appointments	495,332
Total revenue (in euro)	237,954,000

<sup>a</sup> statistics are constantly updated on the following website: <https://www.bravisziekenhuis.nl/over-bravis/over-bravis>

Patients diagnosed with NMSCs in our dataset are being treated by specialists from the specialisms: 1. dermatology and 2. surgery. Hospital locations of these specialisms are located in Bergen op Zoom and Roosendaal. During treatment of NMSCs the clinicians work together with the specialisms Pathology and Radiology as shown in chapter 3.

## 4.2 Data sources

In this study data from two sources is used. The first resource is the Dutch national pathological anatomic database (PALGA <sup>1</sup>). This system is used by the pathologists in the Bravis hospital. PALGA is used in order to identify all patients diagnosed with BCC and SCC in the Bravis hospital. The second resource is ChipSoft’s database behind the HIS/EPR (HiX) as used by the Bravis hospital.

### 4.2.1 PALGA

Pathology reports in the Netherlands are digitally archived in the “Pathologisch-Anatomisch Landelijk Geautomatiseerd Archief (PALGA) (National Pathology Automated Archive)”. This archive was founded in 1971 and since 1991 covers all pathology cases in the Netherlands [14]. This means that 55 pathology laboratories are connected to the national infrastructure. The pathology laboratories together add 2.4 million new results of cytology, histology and autopsies [44].

The PALGA infrastructure consists of 55 decentralized databases in the laboratories. The 55 laboratories send their daily automated pathology results to the national database. As a result the national database stay up-to-date [44].

From this first recourse several comma-separated values (CSV) files were extracted with all available information for all patients that have been diagnosed with BCC and SCC in the time period from 2009 until 2015 by a pathologist and is treated by a dermatologist or surgeon in the Bravis hospital.

### 4.2.2 ChipSoft: HiX - Datawarehouse

Nowadays hospitals have the to opportunity to save and use a wealth of data. The Bravis hospital uses a HIS/EPR supplied by ChipSoft. HiX is the main component of ChipSoft’s software. Hospitals have the choice to use a separate datawarehouse module in order to report and analyse the data without negatively influencing the operational HiX software in the hospital. The HiX-datawarehouse has a denormalised, rational database structure. Every data model has one or more dimension tables. The HiX-datawarehouse is based on Microsoft technology and data can be extracted using SQL tools.

---

<sup>1</sup>[www.palga.nl](http://www.palga.nl)

From this second and biggest resource data extraction is more challenging. Not only because there is enormous amount of data available, but also because a solid understanding of the data in the hospital production database is needed. In consultation with ChipSoft and IT staff and clinicians from the Bravis hospital, we extracted the tables within the production database in order to study all activities performed to our patient population. The query that we used can be found in appendix D.

### 4.3 Data extraction

For this study all data related to the events performed during appointments and operations for every patient diagnosed with BCC and SCC were extracted. The most important columns extracted are shown in table 4.2.

Table 4.2: Columns extracted from the HiX - datawarehouse

Column extracted	Explanation
Patient number	The unique patient number used in the hospital
Gender	The patient's gender
Date of birth	The patient's date of birth
Appointment number	The unique number for an appointment
Operation number	The unique number for an operation
NZA <sup>a</sup> proceeding code	National codes used to declare clinical proceedings
NZA proceeding description	National descriptions for the proceeding codes
Date	The date the proceeding is performed

<sup>a</sup>Nederlandse Zorgautoriteit (Dutch Care Authority)

Because we did not have access to PALGA, a pathologist in the Bravis hospital made a standard extraction (provided by the PALGA software) that contained all available information registered by the pathologists in the Bravis hospital for the patients diagnosed in the period of interest. The dataset contained the columns that are explained in table 4.3. Only the most relevant columns for our study are shown.

Table 4.3: Most important columns in PALGA extraction

Column	Explanation
Patient number	The unique patient number used in the hospital
Gender	The patient's gender
Date of birth	The patient's date of birth
Report number	The unique number for one pathological episode (which could contain numerous examinations)
Diagnosis	This column contains a wealth of characteristics of the diagnosis in an unstructured manner (This will be explained into more detail in the next chapter)
Specialism	The specialism of the requesting clinician
Doctor	The clinician that requests the pathological examination

### 4.4 Data pre-processing

The pre-processing of the data is likely to contain unique elements for this case study, not all pre-processing steps will need to be applied in similar studies. This is not surprising because the

data pre-processing is highly dependent on the way the data is registered in the system. Our main objective at the beginning of this data study was to preserve as much valuable data as possible about the characteristics, appointments and operations applicable to patients diagnosed with BCC and SCC. During this data preparation phase data challenges were discovered for the current hospital (information system) environment.

#### 4.4.1 Log preparation

Event logs in a clinical setting are already explained earlier in subsection 2.1.2. In this study, the same structure as explained by [31, 32, 33] is used. This log structure is also shown in table 4.4. For each row in the table a clinical event  $e$  is represented as  $e = (pid, a, t)$ , where  $pid$  is the patient identifier of  $e$ ,  $a$  is the activity type (e.g., first outpatient appointment) of event  $e$ , and  $t$  is the time-stamp of activity  $a$ . A clinical event is a clinical activity occurring at a particular time-stamp. A patient trace ( $\varepsilon_i$ ) consists of one or more clinical events and is represented as  $\varepsilon = \langle e_1, e_2, \dots, e_n \rangle$ . In order to check the compliance for a patient trace through a model, the activities in the log require to have exactly the same name as the activities in the model. This should be taken into account when preparing the log.

Table 4.4: Log structure

PatientID ( $pid$ )	Activity ( $a$ )	Timestamp ( $t$ )
UniqueID	Character string	yyyy/MM/dd hh:mm:ss
$\vdots$	$\vdots$	$\vdots$

After the data is prepared in a table-like structure it should be exported to MXML format for further analyses. In this thesis the academic version of Disco <sup>2</sup> is used to explore the data in the log. Additionally this software is able to export a table like structured data log in the required MXML language. Disco is a tool that is based on parts of process mining theory. Disco's major functionality is process **Discovery** using the Fuzzy miner algorithm [26]. Disco is an easy-to-use tool that still can make sense out of large and complex datasets. This is the reason why it is a valuable tool in the data preparation process in which we want to make sense out of complex clinical datasets. This makes it possible to focus on the actual problems in the process data without diving into scientific theory about process mining discovery approaches.

#### 4.4.2 Dataset preparation

Eventually this study needs to come to a point where understanding about the patients in the dataset (especially whether it is likely that they had to come back for multiple treatments on the same tumor) and relate this to the process diagnostics from the compliance analysis. Table 4.5 gives an idea about the final dataset structure. Every row in the final dataset contains a unique patient that is investigated for the period 2009-2015. The values for the characteristics that change over time are all kept constant at the initial level when the patient first occurred in the pathology dataset.

---

<sup>2</sup><https://fluxicon.com/>

Table 4.5: Dataset structure

PatientID	Gender	Age	Doctor	Specialism	Recurrent tumor
UniqueID	Male (m)/ Female (f)	Integer	Name doctor	Name specialism	yes (y)/ no (n)
⋮	⋮	⋮	⋮	⋮	⋮

## 4.5 Patient population in datasets

As explained earlier in this thesis, two different patient populations are studied.

The first population includes all patients diagnosed with BCC between 1/1/2009 and 31/12/2014 by the pathological laboratory in the Bravis hospital.

In figure 4.1 the number of BCC occurrences over the different years are plotted. As can be seen, the increase is striking. Although this can have a number of reasons (e.g., better registrations, growth of the hospital) this phenomena is also described in literature (see section 3.1). Overall the mean number of BCC cases per year (2009-2014) in this study's dataset is equal to: 1972 (sd = 673.22), total cases in dataset is equal to: 11832. The plot is shown that a small fraction of these cases are under treatment by the specialism surgery instead of dermatology. Moreover it should be noticed that more (unique) cases can follow from one and the same patient. The number of unique patients in this dataset is equal to: 5899.

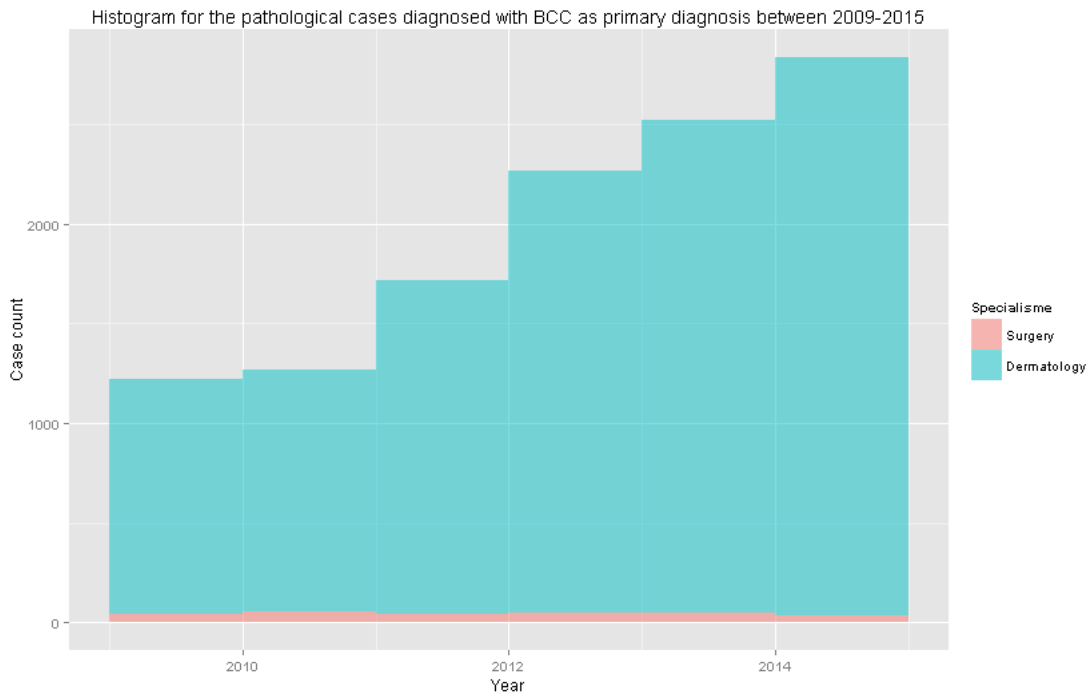


Figure 4.1: Histogram for the pathological cases diagnoses with BCC as primary diagnosis

When looking at the patient's age, all individual patients in the population when they first appear in the dataset are considered. In figure 4.2 the distribution of the different ages in the population are shown. The mean age for the population of patients diagnosed with BCC is equal to: 68.06 years (sd = 13.01 years).

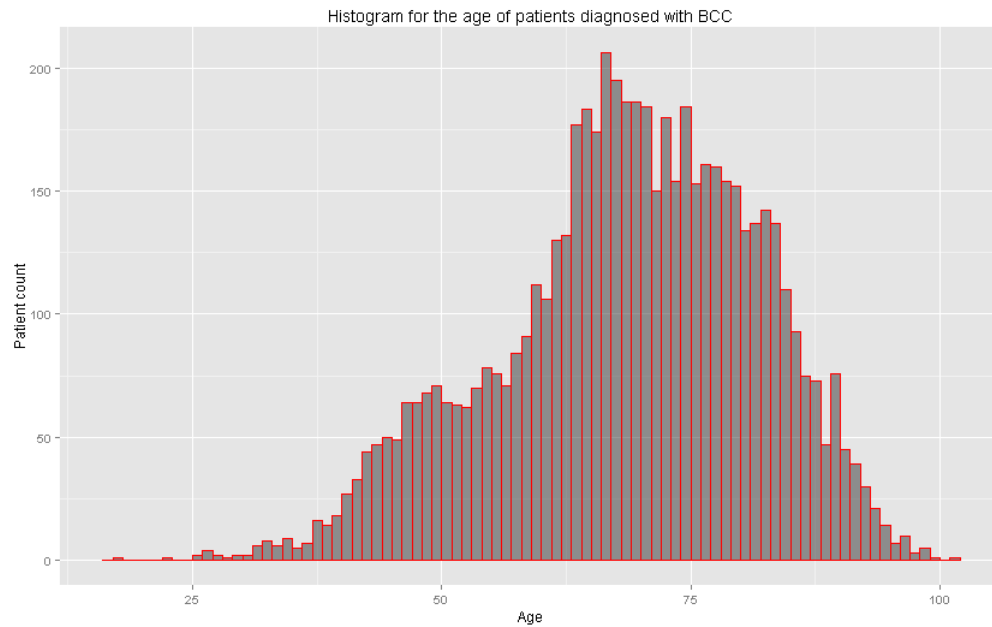


Figure 4.2: Histogram for the age of patients diagnosed with BCC

Finally, we look at some statistics about the gender of this population. It can be seen in figure 4.3, that there is no significant difference in the occurrences of BCC for men or women. The number of males in our BCC population is equal to: 2964 and the number of females is equal to: 2935.

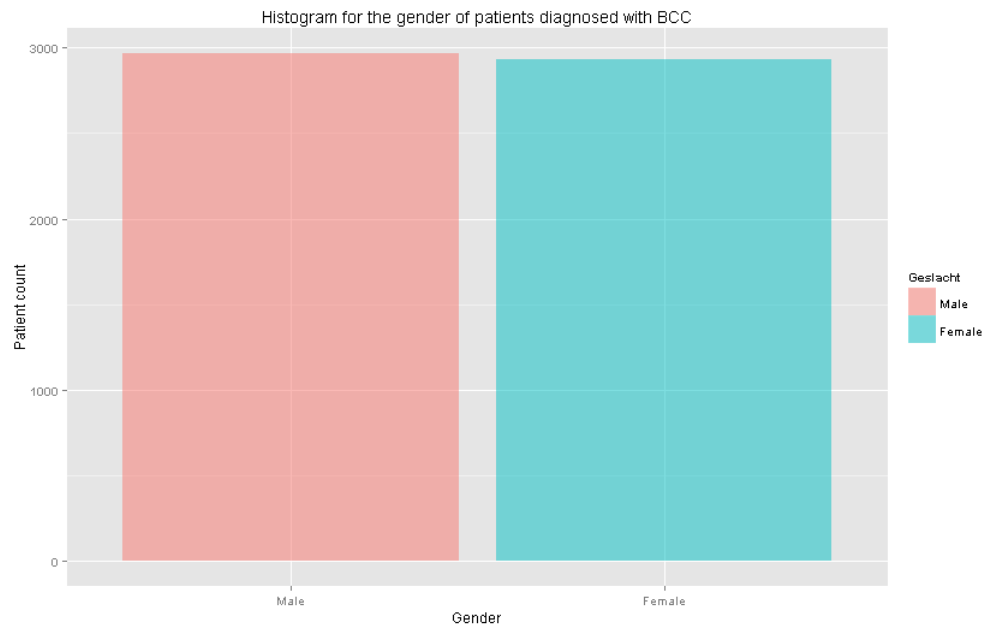


Figure 4.3: Histogram for the gender of patients diagnosed with BCC

The second patient population includes all patients diagnosed with SCC between 1/1/2009 and 31/12/2014 by the the pathological laboratory in the Bravis hospital.

In figure 4.4 the number of SCC occurrences over the different years are plotted. Similar to the graph of BCC, the increase of cases is striking. Although this can have a numerous of reasons, this phenomena is also described in literature for SCC (see section 3.1). Overall the mean number of SCC cases per year (2009-2014) in our dataset is equal to: 329 (sd = 144.93) and total cases in the dataset is equal to: 1974. In the plot is also shown that a small fraction of these cases are under treatment by the specialism surgery instead of dermatology. Moreover it should be noticed that more (unique) cases can follow from one and the same patient. The number of unique patients in this dataset is equal to: 1354.

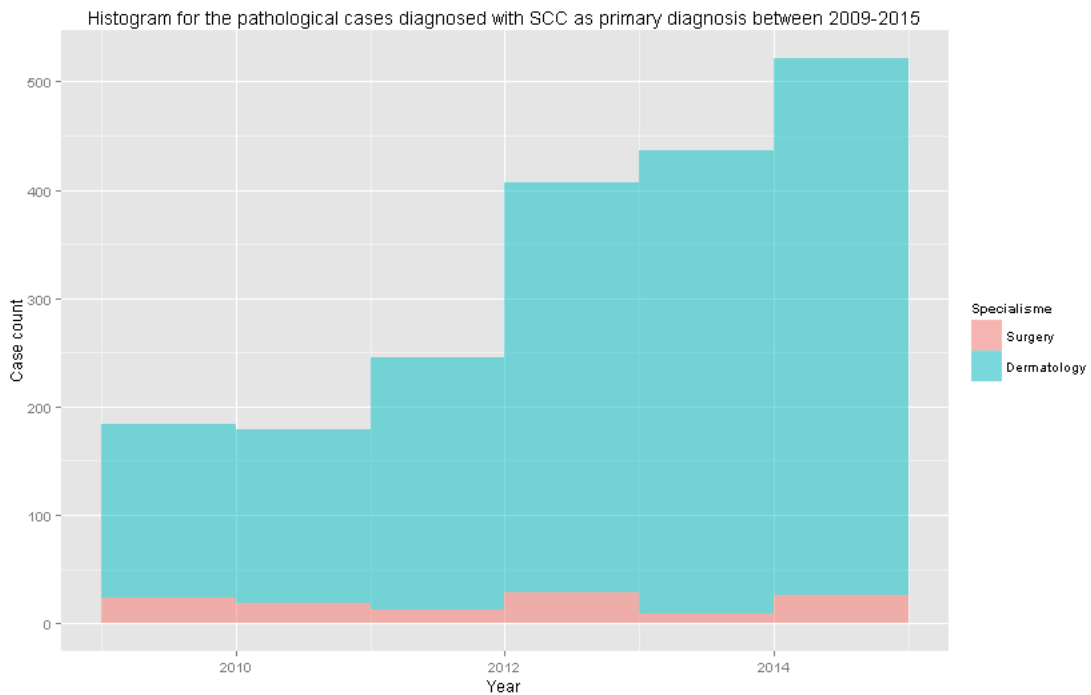


Figure 4.4: Histogram for the pathological cases diagnoses with SCC as primary diagnosis

When looking at the patient's age, all individual patients in the population when they first appear in the dataset are considered. In figure 4.5 the distribution of the different ages in the population are shown. The mean age for the population of patients diagnosed with SCC is equal to: 75.36 years (sd = 10.51 years).

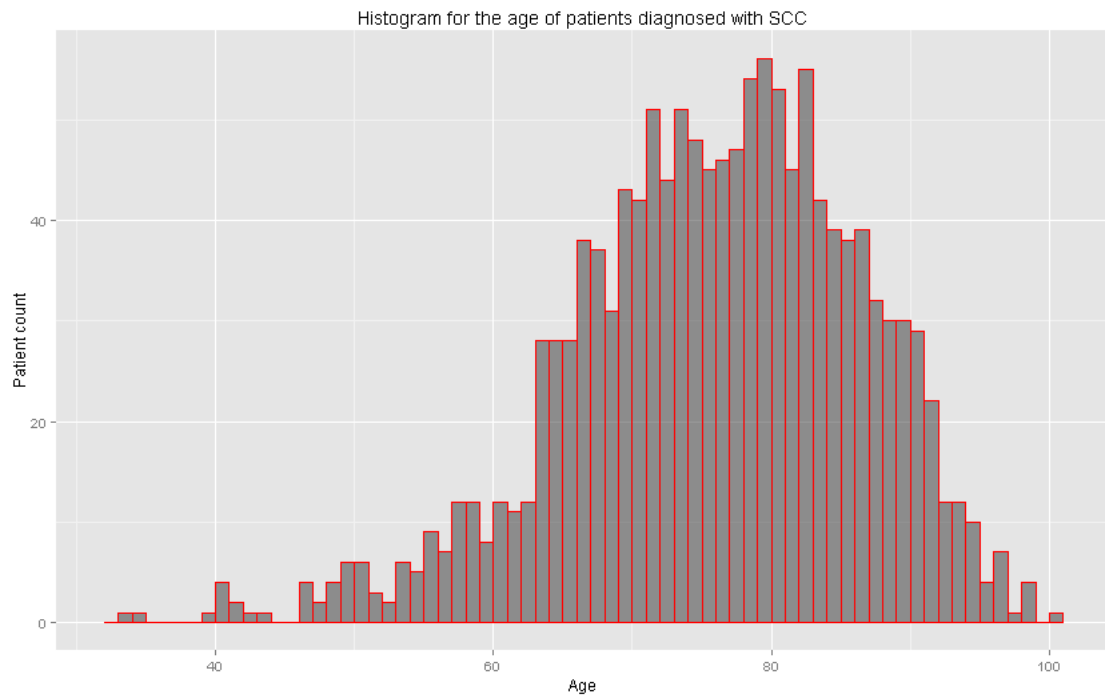


Figure 4.5: Histogram for the age of patients diagnosed with SCC



Finally we look at some statistics about the gender of this population. Figure 4.6 shows, there is a significant difference in the occurrences of SCC for men and women in contrast to the population for BCC. The number of males in our SCC population is equal to: 818 and the number of females is equal to: 536. These numbers are in line with the trends described in medical literature [15].

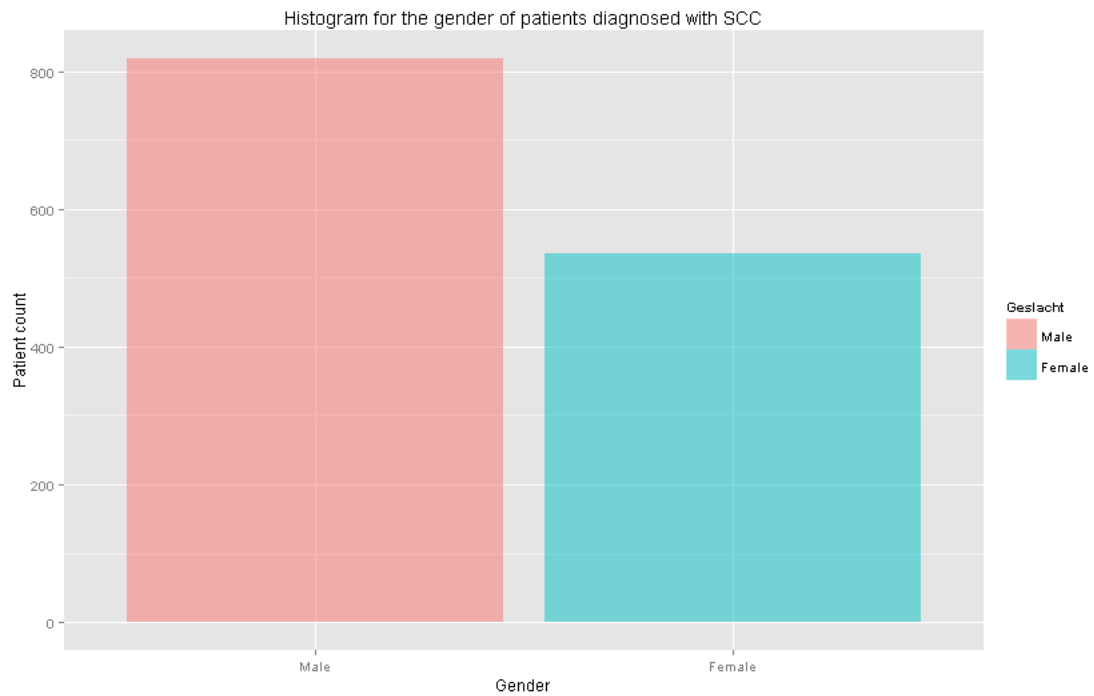


Figure 4.6: Histogram for the gender of patients diagnosed with SCC

## Chapter 5

# Challenges during data collection

Data in hospitals is spread around disparate data sources, as shown in: [39]. In their study they show that these disparate data sources can be divided in: 1. administrative systems, 2. clinical support systems, 3. healthcare logistics systems and 4. medical devices. These four categories of data sources score differently on the level of abstraction, level of accuracy, granularity, directness and correctness. In table 5.1, the four types of data sources are assessed according to these measures [39].

Table 5.1: Spectrum of clinical data sources

	<b>Level of abstraction</b>	<b>Accuracy</b>	<b>Granularity</b>	<b>Directness</b>	<b>Correctness</b>
Administrative systems	High	Low	Low	Low	Average
Clinical support systems	Average	Average	Average	Low	High
Healthcare logistics systems	Average	High/Average	Average/Low	High	High
Medical devices	Low	High/Average	High	High	Average

In this thesis different data sources are used. In HiX an administrative system is coupled to a healthcare logistic systems to support logistic processes (e.g., appointments are coupled with services delivered). Administrative systems take care of the administration and billing of all accountable services including treatments and examinations [39]. Services that have been delivered to patients may be entered manually into the system with a day timestamp. This explains the scores in table 5.1. Because HiX is not a standard administrative system, accuracy and correctness of the data is expected to be higher than for average administrative systems.

The other data source used in this study: PALGA, can be seen as a clinical support system. The pathologists use PALGA, because they have a special needs (i.e., registry of very specific information about certain tissue) that requires a special information system.

For this study as well as for other questions posed by medical professionals, data is required from different data sources. At the moment there are still many challenges in order to answer these questions. Some data challenges were discussed earlier in this thesis (e.g., data extraction). The challenges in this thesis concerning the derived data are addressed below.

**Data challenges in this thesis:**

- dataset - PALGA extraction
  1. Clean: split and arrange columns.
  2. Discover recurrent patients (subsection 5.1.1).
    - (a) Study and define location label tumors.
    - (b) Discover recurrent patients (same diagnosis and location label).
- dataset - HiX extraction
  1. Clean and filter events in the log.
    - (a) Clean: remove duplications.
    - (b) Filter events that are expected to be irrelevant in this study (subsection 5.2.1).
      - DBC codes
      - Account for variability
  2. Align event descriptions in the log with event descriptions in the modeled guidelines 5.2.

In the remaining of this chapter, first two challenging data cleaning tasks are shown that had to be performed to consecutively the PALGA dataset and HiX extraction. Subsection 5.1.1 will be about discovering recurrent patients in the PALGA dataset (including the defining of location labels). Subsection 5.2.1 will treat the topic of filtering events in the log. In the last subsection of this chapter (section 5.2), the naming of events in the models and logs will be discussed.

## 5.1 Log and Data preparation

As stated by many and experienced again in this study, getting the data ready for analysis is a very time consuming task. Various scripts had to be written to perform data transformations. The statistical programming language R is used (a few examples will be explained). To give an impression, a sample of the two datasets is shown (PALGA and an extraction of the HIS (HiX <sup>1</sup>)) as they were before the data transformation and after (anonymization of patient data is already done as well as the construction of a separate dataset to couple the patients in both datasets).

### Dataset: PALGA extraction

Table 5.2: Initial dataset - PALGA extraction

Patient number	Gender	Age	Report number	Diagnosis	Specialism	Doctor
Patient2782	Female	66	Rapport1	huid * neus * rechts * biopt * basaalcelcarcinoom * ...	dermatologie	Gerwen v H
⋮	⋮	⋮	⋮	⋮	⋮	⋮

As can be seen in table 5.2 there is a problem in the *Diagnosis* column in the dataset. This column contains a lot of valuable information. In this initial dataset all this information is kept in one column without much structure. After extracting all unique values from this column, it was discovered that values in this column could be subdivided into the following categories:

---

<sup>1</sup>HiX (Health information eXchange) is the name ChipSoft gave to the HIS/EPR as used in the Bravis hospital.

- Diagnosis
  - Tissue (e.g., huid (skin))
  - Localization (e.g., arm)
  - Side (e.g., links (left))
  - Characteristic (e.g., verdacht maligne (suspected malignant))
  - Resection method (e.g., biopt (biopsy))
  - Primary diagnosis (SCC) (e.g., bowen)
  - Recurrent (e.g., recidief (recurrent))
  - Secondary diagnosis (e.g., melanoom (melanoma))
  - Treatment (e.g., exisie (excision))
  - Confident diagnosis (e.g., geen zekere diagnose (no certain diagnosis))
  - Tumor completely removed (e.g., snijvlak vrij (incision free))
  - Additional examination (e.g., revisie intern discordant (internal revision discordant))
  - Usable material (e.g., geen bruikbaar materiaal (no usable material))
  - Comment (e.g., palga-systeem (palga-system))

After the column is split based on the separator, an R function to assign a value to the corresponding column (category) is used. In this function three datasets are used. The split dataset extracted from PALGA (let's call this dataset  $D1$ ), a dataset in which all unique values are manually divided in one of the categories above (let's call this dataset  $D2$ ) and a dataset which is an exact copy of  $D1$  but the spit columns are replaced by empty columns (filled with  $NA$  values) named after the categories discussed earlier (let's call this dataset  $D3$ ). The function checks for all split columns that originate from the *Diagnosis* column (e.g., *Diagnose\_01*, *Diagnose\_02*, ..., *Diagnose\_n*) and assigns a value in one of the columns if a matching value is found and performs this action for every case. As an illustrative example a part of the R code is shown in which is checked for column *Diagnose\_01* from  $D1$  and assign a value to the *Tissue* column in  $D3$  based on the values in  $D2$  if a matching value can be found for one of the cases in  $i$ .

```
for(i in 1:nrow(D1)){
  if(is.na(D2$Tissue[i])){
    if(D1$Diagnose_01[i] %in% D2$Tissue){
      D3$Tissue[n] <- D1$Diagonse_01[n]
    }
  }
}
```

After this preparation step the initial dataset shown in table 5.2 will transformed to a table as shown in table 5.3.

Table 5.3: Dataset - PALGA extraction (after transformation)

Patient number	...	Doctor	Tissue	Localization	Side	...	Comment
Patient2782	...	Gerwen v H	huid	neus	rechts	...	NA
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

**Dataset: HiX extraction**

Table 5.4: Initial dataset - HiX extraction

Patient number <sup>a</sup>	Event	DeclCode	Timestamp	Start date	End date
Patient2782	Herhaal polikliniekbezoek	190013	2014/01/08	2014/01/01	2014/05/31
Patient2782	HERHAAL POLIKLINKIEKBEZOEK	190013	2009/01/08	2002/04/01	2010/12/31
Patient2782	Pathologisch anatomisch histologisch onderzoek	050501	2009/01/23	2002/04/01	2014/12/31
⋮	⋮	⋮	⋮	⋮	⋮

<sup>a</sup>Appointment number and operation number from the initial extraction are not shown in this table, because they are only used to get an understanding of the data and are not important in this preprocessing example

As can be seen in table 5.4 a common problem while using extractions of the HIS is that the used event description causes duplications in the dataset. This was due to common changes made to these descriptions. The columns *Start date* and *End date* in table 5.4 show the dates between the event description are valid.

Again a R script is used to account for these duplications. The task for this script is to keep only the most recent event description for the corresponding code (*DeclCode*).

The R code looks as follows (let the initial dataset for the HiX extraction be called *D4*):

```
library(dplyr) # required package

D5 <- filter(D4, Event, DeclCode, Startdate, Enddate)
D6 <- filter(D4, Patientnumber, DeclCode, Timestamp)

rows <- c()
for(i in unique(D5$DeclCode)){
  if(nrow(D5[D5$DeclCode == i, ]) > 1){
    rows1 <- D5[(D5$Startdate == max(D5[D5$DeclCode == i, "Startdate"])) &
      D5$DeclCode == i, ]
  }
  else{
    rows1 <- D5[D5$DeclCode == i, ]
  }
  rows <- rbind(rows, rows1)
}
D5 <- rows
D4 <- merge(D6, D5)
```

After this transformation to the dataset (*D4*), duplicate rows are removed and only the most recent event description for an event is used. The dataset in table 5.4 would be transformed in a dataset shown in table 5.5.

### 5.1.1 Discovering recurrent patients within patient populations

In order to find possible negative treatment outcomes for the patients studied in this thesis, clinicians involved in this project were asked what they would define as a negative treatment outcome for patients that were treated for the diseases BCC or SCC. They unanimously explained

Table 5.5: Dataset - HiX extraction (after transformation)

Patient number	Event	DeclCode	Timestamp	Start date	End date
Patient2782	Herhaal polikliniekbezoek	190013	2014/01/08	2014/01/01	2014/05/31
Patient2782	Pathologisch anatomisch histologisch onderzoek	050501	2009/01/23	2002/04/01	2014/12/31
⋮	⋮	⋮	⋮	⋮	⋮

us that the recurrence of a tumor is a reliable indication that the treatment was not effective for the treated patient. This resulted in a new data challenge, because the Bravis hospital does not use an unambiguous manner to register locations for tumors (both in PALGA or HiX).

As shown before the PALGA dataset does contain some data about the location of a tumor. Therefore it was investigated whether it is possible to make a reliable new label for patients with a recurring tumor (assumption: if a patient recurs in the PALGA dataset with a tumor on the same location it is assumed to be a recurring tumor).

Table 5.6 we shows all unique values that are registered for locations of a tumor.

Table 5.6: Terms used to registers tumor locations

Localization SCC	Localization BCC	New location label <sup>a</sup>
aangezicht	aangezicht	25
abdomen	abdomen	28
arm	arm	<b>3</b>
been	been	<b>4</b>
bil	bil	<b>5</b>
borst	borst	4
bovenarm	-	29
bovenbeen	bovenbeen	34
bovenbuik	-	56
bovenlip	-	<b>76</b>
buikhuid	-	28
clavicula	clavicula	<b>6</b>
coeur	coeur	<u>57</u>
duim	duim	<u>63</u>
elleboog	elleboog	31
enkel	enkel	<u>61</u>
flank	flank	<b>8</b>
gelaat	gelaat	25
gewricht	-	<b>9</b>
haar	haar	<b>10</b>
hals	hals	<b>11</b>
hand	hand	<b>7</b>
helix	helix	<b>39</b>
heup	-	<b>13</b>

hiel	-	41
hoofd	hoofd	<b>1</b>
infraclaviculair	-	37
kin	kin	<u>48</u>
knie	knie	36
kruin	kruin	<b>75</b>
kuit	kuit	<u>62</u>
lies	lies	<b>15</b>
lip	lip	<b>66</b>
localisatie onbekend	localisatie onbekend	Exclude
lumbaal	-	42
mamma	mamma	29
mediale ooghoek	-	<b>79</b>
mond	mond	<u>47</u>
navel	-	<u>55</u>
nek	nek	<b>11</b>
neus	neus	<u>50</u>
neuspunt	-	<b>69</b>
neusrug	-	<b>70</b>
neusvleugel	-	<b>71</b>
oksel	oksel	<b>17</b>
onderarm	-	32
onderbeen	onderbeen	35
onderkaak	onderkaak	<u>51</u>
onderooglid	-	<b>78</b>
oog	oog	<u>49</u>
ooghoek	ooghoek	<b>68</b>
ooglid	ooglid	<b>67</b>
oor	oor	<b>12</b>
oorlel	-	<u>65</u>
oorschelp	oorschelp	39
paranasaal	-	<b>72</b>
paravertebraal	-	<b>16</b>
parietaal	-	27
parotis	-	<b>74</b>
pols	pols	33
preputium	preputium	44
presternaal	-	<u>58</u>
retroauriculair	-	40
rib	-	<b>23</b>
rug	rug	<b>16</b>
scapula	-	43
schaambeent	-	<b>19</b>
schedel	schedel	<b>1</b>
scheen	-	<u>60</u>
schouder	schouder	<b>20</b>
slaap	slaap	26
sternum	sternum	<u>58</u>
teen	-	<b>21</b>
tepel	-	<u>59</u>
thorax	thorax	29
tibia	-	<u>60</u>
tragus	-	<u>64</u>
vinger	-	38

voet	-	<b>14</b>
voorhoofd	-	<u>53</u>
vulva	-	<b>22</b>
wang	-	<i>52</i>
wenkbrauw	-	<b>73</b>
-	achterhoofd	<i>27</i>
-	anus	<i>46</i>
-	behaarde hoofd	<i>27</i>
-	glans penis	<i>45</i>
-	labium	<b>66</b>
-	localisatie primaire tumor onbekend	Exclude
-	long	<i>29</i>
-	onderlip	<u>77</u>
-	penis	<b>18</b>
-	perianaal	<i>46</i>
-	romp	<b>2</b>
-	scrotum	<b>24</b>

<sup>a</sup> The appearance of the numbers (e.g., bold, italic, underlined, etc.) gives an indication of the level of precision for the used term. If the the number appears **bold** this means that there are no terms used that have a higher level of abstraction (e.g., hoofd (head)). A number that is of the *italic* type indicates that a term is used that is 1 level less precise (e.g., aangezicht (face)). A number that appears underlined indicated that there are terms used that are 1 and 2 levels less precise (e.g., mond (mouth)). A number that is of the **bold and italic** type indicates that there are terms used that are 1, 2 and 3 levels less precise (e.g., lip/labium). Finally, a number that appears **bold and underlined** indicates that there are terms used that are 1, 2, 3 and 4 levels less precise (e.g., bovenlip (upper lip)).

Based on the new labels defined in table 5.6, patients are assigned to three different groups. 1. Patients that return in the PALGA dataset for a tumor on the same location (same location label and side), 2. Patients that return in the dataset for a tumor on a possible different location and 3. patients that only occur once in the dataset. Because one pathology report may contain more than one event per patient, multiple occurrences in one report should not be counted as recurrent patients. To correct for this, a count function is used in combination with report number. When the number of occurrences for a patient for one report is the same as the total number of occurrences for that patient, this means that a patient only reoccurs in one pathology episode.

In figure 5.1 and 5.2 the total number of patients in the three separate groups are shown for the BCC patient population and the SCC patient population.



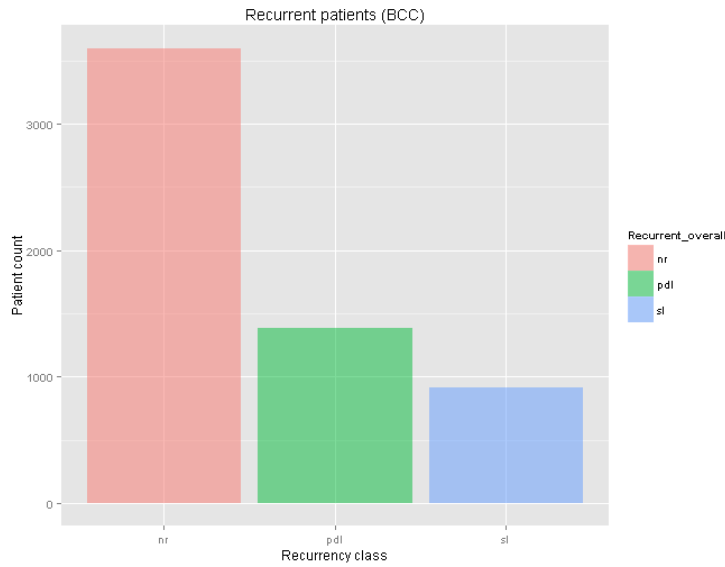


Figure 5.1: Histogram for the number of patients that do not return (nr), return in the pathological dataset with a skin sample from a tumor that was possibly located on a different location on the body (pdl) and patients that reoccur in the dataset with a tumor located on the same location (sl), based on the defined coding.

The number of patients that do not return in the dataset (nr) is: 3599 (61%). The number of patients that do return in the dataset for a possible different tumor (pdl) is: 1384 (23%). The number of patients that do return in the dataset for possibly the same tumor (sl) is: 916 (16%).

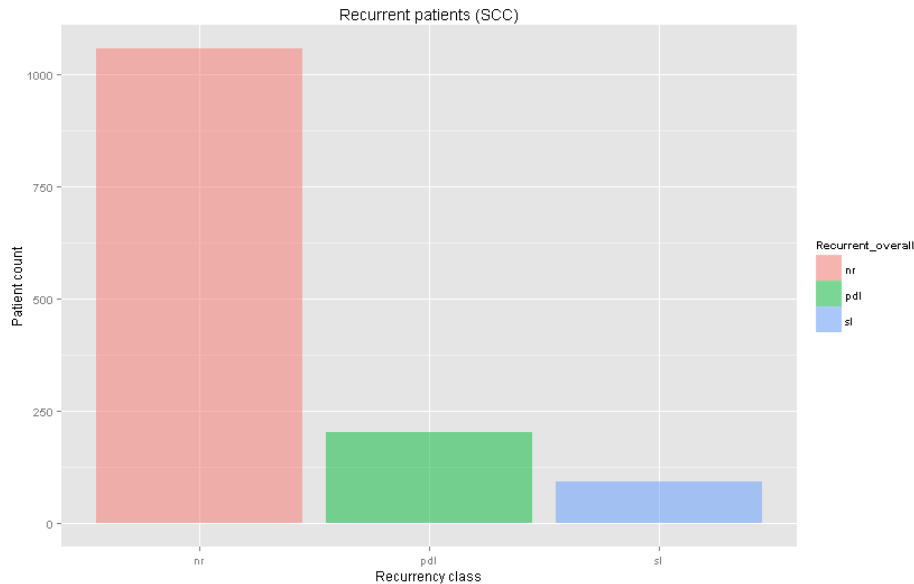


Figure 5.2: Histogram for the number of patients that do not return (nr), return in the pathological dataset with a skin sample from a tumor that was possibly located on a different location on the body (pdl) and patients that reoccur in the dataset with a tumor located on the same location (sl), based on the defined coding.

The number of patients that do not return in the dataset (nr) is: 1057 (78%). The number of patients that do return in the dataset for a possible different tumor (pdl) is: 204 (15%). The number of patients that do return in the dataset for possibly the same tumor (sl) is: 93 (7%).

## 5.2 Aligning events in the modeled guidelines with events in the log

Process models determine a representation of organizational processes with the goal of analysing and studying various aspects of the real-world business process such as the events and the actors related to the process execution [34]. The challenge is to model the behaviour in a process as close to reality as possible. This thesis uses process models to capture the behaviour described in clinical guidelines, while corresponding to events that can be discovered in the hospital's HIS. In section 3.2 the clinical guidelines are translated to process models in BPMN 2.0. During the construction of these models there was no knowledge about the data in the HIS yet. In this section, the activities in the initial process models are compared against the events in the log. To get a good understanding of the log data, process discovery is performed using the Fuzzy Miner algorithm [2, 26]. This algorithm can leave out less important events because it uses significance/correlation metrics to interactively simplify the process model. The discovered processes for patients diagnosed with BCC and SCC can be consecutively seen in figure 5.3 and 5.4 (without simplification).

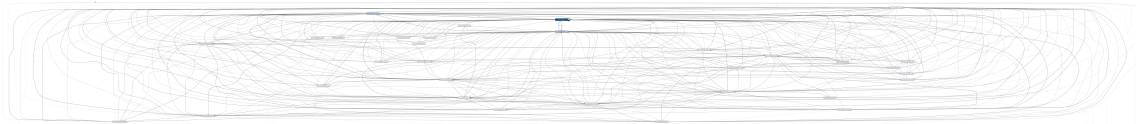


Figure 5.3: Discovered process using process discovery for patients diagnosed with BCC. The log used contains all events that the patients were subject to during the period 2009-2015.

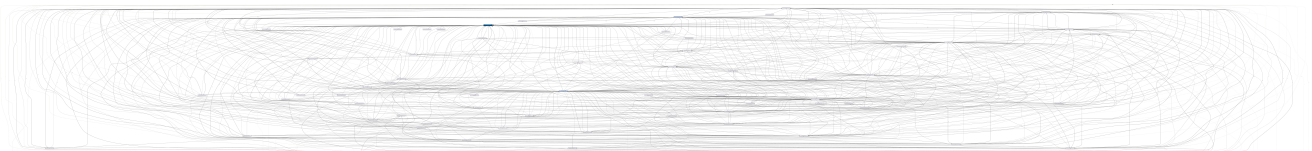


Figure 5.4: Discovered process using process discovery for patients diagnosed with SCC. The log used contains all events that the patients were subject to during the period 2009-2015.

Although the process are barely visible on the pictures, they give a good impression of the variation of the paths that patients with the same diagnosis take through a clinical process.

The models shown in this section (5.2) are data-based models. In an ideal situation, these data-based models have the same activities and control-flow as the knowledge-based models derived from the clinical guidelines. In that case the process would be perfectly compliant. The described ideal situation is of course not realistic. The complexity of the models in this section is due to the large number of events that extend over multiple disciplines and the lack of structure. The used algorithm (Fuzzy Miner) cannot make sense of the data and keeping all details at the same time. Currently there are no existing process mining algorithms that can identify good process models in healthcare [34]. Although the Fuzzy Miner cannot make sense of the whole process including all the details, it includes numerous filters to (visually) explore the log data. In this study, data-based models are not used to replace the knowledge-based models. Instead the data-based models are used to adjust the knowledge-based models in order to make them suitable for the compliance analysis using the event logs. The knowledge-based models are used as basis to keep the control flow as prescribed by the clinical guidelines.

### 5.2.1 Filtering events in the log

Once the relevant data has been located, the extraction of data is fairly straightforward, the cleaning is somewhat more challenging, but the real challenge is to select event data that is relevant to the question tried to answer.

Because this study aims to evaluate the guidelines, it is chosen to filter all the events from the log data that do not relate directly to the guidelines. The selection of events is done manually based on the events in the initial knowledge-based models.

Due to the filtering, only 18% (151,823) of the events, but 100% of the cases (5,689<sup>2</sup>) were kept for patients diagnosed with BCC (32 events were selected from the 2496 in the log). For patients diagnosed with SCC, 21% (45,741) of the events for 99% (1,339<sup>3</sup>) of the cases were kept (65 events were selected from the 1679 in the log). This caused a significant reduction in variability and complexity, although still a fraction of the investigated patients followed similar paths through the process. For BCC 5,325 from the total 5,869 cases are unique. For SCC a similar trend can be noticed, 1,305 from the total 1,339 cases are unique. Again process discovery was used to study the effect of the filtering.

The variability of the discovered model after selection is still high. When only the most frequent events that relate to the clinical guidelines for BCC and SCC are kept (20% of all events), a better visual impression of the events is gained, as can be seen in figure 5.5 for patients diagnosed with BCC and 5.6 for patients diagnosed with SCC. Although the discovered models are easier to (visually) interpret, they also neglects events that are important for clinicians but are not frequent enough to be shown in this models (e.g., Mohs surgery). One should therefore be careful while interpreting the discovery models in this context.

While studying the discovered processes, one can see that their consecutive events can be recognized in the (modeled) guidelines. For example, most patients start their treatment with “first outpatient appointment (eerste polikliniekbezoek)” followed by a treatment (excision) after which a histological examination is performed. This is a frequent path in the log that is also allowed by the (modeled) guidelines.

---

<sup>2</sup>This number is not equal to the total number of patients in the patient population in PALGA, because this study only investigated the patients in HiX that could be joined on the PALGA dataset both on patient number, gender and date of birth. This decreased the number of unique patients from 5,899 to 5,689

<sup>3</sup>The number of unique patients in the log decreased in comparison to the patient population in PALGA (from 1354 to 1,339). This is due to filtering of the log and because some patients could not be joined reliably

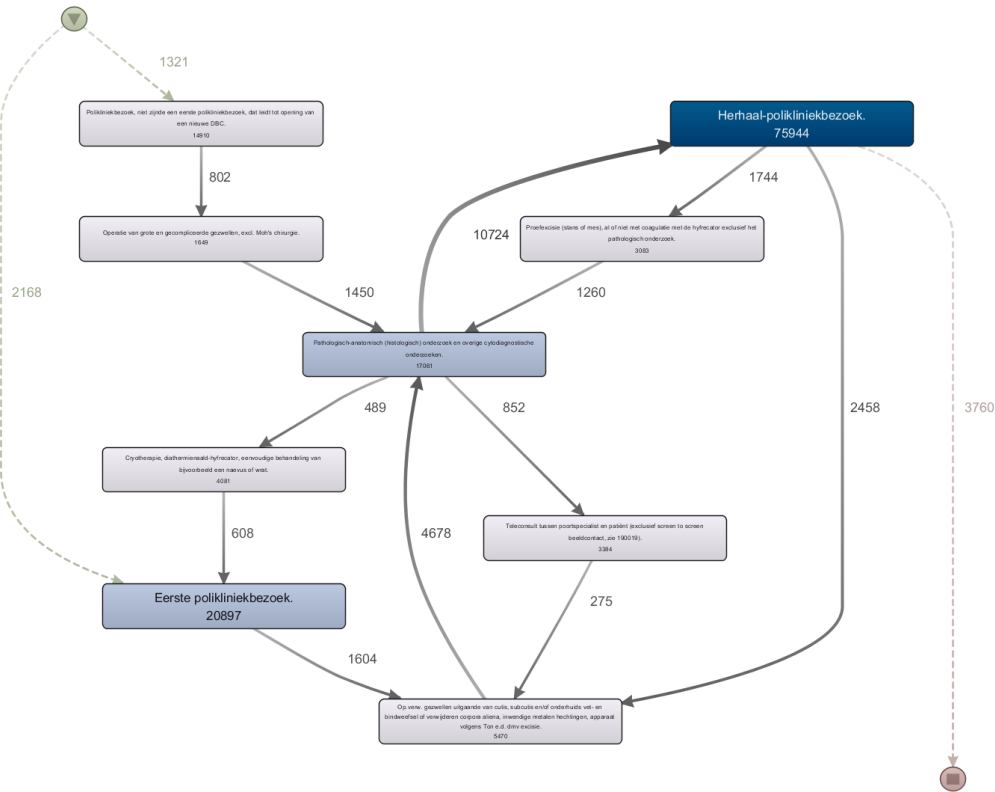


Figure 5.5: Discovered process using process discovery for patients diagnosed with BCC. The log used contains 20% of the most events that the patients were subject to during the period 2009-2015 and are related to the treatment of BCC.

## 5.2.2 Activities in the guidelines vs. activities in the log

The naming of events in the knowledge-based models and the data-based models contain significant differences, both on label as on level of granularity. In order to perform a compliance analysis it would be crucial that the naming is exactly the same in both model and log. Additionally it is often unknown if activities are performed to treat the diagnosis investigated. Taking this into account, it was chosen to adopt the naming of events in the log. The activities are based on care activities as compiled by the Dutch Care authority (“Nederlandse Zorgautoriteit (NZa)”). This independent institution negotiates about care (activities) as well as their quality and price with care providers and health insurances [42]. These activities are used by all hospitals in the Netherlands to invoice treatments based on a diagnosis (“Diagnose Behandel Combinatie (DBC)”). Moreover, the care activities also prove to be the most reliable current description of activities performed in a hospital (audited by the NZa).

A serious issue with the activities is that it is unknown whether an activity that is assessed to be applicable to the investigated diagnosis also is executed to treat that diagnosis. Namely, multiple diagnosis are classified within one financial product (DBC) that a patient can be treated for. Appointments and treatments can intermingle.

To account for as much variation as possible caused by the issues discussed above, events in the log are filtered as discussed in 5.2.1.

Due to the difference in level of granularity of activities in the initial (knowledge-based) models

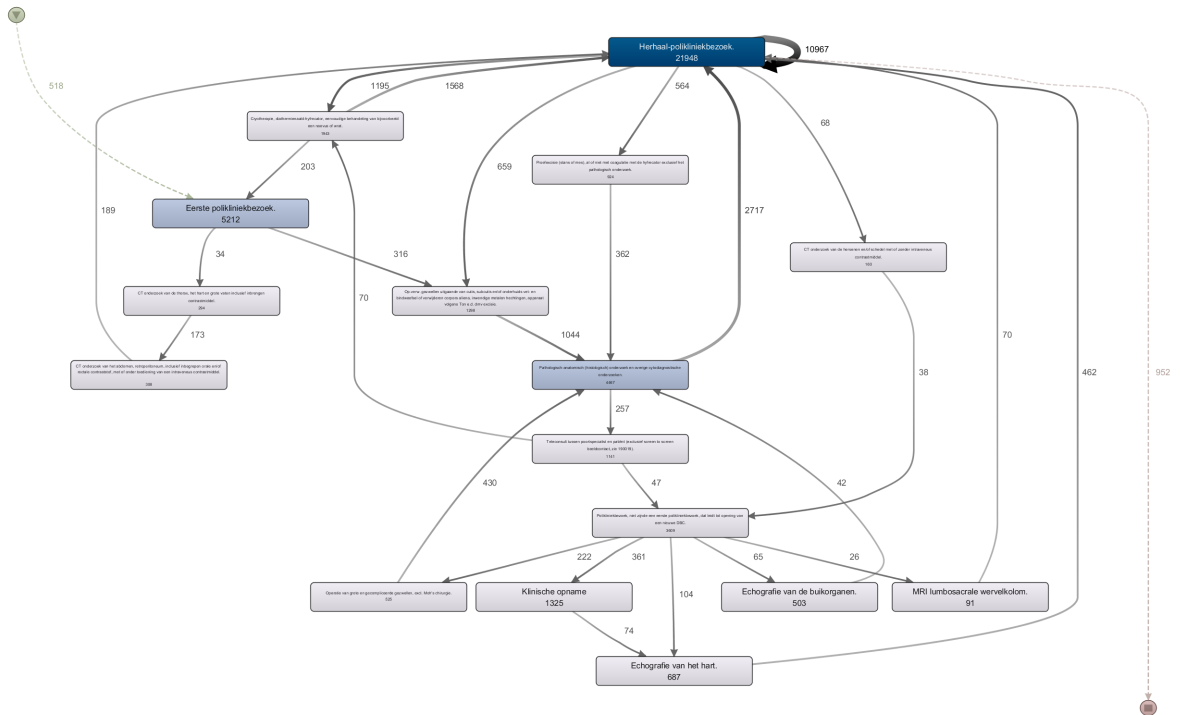


Figure 5.6: Discovered process using process discovery for patients diagnosed with SCC. The log used contains 20% of the most events that the patients were subject to during the period 2009-2015 and are related to the treatment of SCC.

and activities in the log (as shown by the data-based models). Adjustments had to be made to the initial process models (both structure and naming), nevertheless their main structure could be kept. In the remainder of this chapter, first the activities are shown that could be discovered in the log that could on its turn be compared to the initial activities in the guidelines. Afterwards the influence on the process models is shown. In the next chapter it will be discussed whether the data adjusted process models are usable in a compliance analysis.

In order to get a good understanding of the events that should be adopted in the process model we make a comparison between the initial events and events that could identified in the log were made. This comparison was subdivided in the same way as the process models in section 3.2 were subdivided. The adjusted events in the model correspond to the filtered events in the log.

In the tables 5.7, 5.8, 5.9, 5.10, 5.11 and 5.12, the events in the initial model are listed in the right column and the corresponding events in the log are listed in the left column. The descriptions of the events are in Dutch, but when needed, a translation is given between brackets.

## Diagnostics process for patients diagnosed with BCC

Table 5.7: Events in model vs. events in log for guidelines BCC (Diagnostics)

Events guideline	Events log
Perform anamnesis	Eerte polikliniekbezoek (first outpatient appointment)
Perform physical examination	Polikliniekbezoek, niet zijnde het eerste polikliniekbezoek, dat leidt tot opening van een nieuw DBC. (Outpatient appointment, not first, that results in opening of a new DBC <sup>a</sup> )
Request biopsy	–
Take shave biopsy	
Take punch biopsy	Proefexcisie (stans of mes), al of niet met coagulatie met hyfrecator (Biopsy (stans or knife), possibly includes coagulation with hyfrecator)
Take excision biopsy	
Request pathological examination biopsy	–
Examine biopsy (histological examination)	Pathologisch-anatomisch (histologisch) onderzoek en overige cytodiagnostische onderzoeken (Pathological anatomical (histological) examination and other cytodiagnostic examinations)
Classify BCC	Herhaalpolikliniekbezoek (repeating outpatient appointment)
	Screen to screen beeldcontact ter vervanging van een fysiek herhaalconsult (Screen to screen video contact to replace a repeating physical consultation)
	Teleconsult tussen poortspecialist en patient (teleconsultation between specialist and patient)

<sup>a</sup>a DBC (Diagnose Behandel Combinatie (Diagnosis Treatment Combination)) is a concept that is used in the Netherlands to invoice a combination of treatments one is entitled to receive for a certain diagnosis

**Treatment process for patients diagnosed with BCC**

Table 5.8: Events model vs. events in log for guidelines BCC (Treatment)

Events guideline	Events log
Choose treatment	Eerte polikliniekbezoek (first outpatient appointment)  Polikliniekbezoek, niet zijnde het eerste polikliniekbezoek, dat leidt tot opening van een nieuw DBC (Outpatient appointment, not first, that results in opening of a new DBC <sup>a</sup> )
Perform surgical excision	Op.verw. gezwellen uitgaande van cutis, subcutis en/of onderhuids vet- en bindweefsel of verwijderen corpora aliena, inwendige metalen hechtingen, apparaat volgens Ton e.d. dmv excisie (Surgical removal tumors starting from cutis, subcutis and / or subcutaneous fat and connective tissue or remove corpora aliena, internal metal sutures)  Operatie van grote en gecompliceerde gezwellen, excl. Mohs chirurgie (Surgery of large and complicated tumors, excl. Mohs surgery)  Operatie van grote en gecompliceerde gezwellen (Surgery of large and complicated tumors)  Operatieve verwijdering van gezwellen, corpora aliena en dergelijke (Surgical excision tumors, corpora aliena and such)
Perform Mohs micrographic surgery	Operatieve verwijdering van gezwellen door middel van Mohs chirurgie (surgical removal of tumors through Mohs surgery)
Perform radiotherapy	Verstrekking chemo-immunotherapie per infuus of per injectie (Provision chemo-immunotherapy by infusion or by injection)  Verstrekking chemotherapie per infuus of per injectie bij niet-gemetastaseerde tumoren (Provision of chemotherapy by infusion or by injection with non-metastatic tumors)  Verstrekking immunotherapie per infuus of per injectie (excl. desensibilisatie middels immunotherapie bij kinderen, excl. behandeling met methotrexaat (MTX) bij kinderen) (Provision immunotherapy by infusion or by injection (excl. Desensitization through immunotherapy for children, excl. treatment with methotrexate (MTX) for children))
Perform cryosurgery	Cryotherapie, diathermienaald-hyfreacator, eenvoudige behandeling van bijvoorbeeld naevus of wrat (cryotherapy, easy treatment for neavus or wart)
Perform photodynamic therapy	Foto-therapie van chronische huidziekten, al dan niet ondersteund door medicamenteuze fotosensibiliserende therapie, behandeling gedurende de eerste maand (Photo-therapy for chronic skin diseases, possibly supported by photosensitizer drug therapy, treatment during the first month)

<sup>a</sup>a DBC (Diagnose Behandel Combinatie (Diagnosis Treatment Combination)) is a concept that is used in the Netherlands to invoice a combination of treatments one is entitled to receive for a certain diagnosis

	Foto-therapie van chronische huidziekten, al of niet ondersteund door medicamenteuze fotosensibiliserende therapie, behandeling gedurende de volgende elf maanden (Photo-therapy of chronic skin diseases, possibly supported by photosensitizer drug therapy, treatment during the next eleven months)
	Fotodynamische therapie (fotochemische lichttherapie van (pre-)maligniteiten) (Photodynamic therapy (photochemical therapy of (pre-) malignancies))
Perform curettage en cautery	Curettage
Perform local medicinal therapy	Methylaminolevulaat, per toedieningseenheid van 10 mg bij indicaties welke voldoen aan de beleidsregel dure geneesmiddelen ((Methylaminolevulaat, used per unit of 10 mg for indications which meet the policy and rates specialist medical care)
	Methylaminolevulaat, toedieningsvorm creme, per gebruikte eenheid van 10 mg bij indicaties welke voldoen aan de beleidsregel prestaties en tarieven medische specialistische zorg (Methylaminolevulaat, cream, used per unit of 10 mg for indications which meet the policy and rates specialist medical care)
Perform systematic medicinal therapy	–
Request pathological examination biopsy	–
Examine biopsy (histological examination)	Pathologisch-anatomisch (histologisch) onderzoek en overige cytodiagnostische onderzoeken (Pathological anatomical (histological) examination and other cytodiagnostic examinations)

### Follow-up process for patients diagnosed with BCC

Table 5.9: Events in model vs. events in log for guidelines BCC (Follow-up)

Events guideline	Events log
Instructions self-examination	Herhaalpolikliniekbezoek (repeating outpatient appointment)
(full body) Skin check	Teleconsult tussen poortspecialist en patient (teleconsultation between specialist and patient)
	Screen to screen beeldcontact ter vervanging van een fysiek herhaalconsult (Screen to screen video contact to replace a repeating physical consultation)



**Diagnostics process for patients diagnosed with SCC**

Table 5.10: Events in model vs. events in log for guidelines SCC (Diagnostics)

Events guideline	Events log
Perform anamnesis	Herhaalpolikliniekbezoek (repeating outpatient appointment)
Determine size, location and induration of abnormality	Teleconsult tussen poortspecialist en patient (teleconsultation between specialist and patient)
Palpation: determine degree of invasion in underlying tissue	Screen to screen beeldcontact ter vervanging van een fysiek herhaalconsult (Screen to screen video contact to replace a repeating physical consultation)
Request biopsy	–
Take shave biopsy	Proefexcisie (stans of mes), al of niet met coagulatie met hyfrecator (Biopsy (stans or knife), possibly includes coagulation with hyfrecator)
Take punch biopsy	
Take excision biopsy	
Request pathological examination biopsy	–
Examine biopsy (histological examination)	Pathologisch-anatomisch (histologisch) onderzoek en overige cytodiagnostische onderzoeken (Pathological anatomical (histological) examination and other cytodiagnostic examinations)
Examine regional lymph nodes	
Inspect entire skin suitable for patient's sun damage and skin type	Herhaalpolikliniekbezoek (repeating outpatient appointment)
Determine prognostic grouping (TMN level)	
Request palpation with cytologic puncture for suspected lymph nodes	–
Conduct palpation with cytologic puncture for suspected lymph nodes	Pathologisch-anatomisch (histologisch) onderzoek en overige cytodiagnostische onderzoeken (Pathological anatomical (histological) examination and other cytodiagnostic examinations)
	Eenvoudig biopt, eenvoudige cytologie (excl. bepalingen op de aanwezigheid van micro-organismen) (Simple biopsy, simple cytology (excl. Examination for the presence of micro-organisms))
	Biopt, matig complexe cytologie (Biopsy, moderately complex cytology)
	Naaldbiopt, complexe cytologische punctie (punch biopsy, complex cytologic puncture)

Request echographic screening for unsuspected regional lymph nodes	–
Perform echographic screening for unsuspected regional lymph nodes	Echografie onderste extremiteit
	Echografie van de schildklier en/of hals
	Echografie van de bovenste extremiteiten
	Echografie van het bewegingsapparaat
	Echografie van de buikorganen
	Echografie van het hart en/of de thorax
	Echografie van de schedel
	Echografie van het hart
	Echografie van mamma
Request echographic screening of neck including parotid area with cytologic puncture for suspected lymph nodes (> 5 mm)	–
Perform echographic screening of neck including parotid area with cytologic puncture for suspected lymph nodes (> 5 mm)	Diagnostische puncties van niet palpabele afwijkingen of organen, onder echografische controle (Diagnostic puncture of non palpable abnormalities or organs with echographic guidance)
Perform additional diagnostics (MRI and/or CT with UV-Contrast)	MRI thorax(wand), mamma en mediastinum
	MRI thoracale wervelkolom
	MRI schouder(s)/bovenste extremiteit(en)
	MRI lumbosacrale wervelkolom
	MRI heup(en)/ onderste extremiteit(en)
	MRI hersenen - standaard
	MRI hersenen - met contrast
	MRI hersenen
	MRI cervicale wervelkolom en/of hals inclusief craniovertebrale overgang
	MRI bekken
	MRI achterste schedelgroeve
	MRI abdomen
	CT onderzoek van de wervelkolom
	CT onderzoek van de aangezichtsschedel, met of zonder intraveneus contrast

CT onderzoek van de hersenen en/of schedel met of zonder intraveneus contrastmiddel

CT onderzoek van de bovenste extremiteiten, met of zonder intraveneus contrast

CT onderzoek van de onderste extremiteiten, met of zonder intraveneus contrast

CT van het bekken inclusief inbrengen orale en/of rectale contraststof. Met of zonder toediening van een intraveneus contrastmiddel

CT onderzoek van het abdomen, retroperitoneum, inclusief inbegrepen orale en/of rectale contraststof, met of onder toediening van een intraveneus contrastmiddel

CT onderzoek van de thorax, het hart en grote vaten inclusief inbrengen contrastmiddel

Request CT-Thorax	–
Perform CT-Thorax	CT onderzoek van de thorax, het hart en grote vaten inclusief inbrengen contrastmiddel
Request PET-CT	–
Perform PET-CT	PET WB (whole body), oncologie.

**Treatment process for patients diagnosed with SCC**

Table 5.11: Events in model vs. events in log for guidelines SCC (Treatment)

Events guideline	Events log
Choose treatment	Herhaalpolikliniekbezoek (repeating outpatient appointment)  Teleconsult tussen poortspecialist en patient (teleconsultation between specialist and patient)  Screen to screen beeldcontact ter vervanging van een fysiek herhaalconsult (Screen to screen video contact to replace a repeating physical consultation)
Perform surgical excision	Op.verw. gezwellen uitgaande van cutis, subcutis en/of onderhuids vet- en bindweefsel of verwijderen corpora aliena, inwendige metalen hechtingen, apparaat volgens Ton e.d. dmv excisie (Surgical removal tumors starting from cutis, subcutis and / or subcutaneous fat and connective tissue or remove corpora aliena, internal metal sutures)
	Operatie van grote en gecompliceerde gezwellen, excl. Mohs chirurgie (Surgery of large and complicated tumors, excl. Mohs surgery)
	Operatie van grote en gecompliceerde gezwellen (Surgery of large and complicated tumors)
	Operatieve verwijdering van gezwellen, corpora aliena en dergelijke (Surgical excision tumors, corpora aliena and such)
Perform Mohs micrographic surgery	Operatieve verwijdering van gezwellen door middel van Mohs chirurgie (surgical removal of tumors through Mohs surgery)
Perform radiotherapy	Verstrekking chemo-immunotherapie per infuus of per injectie (Provision chemo-immunotherapy by infusion or by injection)
	Verstrekking chemotherapie per infuus of per injectie bij niet-gemetastaseerde tumoren (Provision of chemotherapy by infusion or by injection with non-metastatic tumors)
	Verstrekking immunotherapie per infuus of per injectie (excl. desensibilisatie middels immunotherapie bij kinderen, excl. behandeling met methotrexaat (MTX) bij kinderen) (Provision immunotherapy by infusion or by injection (excl. Desensitization through immunotherapy for children, excl. treatment with methotrexate (MTX) for children))
Perform cryosurgery	Cryotherapie, diathermienaald-hyfreator, eenvoudige behandeling van bijvoorbeeld naevus of wrat (cryotherapy, easy treatment for neavus or wart)
Request pathological examination biopsy	–
Examine biopsy (histological examination)	Pathologisch-anatomisch (histologisch) onderzoek en overige cytodiagnostische onderzoeken (Pathological anatomical (histological) examination and other cytodiagnostic examinations)

**Follow-up process for patients diagnosed with SCC**

Table 5.12: Events in model vs. events in log for guidelines SCC (Follow-up)

Events guideline	Events log
Instructions self-examination	Herhaalpolikliniekbezoek (repeating outpatient appointment)
(full body) Skin check	Teleconsult tussen poortspecialist en patient (teleconsultation between specialist and patient)
	Screen to screen beeldcontact ter vervanging van een fysiek herhaalconsult (Screen to screen video contact to replace a repeating physical consultation)

Ideally it would be possible to use the initial models that followed from the clinical guidelines. These models reflect an interpretation of the exact level of detail described in the current guidelines.

**5.2.3 Data adjusted process models**

Because the naming in the log is adopted, it ensured to change the events in the process models, while changing as less as possible to their original structure. A comparison of all activities that remain in the log after filtering per sub process are shown in tables 5.7, 5.8, 5.9 (BCC) and 5.10, 5.11, 5.12 (SCC).

The main differences is that due to different levels of granularity more choices in the process model have to be allowed (e.g., for an event CT scan a lot of different events could be found in the event log). On the other hand we cannot find detailed information of what happened during appointments form the events in the event log cannot be found.

Models that were constructed with the new event labels are shown in Appendix A, section A.3.

## Chapter 6

# Case study results & Evaluation

### 6.1 Compliance analysis

Compliance analysis of treatment behavior is essential for healthcare organizations in bringing clinical guidelines to a higher level of maturity. In [30] an example is shown for compliance analysis in a clinical context. The authors apply compliance checking to the clinical pathway for unstable angina. The compliance checking technique used in this study (explained in section 2.1) is different in essence, but the goal of checking compliance stays the same.

As input for the compliance analysis process models are needed that reflect the original paper-based clinical guidelines for BCC and SCC. At the same time the descriptions for the activities in the process models should correspond to the activity descriptions in the HIS. The initial process models that reflect the clinical guidelines are explained in section 3.2. In chapter 4 it was studied whether the initial process models are able to match activities in the HIS. The second input for the compliance analysis are the logs that contain all activities for patients diagnosed with BCC (for the BCC case) and diagnosed with SCC (for the SCC case). It should be taken into account that the activities in the log could also contain activities that are performed for the purpose of different diagnoses. In chapter chapter 4 it was studied how to cope with this fact and how to filter the log to account for as much variability as possible.

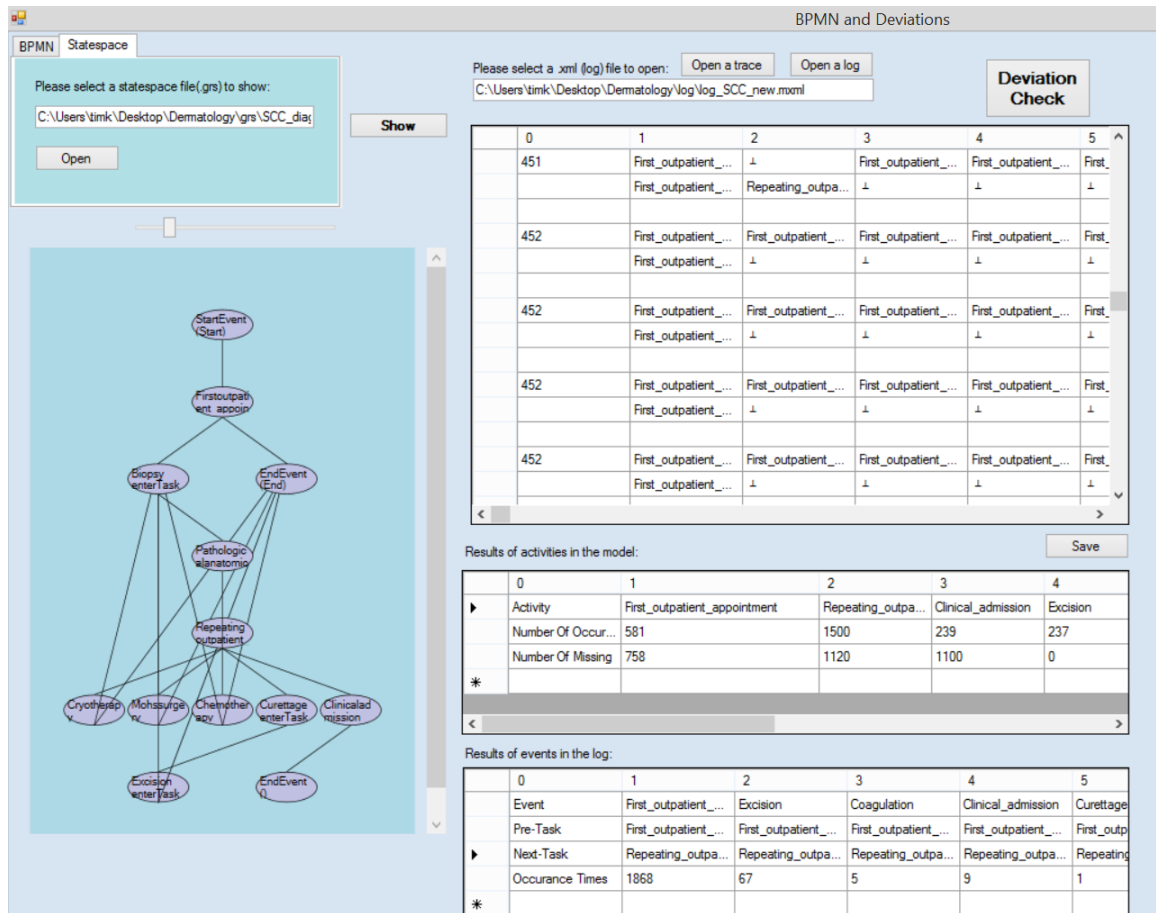


Figure 6.1: Compliance tool output

### 6.1.1 Compliance analysis results BCC

In table 6.1 and table 6.2 the results of the compliance analysis for the events in the guidelines of BCC are summarized. The corresponding model for table 6.1 can be seen in figure A.7. The corresponding model for table 6.2 can be seen in figure A.8. Note that the message event control flow is replaced by standard connections, because the tool did not support this structure in a proper way. This alternative structure still behaves approximately the same.

Table 6.1: Results of activities in the model for guidelines BCC (diagnostics)

<b>Activity</b>	<b>Number of occurrences</b> synchronous move	<b>Number of Missing</b> move on model
Eerstepolikliniekbezoek	2320	2204
Herhaalpolikliniekbezoek	2980	2819
PolikliniekbezoeknietzijndeheteerstepolikliniekbezoekdatleidttotopeningvaneennieuwDBC	1345	0
Proefexcisiemetstansofmesalofnietmetcoagulatiemethylfrecatorexclusiefhetpathologischonderzoek	149	243
Pathologisch anatomisch onderzoek overigecytologische onderzoeken	392	0
Teleconsulttussenpoortspecialisten-patientexclusiefscreentootscreenbeeldcontactzie190019	158	0

The three most frequent patterns in the log:

- Number of occurrences: 922
  - Event: Pathologisch anatomisch onderzoek overigecytologische onderzoeken
  - Pre-task: Eerstepolikliniekbezoek
  - Next task: Herhaalpolikliniekbezoek
- Number of occurrences: 336
  - Event: Pathologisch anatomisch onderzoek overigecytologische onderzoeken
  - Pre-task: PolikliniekbezoeknietzijndeheteerstepolikliniekbezoekdatleidttotopeningvaneennieuwDBC
  - Next task: Herhaalpolikliniekbezoek
- Number of occurrences: 213
  - Event: Eerstepolikliniekbezoek
  - Pre-task: Eerstepolikliniekbezoek
  - Next task: Herhaalpolikliniekbezoek



Table 6.2: Results of activities in the model for guidelines BCC (Treatment + follow-up)

Activity	Number of occurrences synchronous move	Number of Miss- ing move on model
Fototherapie van chronische huidaandoeningen die niet- ondersteund door medicamenteuze fotosensibiliserende- therapiebehandeling gedurende de eerste maand	0	5814
Pathologisch anatomisch histologisch onderzoek- en overige cytodiagnostische onderzoeken	214	5655
Herhaal polikliniekbezoek	192	0
Fotodynamisch therapiefotochemisch licht- therapie van premaligniteiten	3	0
Methylaminolevulaat toedieningseenheid van 10 mg bij- indicaties welke voldoen aan de beleidsregel- dure geneesmiddelen	3	0
Fotodynamisch therapiefotochemisch licht- therapie van premaligniteiten	11	0
Methylaminolevulaat toedieningsvorm creme pergebruikt- eenheid van 10 mg bij indicaties welke voldoen aan de- beleidsregel prestatiesentarieven medische- specialistische zorg	10	0
Operatie van grote en gecompliceerde gezwellen	6	0
Operatieve verwijdering van gezwellen corpora aliena en- dergelijke uitgaande van of zich bevindende in- dieper liggende structuren dan in code 038911 is omschreven	25	0
Dermatologische behandeling met laser tot ongeveer 12- van het lichaamsoppervlak 4x5 cm onder lokale anesthesie	8	0
Dermatologische behandeling met laser groter dan 1 van- het lichaamsoppervlak onder algehele anesthesie	1	0
Teleconsult tussen poortspecialisten patiënt exclusief- screentot screenbeeld contact zie 190019	3	0
Dermatologische behandeling met laser tussen 12- en 1 van het lichaamsoppervlak onder lokale of algehele anesthesie	1	0

The three most frequent patterns in the log:

- Number of occurrences: 60
  - Event: Opverwege zwellenuitgaande van cutis subcutis en of onderhuidsveten bindweefsel
  - Event: of verwijdering corpora aliena inwendig metalen hechtingen apparaat volgens-  
Tone dmv excisie
  - Pre-task: Herhaal polikliniekbezoek
  - Next task: Fototherapie van chronische huidaandoeningen die niet-  
ondersteund door medicamenteuze-  
fotosensibiliserende therapiebehandeling gedurende de eerste maand
- Number of occurrences: 59
  - Event: Herhaal polikliniekbezoek
  - Pre-task: Herhaal polikliniekbezoek
  - Next task: Fototherapie van chronische huidaandoeningen die niet-  
ondersteund door medicamenteuze-  
fotosensibiliserende therapiebehandeling gedurende de eerste maand

- Number of occurrences: 13
  - Event: Pathologisch anatomisch histologisch onderzoek overige cytodiagnostische onderzoeken
  - Pre-task: Herhaalpolikliniekbezoek
  - Next task: Fototherapie van chronische huidziekten al dan niet ondersteund door medicamenteuze fotosensibiliserende therapie behandeling gedurende eerste maand

### 6.1.2 Compliance analysis results SCC

Because usage of the data adjusted models as shown in figure A.9, figure A.10 and figure A.11 proved to result in too complex structures for our tool to compute a statespace, we aggregated the events in the log and corresponding activities in the model (to get an idea of their cohesion we made new process discovery plots, see figure A.5 and A.6). The following levels were remained in the log and model (The R script used to aggregated the activities in the log can be found in appendix E.):

- First\_outpatient\_appointment
- Repeating\_outpatient\_appointment
- Clinical\_admission
- Biopsy
- Pathological\_anatomical\_histological\_examination
- Excision
- Cryotherapy
- Mohs\_surgery
- Curettage
- Chemotherapy
- Echography
- MRI
- CT\_examination
- Echographic\_screening\_of\_diagnostic\_biopsy\_nonpalpable
- CT\_examination\_thorax

In table 6.3 and table 6.4 the results of the compliance analysis for the events in the guidelines of SCC are summarized. The corresponding model for table 6.3 can be seen in figure A.12. The corresponding model for table 6.4 can be seen in figure A.13.

Table 6.3: Results of activities in the model for guidelines SCC (excl. treatment)

<b>Activity</b>	<b>Number of occurrences</b> synchronous move	<b>Number of Missing</b> move on model
First_outpatient_appointment	583	756
Repeating_outpatient_appointment	246	1150
Clinical_admission	61	1278
Biopsy	28	1433
Pathological_anatomical_histological_examination	196	0
Echography	123	69
MRI	57	0
CT_examination	80	0
Echographic_screening_of_diagnostic_biopsy_nonpalpable	0	2
CT_examination_thorax	2	0

The three most frequent patterns in the log:

- Number of occurrences: 2233
  - Event: First\_outpatient\_appointment
  - Pre-task: Pathological\_anatomical\_histological\_examination
  - Next task: Echography
- Number of occurrences: 801
  - Event: First\_outpatient\_appointment
  - Pre-task: Echography
  - Next task: Repeating\_outpatient\_appointment
- Number of occurrences: 441
  - Event: First\_outpatient\_appointment
  - Pre-task: MRI
  - Next task: Repeating\_outpatient\_appointment

Table 6.4: Results of activities in the model for guidelines SCC (excl. additional diagnostics)

<b>Activity</b>	<b>Number of occurrences</b> synchronous move	<b>Number of Missing</b> move on model
First_outpatient_appointment	581	758
Repeating_outpatient_appointment	1500	1120
Clinical_admission	239	1100
Excision	237	0
Pathological_anatomical_histological_examination	594	14
Cryotherapy	158	0
Biopsy	76	0
Mohs_surgery	1	0
Curettage	1	0
Chemotherapy	3	0

The three most frequent patterns in the log:

- Number of occurrences: 6326
  - Event: First\_outpatient\_appointment
  - Pre-task: Pathological\_anatomical\_histological\_examination
  - Next task: Repeating\_outpatient\_appointment
- Number of occurrences: 2191
  - Event: First\_outpatient\_appointment
  - Pre-task: Repeating\_outpatient\_appointment
  - Next task: Clinical\_admission
- Number of occurrences: 1868
  - First\_outpatient\_appointment
  - First\_outpatient\_appointment
  - Repeating\_outpatient\_appointment

## 6.2 Evaluation of compliance results

What do the results in section 6.1.1 and 6.1.2 tell us.

As can be seen in the analysis above, aggregation of the activities for the BCC process was not chosen. For the SCC process this could not be avoided, because the models became too complex to be usable in the software tool. In general, our vision is to limit the use of aggregation with limited clinical knowledge. As can be seen in Appendix E, interpretation of every individual clinical activity is needed in order to aggregate them.

The results above show compliance metrics of the execution of activities in the data-adjusted knowledge-based process models that represent the clinical guidelines for the two diagnoses (BCC and SCC). Algorithm 1 optimizes the routing of real-life patient behavior (captured in an event log) through the process models in figures A.7, A.8, A.12 and A.13.

The tables 6.1, 6.2 (for the BCC patient population), 6.4 and 6.3 (for the SCC patient population) give information about the patients' processes in comparison to the guidelines. During

a synchronous or compliant move, the patient behavior (as registered in the event log) is allowed by the guidelines (as registered in the BPMN process model) and expected by the algorithm to happen. When the algorithm identifies a move on model, an activity from the calculated path through the process model is missing in the log.

While interpreting the compliance results above one should take into account problems that currently exist and could have influenced the results.

Three possible issues with the event log data:

1. *Uncertainty whether activities are performed to treat the diagnosis investigated.* This is one of the main concerns. As shown in section 4.5 the average population in that is diagnosed with BCC or SCC can be described as elderly people for whom this diagnosis is often not the only reason to come to a hospital. Although much of the irrelevant activities have been filtered out as described in subsection 5.2.1, activities such as “outpatient appointment” could relate to different diseases. When activities are introduced in the event log that are not part of the process investigated it will likely influence the results.
2. *Registration errors.* Reliability of registrations are dependent on declarations by hospitals. Usage of a different declaration code for hospitals does not have to make a big difference. Because declaration codes are used to invoice, the biggest risk is that a wrong amount for a treatment will be paid. In this analysis on the other hand, it will influence the path taken (e.g., first outpatient appointment vs. repeating outpatient appointment).
3. *During aggregation and filtering, clinical activity descriptions had to be interpreted* (room for error).

Two issues with the current version of the tool that supports the compliance analysis technique we used in this thesis:

1. Difficulty with large amounts of data.
2. Inability to process complex model structures (large amount of events or possible paths).
3. This study is one of the first real applications of the tool.

Due to remarks about the tool addressed above, many variants of models needed to be tested to get usable results.

Taking this into account, some interesting observations are discussed below:

- table 6.1:

- Many missing “Eerstepolikliniekbezoeken” (First outpatient appointments) and “Herherhaalpolikliniekbezoeken” (Repeating outpatient appointments). This looks strange and can be explained by the data issues explained above. Many of the patients in the investigated patient population are treated for multiple diagnoses at the same time. The missing activities could have taken place at a different hospital, a different point in time or different specialism. Another possible explanation is that declaration codes have been used that were filtered out of our dataset.
- Missing “Proefexcisie met stans of mes al of niet met coagulatie met hylfrecator exclusief het pathologischonderzoek” (biopsy). This could be explained by the fact that a histological examination can also take place after a treatment. In that case no biopsy is required.

- table 6.2

- These results look very strange. According to the output the tool, nearly all patients are supposed to be treated with a “Fototherapie van chronische huidziekten al dan niet ondersteund door medicamenteuze fotosensibiliserendetherapie behandeling gedurende de eerste maand” and “Pathologisch anatomisch histologisch onderzoek en overige cyto-diagnostischeonderzoeken”. This is not what is intended by the used model shown in figure A.8. Additionally this pattern could not be recognized in the discovery models. Unfortunately this was the best output that could be derived from the tool. After numerous trail, we assume that the model especially experienced difficulty with the size of the dataset used for the logs.
- table 6.3
  - The missing “First\_outpatient\_appointments”, “Repeating\_outpatient\_appointments”. These results are similar as those in table 6.1 and the explanation is applicable here as well.
  - Missing “Clinical\_admissions”. This is strange, because according to the model this activity is not compulsory. An similar situation is show for “Pathologisch anatomisch histologisch onderzoek en overige cytodiagnostischeonderzoeken” in table 6.2. It could be that the current tool does not interpret this control flow properly. The same explanation is applicable to the missing activities: Clinical\_admission, Echography, Echographic\_screening\_of\_diagnostic\_biopsy\_nonpalpable and partly Biopsy.
  - An different part of the missing Biopsies (196 - 28) can be explained by the fact that histological examinations can also take place after the removal of a tumor during treatment in stead of a biopsy.
- table 6.4
  - Many missing “First\_outpatient\_appointments”, “Repeating\_outpatient\_appointments”. These results are similar as those in tables 6.1 and 6.3 and the explanation is applicable here as well.
  - Missing “Clinical\_admissions”. This is strange, because according to the model this activity is not compulsory. An similar situation is show for “Pathologisch anatomisch histologisch onderzoek en overige cytodiagnostischeonderzoeken” in table 6.2. It could be that the current tool does not interpret this control flow properly.

If we show the results in percentages, For the BCC case: 42% of all events that are expected to happen according to the guidelines are skipped during diagnosis and 96% of all events that are expected to happen according to the guidelines are skipped during treatment/follow-up.

For the SCC case: 47% of all events that are expected to happen according to the guidelines are skipped during diagnosis and 44% of all events that are expected to happen according to the guidelines are skipped during treatment/follow-up.

Due to the issues addressed earlier these results should be interpreted carefully.

### 6.3 Evaluation of the approach to evaluate clinical guidelines for NMSCs

During the development and application of this study’s approach, ways to use a theoretical approach in a complex practical setting had to be explored. As a result, diagnostics about patient behavior in combination with the guidelines for NMSCs were found. Unfortunately the reliability of the results is questionable due to problems with the data and immaturity of the used software tool.

In this section two different parts of this approach will be evaluated:

- The approach chosen in this study
- The current BPMN compliance analysis tool

The approach chosen in this study consisted of three phases:

- Translating the clinical guidelines for NMSC into BPMN 2.0 process models
  - Was the first research question answered? Did we show a good translation of the clinical guidelines into BPMN process models?

In chapter 3, a careful translation of the clinical guidelines has been made. To look at each phase of the guidelines as careful as possible the guidelines have been split into small understandable parts. In order to understand the clinical guidelines, medical literature was used. Additionally, clinical professionals at the Bravis hospital assisted. To optimize the control flow of the process models, the 7PMGs [40] were used as guidance. These guidelines are summarized in table 3.3. The knowledge-based models developed in this section are influenced by interpretation and contain design choices. Different models could execute approximately the same behavior as described in the text based guidelines. During the development of these models there was no knowledge about the data yet (e.g., level of granularity). In order to perform a compliance analysis it would be of importance that models and data are comparable. The disadvantage of this approach is that it could have been expected that the initial models had to be redesigned later. In short, the initial models in chapter 3 give a good translation of the clinical guidelines for BCC and SCC, although these models were not usable during the compliance analysis. In order to execute a compliance check comprised adjustments had to be made to the initial models.

- Study data challenges in HISs nowadays in order to do a data driven analysis of patients and processes.
  - Was the second research question answered? Did we show whether the data currently available electronically is suitable for a compliance analysis?

As we showed in this study it is not straightforward to use electronically available hospital data for analyses. We have been involved in the whole data collection and preparation process. Meanwhile it was possible to address the specific difficulties that were faced. These difficulties are specific for the context investigated and have not been investigated (often) before. An attempt was made to use the data currently available in an hospital in an compliance analysis. Although we were able to find solutions for a lot of issues in the data, unfortunately many inconsistencies in the results could be explained by currently unsolvable issues in the data. In short, interesting data challenges have been discussed and solutions have been proposed. The data currently available for the processes studied could not be used for a reliable compliance analysis without individually examining every patient case.

- Apply compliance analysis to real-life clinical processes for NMSCs and evaluate its results.
  - Was the third research question answered? Did we show a good attempt to apply a compliance analysis and give a reliable evaluation of its results?

Although not all issues in the data could be solved, a compliance analysis was performed and its results evaluated in chapter 6. Before the tool could show results, the initial process models from chapter 3 had to be adjusted according to the data as discussed

in the chapter before. Everything was done to compromise the models in such a way that the models could be kept as close to the real clinical guidelines while minimizing its complexity. Unfortunately, final results raise questions that can be traced back to data issues and immaturity of the used software tool.

In hindsight, it would have been better to investigate a small part of the guidelines with a limited number of patients, this could have enabled us to solve the “unsolvable data issues” experienced manually. Additionally, the demands of our simple software tool would have been limited.

## 6.4 Evaluation of BPMN compliance analysis technology

At this point in time the tool applied in this study is not ready to be evaluated based on usability and acceptability using for example the technology acceptance model (TAM) [56]. As we showed in this study, there are numerous situation in which the tool shows results that raise questions. In this study the focus has been on the application of a new compliance analysis technique supported by this new tool. During this process it was experienced that a reliable application of compliance checking in BPMN for the current version of the tool is not possible for the complex context investigated.





## Chapter 7

# Conclusion & Recommendations

There is still a lot to win to make data driven evaluation of clinical guidelines better. In this thesis is shown that it is possible to check the compliance of processes that are based on and scientifically supported by clinical guidelines. Due to existing issues in the data and lack of testing of the software tool, the current results have not much value in practice.

The goal of this study was to “develop and test an approach in order to evaluate patients’ adherence to clinical guidelines for NMSCs”. At the start of this study it was uncertain whether the available data would enable us to present usable compliance results that could be used in practice. Of course, providing the clinical professionals with usable results has been a motivation during this study.

The first research (sub)question has been concerned with the translation of the clinical guidelines for NMSCs into process models. Descriptions in paper-based guidelines are sometimes vague (e.g., “..we advise to..”, “..one or more treatments..”). The BPMN language can deal with these descriptions. The disadvantage: the vague descriptions allow freedom in the models which decrease the value of compliance checking. Additional medical research is needed to make the guidelines more prescriptive. This would increase the value of compliance checking. An alternative is to study behavior of patients and relate this to negative treatment outcomes. In appendix C, data mining techniques are discussed that would make this possible. In the current study, reliable patient behavior could not be reconstructed using the data discussed in chapter 4. Modeling clinical processes for a compliance analysis based on knowledge written down in clinical guidelines will almost certainly lead to activities that do not match the events in the data (as showed in chapter 5). Future research could consider to start with models discovered from data and keep only those paths that match the clinical guidelines. In order to discover reliable models in a clinical environment, improvements have to made to current process discovery algorithms [34].

The second research (sub)question has been considered with the data that is currently available in hospitals electronically. Thorough understanding of database structure needed for data extraction. Most of this knowledge is owned by the supplier of the hospital’s datawarehouse (ChipSoft in this case). Currently the Bravis hospital (and probably most hospitals) do not invest in improving their understanding of its datawarehouse in order to optimize their usage of it. Much knowledge to clean the data for analyses is available in a hospital, because it requires domain specific knowledge. Moreover, in order to improve reliability of recurrency classification for NMSCs better registrations are necessary for the location of tumors. In our approach we were able to give an idea about the recurrency of tumors for the investigated patients. Suppliers like ChipSoft are constantly improving their software to make data registration easier and react to other demands from hospitals. Much of the issues experiences in this thesis are acted upon nowadays. In should be taken into account that this study looks back data five years up to ago.

The third and last research (sub)question has been considered with the application of compliance checking of BPMN process models, using the technique developed by [62]. After the case study was completed, it could be concluded that the method was applicable in practice. Unfor-

Unfortunately the data currently available combined with immaturity of compliance checking software was not able to give reliable results for the complex situation investigated. Future research should focus on using this technique on less complicated processes and try to relate real behavior to treatment outcomes. In this study it was not valuable to relate questionable compliance results with treatment outcomes.

More general remarks that should be considered in future research are:

First, much communication in hospitals in current hospitals is done by digital orders. This behavior can be described in BPMN 2.0 with message events. The current software in combination with the data is not able to adequately check the compliance of these digital orders (and therefore more complex structures). Messages are not registered as such in the event logs, therefore the software should account for this. If a solution can be found in the future, it will be able to benefit even more from the benefits of the BPMN process modeling language over for example Petri nets. Additionally, it seems logical to make the technique used in this study compatible with process mining research. Think for example of integrating the technique in ProM. Mutual benefits could arise (log preparation in XES or MXML). This prevents researchers from reinventing the wheel. An example of non-compatibility: during this study the standard MXML format that is used as import for ProM was used, however by default event types are named: "complete" and our tool needs the event types to be named "start".

Most of all, it is important to keep applying data and process mining techniques in a real-life healthcare environment. Benefits that arise include: medical professionals that become more aware of the possibilities of data. On the other side could researchers that develop new algorithms for healthcare applications experience its complexity.

# Bibliography

- [1] W.M.P van der Aalst. Patterns and XPD L: A critical evaluation of the XML process definition language. *BPM Center Report BPM-03-09*, *BPMcenter.org*, pages 1–30, 2003. 11
- [2] W.M.P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer Verlag, 2011. 5, 6, 9, 45
- [3] W.M.P. van der Aalst, A. Adriansyah, and B. van Dongen. Conformance checking of processes based on monitoring real behavior. *Information Systems*, 33(1):64–95, 2008. 5, 6
- [4] W.M.P. van der Aalst, A. Adriansyah, and B. van Dongen. Replaying history on process models for conformance checking and performance analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(2):182–192, 2012. 8, 9
- [5] A. Adriansyah, B.F. van Dongen, and W.M.P. van der Aalst. Conformance checking using cost-based fitness analysis. In *Enterprise Distributed Object Computing Conference (EDOC), 2011 15th IEEE International*, pages 55–64, 2011. 9
- [6] C.C. Aggarwal. *An Introduction to Frequent Pattern Mining*, book section 1, pages 1–17. Springer International Publishing, 2014. 94
- [7] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994. 94
- [8] T. Allweyer. *BPMN 2.0: Introduction to the Standard for Business Process Modeling*. BoD-Books on Demand, 2011. 6
- [9] C. C. Blackmore. Clinical prediction rules in trauma imaging: Who, how, and why? *Radiology*, 235(2):371–374, 2005. 92
- [10] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 94
- [11] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and regression trees*. CRC press, 1984. 93, 94
- [12] R.J. Brooks and A.M. Tobias. Choosing the best model: Level of detail, complexity, and model performance. *Mathematical and computer modelling*, 24(4):1–14, 1996. 25
- [13] J.C.A.M. Buijs, B.F. van Dongen, and W.M.P. van der Aalst. Quality dimensions in process discovery: The importance of fitness, precision, generalization and simplicity. *International Journal of Cooperative Information Systems*, 23(01), 2014. 6
- [14] M. Casparie, A.T.M.G. Tiebosch, G. Burger, H. Blauwgeers, A. Van de Pol, J.H.J.M. van Krieken, and G.A. Meijer. Pathology databanking and biobanking in the Netherlands, a central role for PALGA, the nationwide histopathology and cytopathology data network and archive. *Analytical Cellular Pathology*, 29(1):19–24, 2007. 28

- [15] E. de Vries, H. van der Rhee, and J. Coebergh. Trends, oorzaken, aanpak en gevolgen van de huidkankerepidemie in Nederland en Europa. *Ned Tijdschr Geneeskd.*, 150:1108–15, 2006. 15, 36
- [16] R. Dechter and J. Pearl. Generalized best-1st search strategies and the optimality of A. *Journal of the ACM*, 32(3):505–536, 1985. 9
- [17] B. van Dongen. *Process Mining and Verification*. PhD thesis, Eindhoven University of Technology, July 2007. 5
- [18] A. Dwivedi, R.K. Bali, A.E. James, R.N.G. Naguib, and D. Johnston. Merger of knowledge management and information technology in healthcare: opportunities and challenges. In *Electrical and Computer Engineering, 2002. IEEE CCECE 2002.*, volume 2, pages 1194–1199 vol.2, 2002. 1
- [19] D.M. Eddy. Practice policies: Where do they come from? *JAMA*, 263(9):1265–1275, 1990. 13
- [20] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996. 92
- [21] S.C. Flohil, S. Koljenovic, E.R. de Haas, L.I. Overbeek, E. de Vries, and T. Nijsten. Cumulative risks and rates of subsequent basal cell carcinomas in the Netherlands. *Br J Dermatol*, 165(4):874–81, 2011. 18
- [22] E. Frank and I. H. Witten. Generating accurate rule sets without global optimization. In J. Shavlik, editor, *Fifteenth International Conference on Machine Learning*, pages 144–151. Morgan Kaufmann, 1998. 94
- [23] S. van der Geer-Rutten. *Disease Management for Chronic Skin Cancer*. PhD thesis, Erasmus University Rotterdam, April 2012. 15, 18
- [24] P.M.E. van Gorp and R. Dijkman. A visual token-based formalization of BPMN 2.0 based on in-place transformations. *Information and Software Technology*, 55(2):365–394, 2013. 6
- [25] M. Gray. *Evidence-based health care and public health: how to make decisions about health services and public health*. Elsevier Health Sciences, 2014. 13
- [26] C.W. Günther and W.M.P. van der Aalst. Fuzzy mining–adaptive process simplification based on multi-perspective metrics. In *Business Process Management*, pages 328–343. Springer, 2007. 30, 45
- [27] R. C. Hawkins. The evidence based medicine approach to diagnostic testing: practicalities and limitations. *Clinical Biochemist Reviews*, 26(2):7–18, 2005. 1, 13
- [28] R.E. Herzlinger. Why innovation in health care is so hard. *Harvard business review*, 84(5):58, 2006. 1
- [29] K. Hornik, C. Buchta, and A. Zeileis. Open-source machine learning: R meets Weka. *Computational Statistics*, 24(2):225–232, 2009. 94
- [30] Z. Huang, Y. Bao, W. Dong, X. Lu, and H. Duan. Online treatment compliance checking for clinical pathways. *Journal of Medical Systems*, 38(10):1–14, 2014. 57
- [31] Z. Huang, X. Lu, and H. Duan. Latent treatment pattern discovery for clinical processes. *Journal of Medical Systems*, 37(2):1–10, 2013. 6, 30
- [32] Z. Huang, X. Lu, and H. Duan. *Similarity Measuring between Patient Traces for Clinical Pathway Analysis*, volume 7885 of *Lecture Notes in Computer Science*, book section 38, pages 268–272. Springer Berlin Heidelberg, 2013. 6, 30

- 
- [33] Z. Huang, X. Lu, H. Duan, and W. Fan. Summarizing clinical pathways from event logs. *Journal of Biomedical Informatics*, 46(1):111–127, 2013. 6, 30
- [34] U. Kaymak, R. Mans, T. van de Steeg, and M. Dierks. On process mining in health care. In *Systems, Man, and Cybernetics (SMC)*, pages 1859–1864, Oct 2012. 5, 45, 46, 69
- [35] M. Kuhn. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008. 94
- [36] M. Kuhn and K. Johnson. *Applied predictive modeling*. Springer, 2013. 93
- [37] R. Lenz, R. Blaser, M. Beyer, O. Heger, C. Biber, M. Bumlein, and M. Schnabel. IT support for clinical pathways lessons learned. *International Journal of Medical Informatics*, 76:S397–S402, 2007. 1
- [38] W. Liu, Y. Hsu, and B. Ma. Integrating classification and association rule mining. In *KDD'98*, 1998. 94
- [39] R.S. Mans, W.M.P. van der Aalst, R.J.B. Vanwersch, and A.J. Moleman. Process mining in healthcare: Data challenges when answering frequently posed questions. In R. Lenz, S. Miksch, M. Peleg, M. Reichert, D. Riao, and A. ten Teije, editors, *Process Support and Knowledge Representation in Health Care*, volume 7738 of *Lecture Notes in Computer Science*, pages 140–153. Springer Berlin Heidelberg, 2013. 37
- [40] J. Mendling, H. A. Reijers, and W.M.P. van der Aalst. Seven process modeling guidelines (7PMG). *Information and Software Technology*, 52(2):127–136, 2010. 25, 27, 66
- [41] R. Motley, P. Kersey, and C. Lawrence. Multiprofessional guidelines for the management of the patient with primary cutaneous squamous cell carcinoma. *British Journal of Dermatology*, 146(1):18–25, 2002. 19, 20, 23
- [42] NZa. Nederlandse Zorgautoriteit, 2015. [Online; accessed 10-November-2015]. 47
- [43] OMG. Business process management and notation (BPMN) version 2.0. *OMG Specification*, 2011. 6
- [44] PALGA. Stichting PALGA, 2015. [Online; accessed 21-October-2015]. 28
- [45] C.A. Petri and W. Reisig. Petri net. *Scholarpedia*, 3(4):6477, 2008. 6, 9
- [46] E. Ramezani, D. Fahland, and W.M.P. van der Aalst. *Where Did I Misbehave? Diagnostic Information in Compliance Checking*, volume 7481 of *Lecture Notes in Computer Science*, book section 21, pages 262–278. Springer Berlin Heidelberg, 2012. 9
- [47] A. Ramudhin, E. Chan, R. Benziane, and A. Mokadem. Modeling and optimization of health care processes. In *IIE Annual Conference Proceedings*, pages 1–6. Institute of Industrial Engineers-Publisher, 2006. 25
- [48] D.E. Rowe, R.J. Carroll, and C.L. Day. Prognostic factors for local recurrence, metastasis, and survival rates in squamous cell carcinoma of the skin, ear, and lip: implications for treatment modality selection. *Journal of the American Academy of Dermatology*, 26(6):976–990, 1992. 24
- [49] D.L. Sackett, W. Rosenberg, J.A. Gray, R.B. Haynes, and W.S. Richardson. Evidence based medicine: what it is and what it isn't. *Bmj*, 312(7023):71–72, 1996. 13
- [50] W. Sermeus and K. Vanhaecht. Wat zijn klinische paden? *Acta Hospitalia*, 42(3):5–12, 2002. 13
-

- [51] C. Shearer. The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22, 2000. 14, 92
- [52] N.R. Telfer, G.B. Colver, and C.A. Morton. Guidelines for the management of basal cell carcinoma. *British Journal of Dermatology*, 159(1):35–48, 2008. 16, 17
- [53] A. H. Tjora and G. Scambler. Square pegs in round holes: Information systems, hospitals and the significance of contextual awareness. *Social Science & Medicine*, 68(3):519–525, 2009. 1
- [54] A. Vance. Data analysts captivated by Rs power. *New York Times*, 6, 2009. 14
- [55] K. Vanhaecht, K. De Witte, and W. Sermeus. *The impact of clinical pathways on the organisation of care processes*. PhD thesis, Katholieke Universiteit Leuven, 2007. 13
- [56] V. Venkatesh, M.G. Morris, G.B. Davis, and F.D. Davis. User acceptance of information technology: Toward a unified view. *MIS quarterly*, pages 425–478, 2003. 67
- [57] L.R.E. Vermeulen. A process modelling method for care pathways. Master’s thesis, University of Technology Eindhoven, November 2013. 13, 25
- [58] E. de Vries, L.V. de Poll-Franse, W.J. Louwman, F.R. de Gruijl, and J.W.W. Coebergh. Predictions of skin cancer incidence in the Netherlands up to 2015. *British Journal of Dermatology*, 152(3):481–488, 2005. 15
- [59] A.K. Waljee, P.D.R. Higgins, and A.G. Singal. A primer on predictive models. *Clinical and translational gastroenterology*, 5(1):e44, 2014. 94
- [60] H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. 14
- [61] J.J. Wu and I.F. Orenco. Squamous cell carcinoma in solid-organ transplantation. *Dermatology online journal*, 8(2), 2002. 15
- [62] H. Yan, P.M.E van Gorp, U. Kaymak, X. Lu, R. Vdovjak, H.H.M. Korsten, and H. Duan. Analyzing conformance to clinical protocols involving advanced synchronizations. In *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on Bioinformatics and Biomedicine*, pages 61–68, 2013. 2, 5, 9, 12, 69

## Appendix A

# BPMN 2.0 process models

### A.1 Initial process models



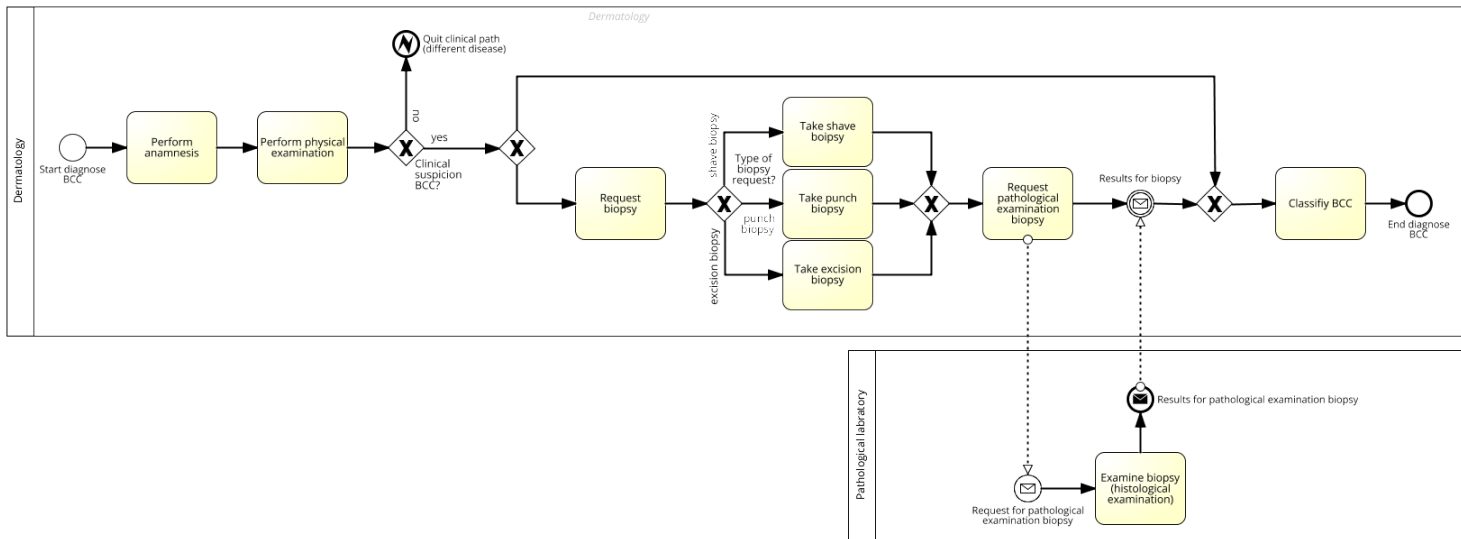


Figure A.1: Initial process model for diagnostics BCC



## A.2 Discovered process models

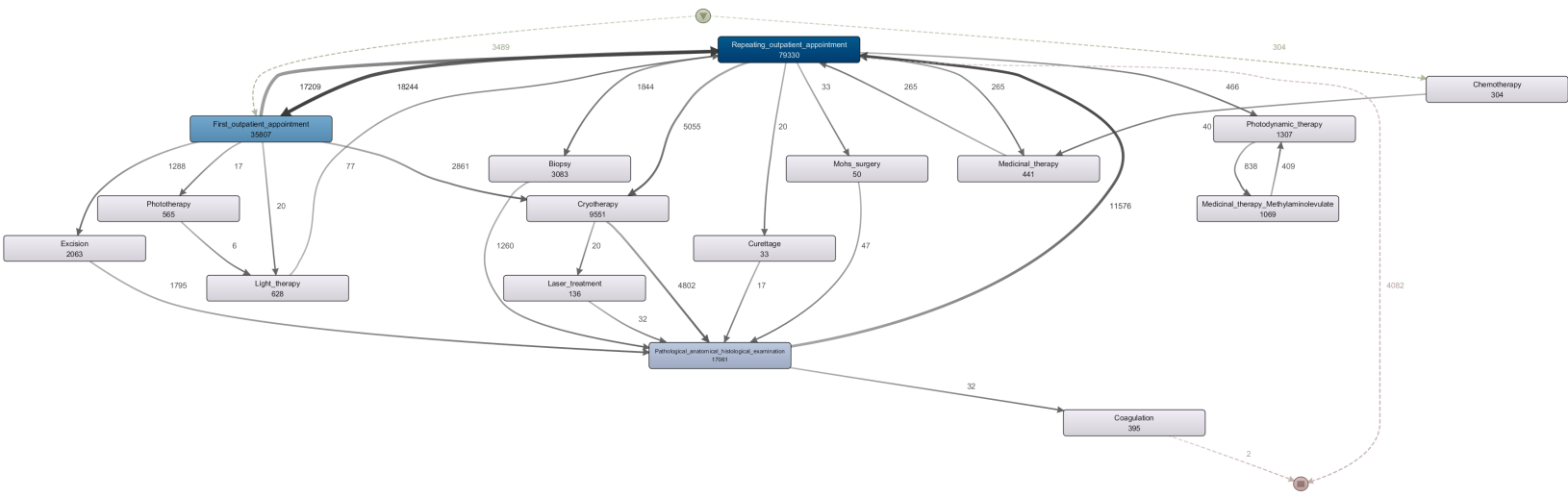


Figure A.3: Discovered process model for diagnostics BCC (aggregated activities)

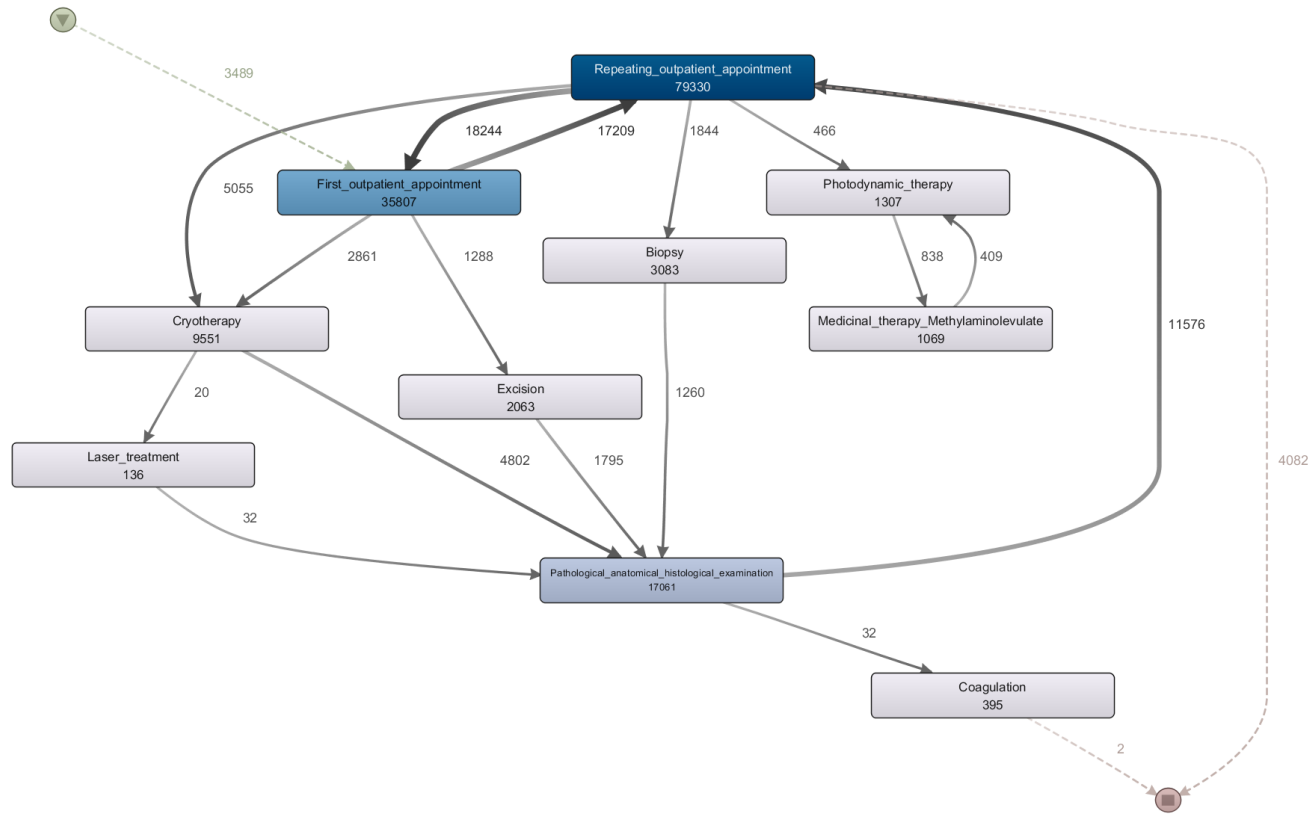


Figure A.4: Discovered process model for diagnostics BCC (filter 50%, aggregated activities)

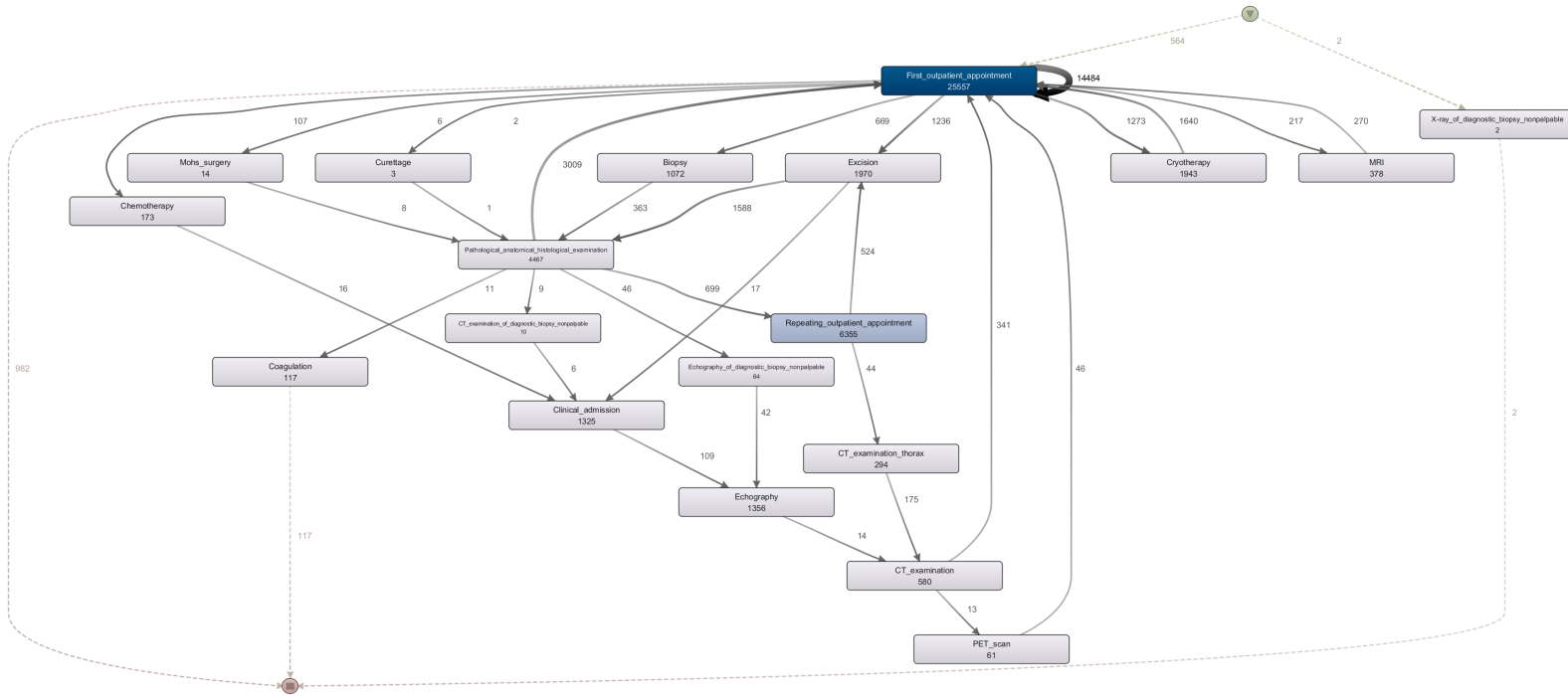


Figure A.5: Discovered process model for diagnostics SCC (aggregated activities)

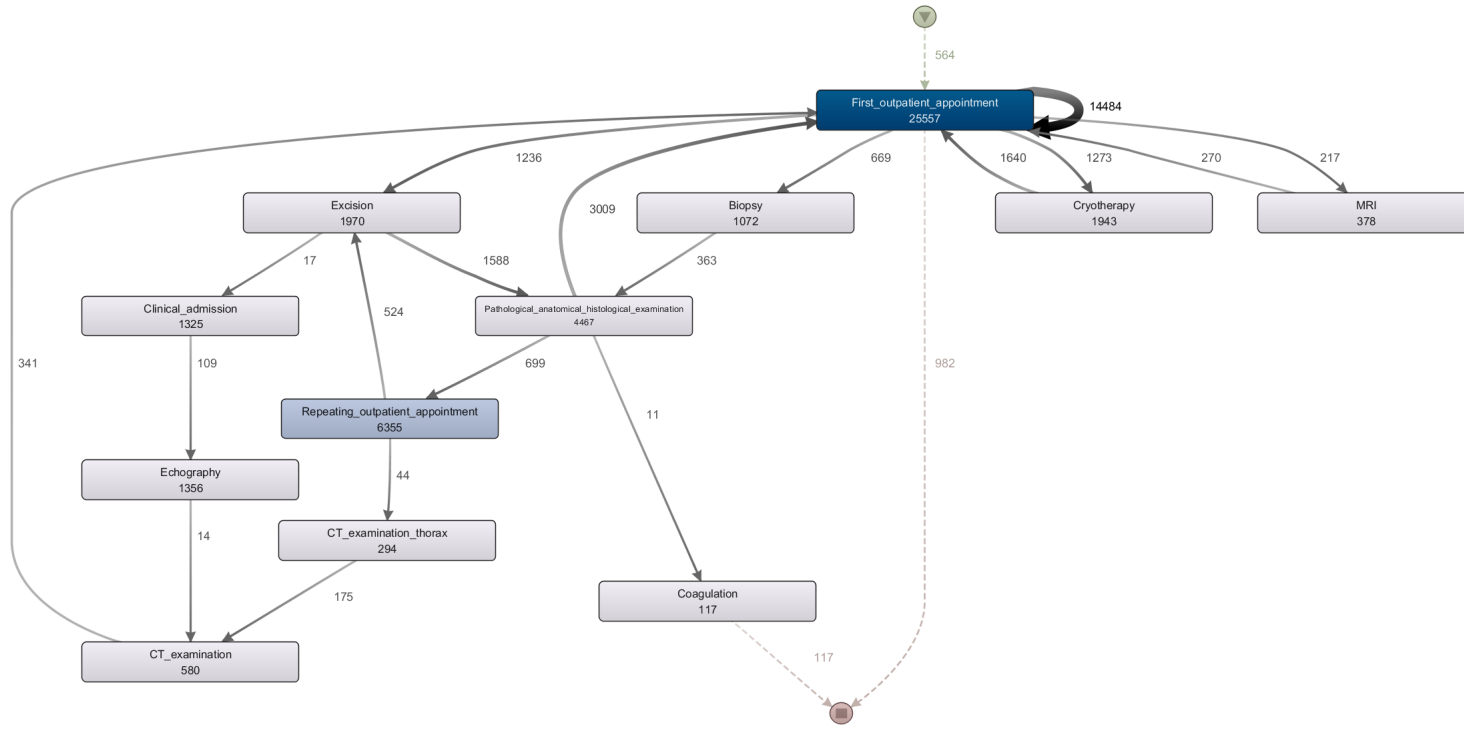


Figure A.6: Discovered process model for diagnostics SCC (filter 50%, aggregated activities)

### A.3 Data adjusted process models



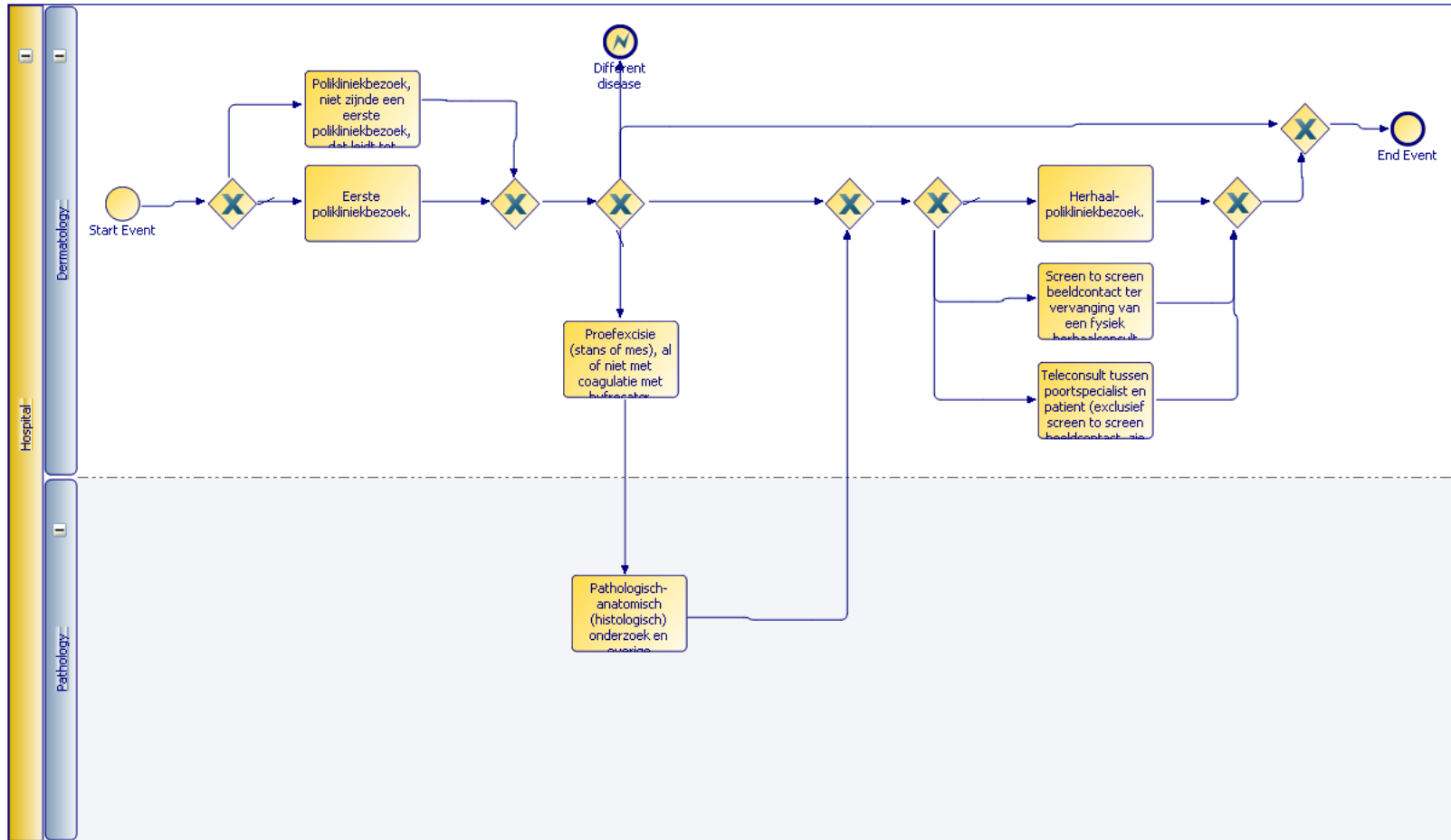


Figure A.7: Data adjusted process model for diagnostics BCC

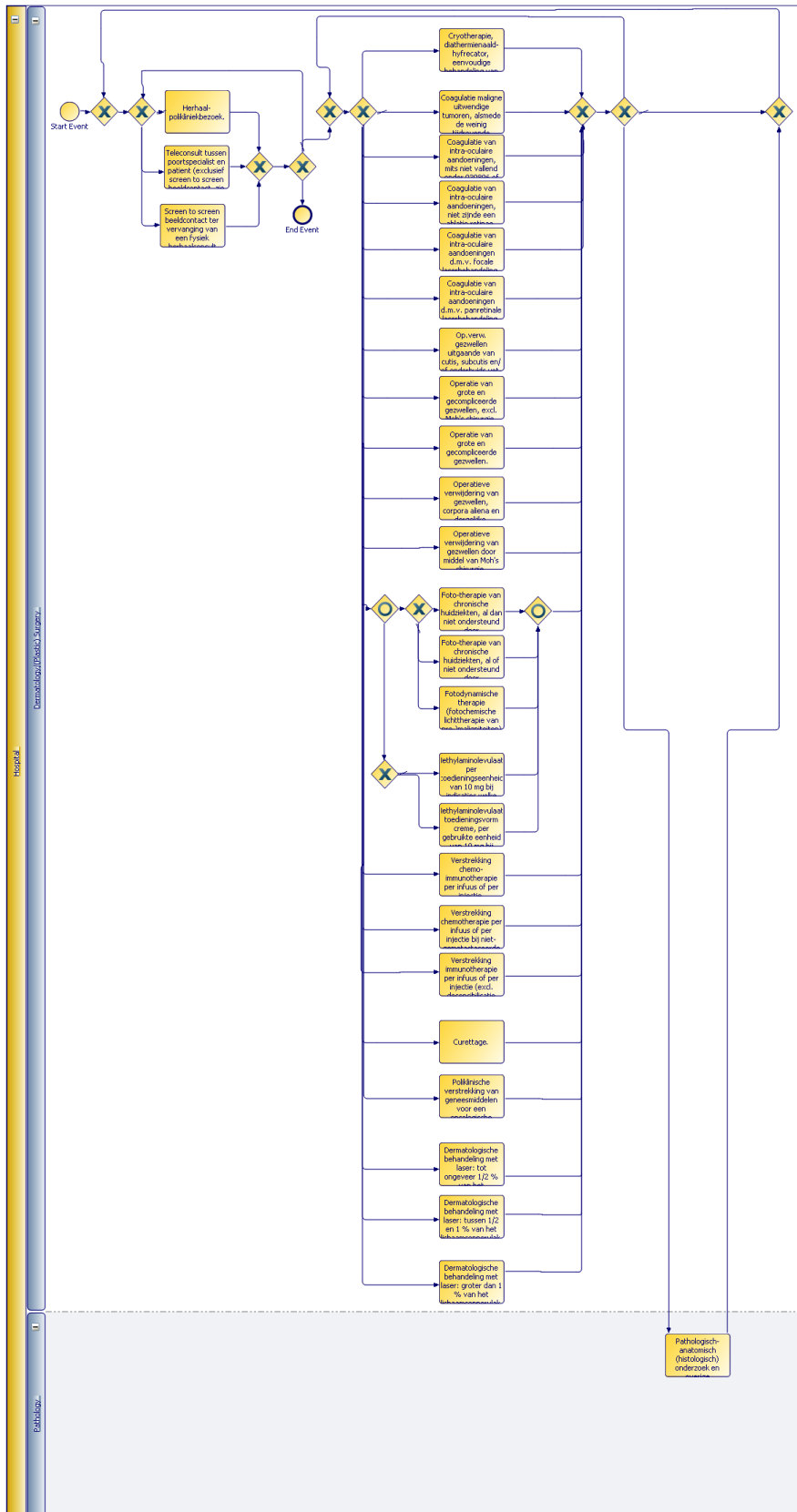


Figure A.8: Data adjusted process model for treatment and follow-up BCC

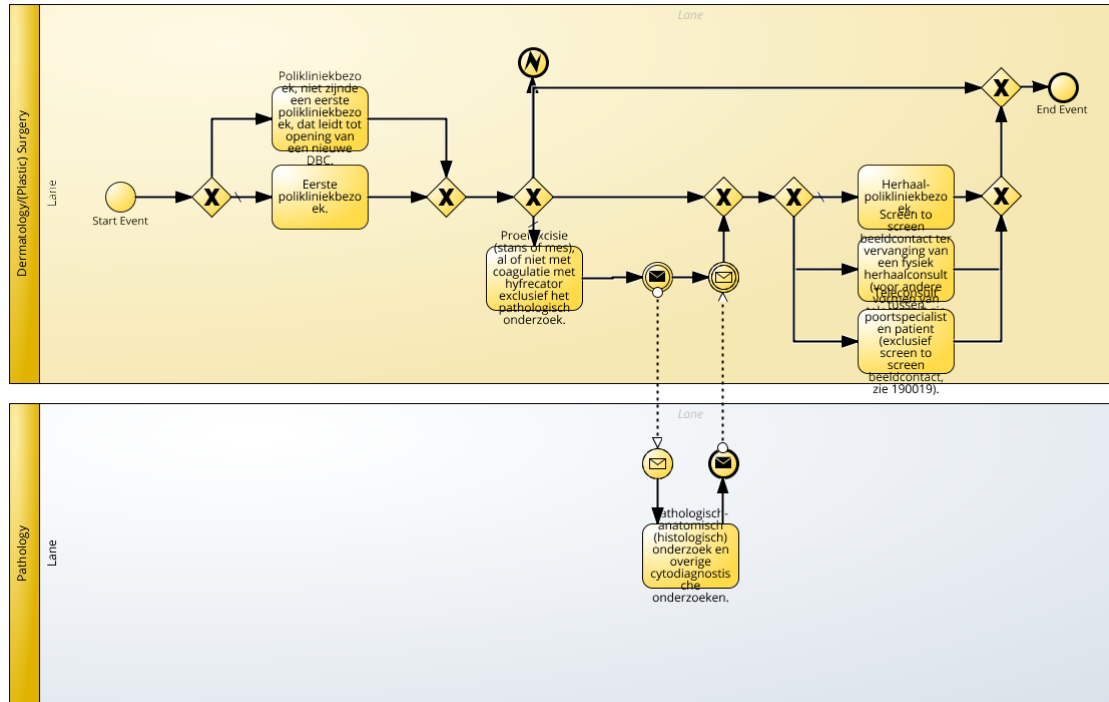


Figure A.9: Data adjusted process model for diagnostics SCC





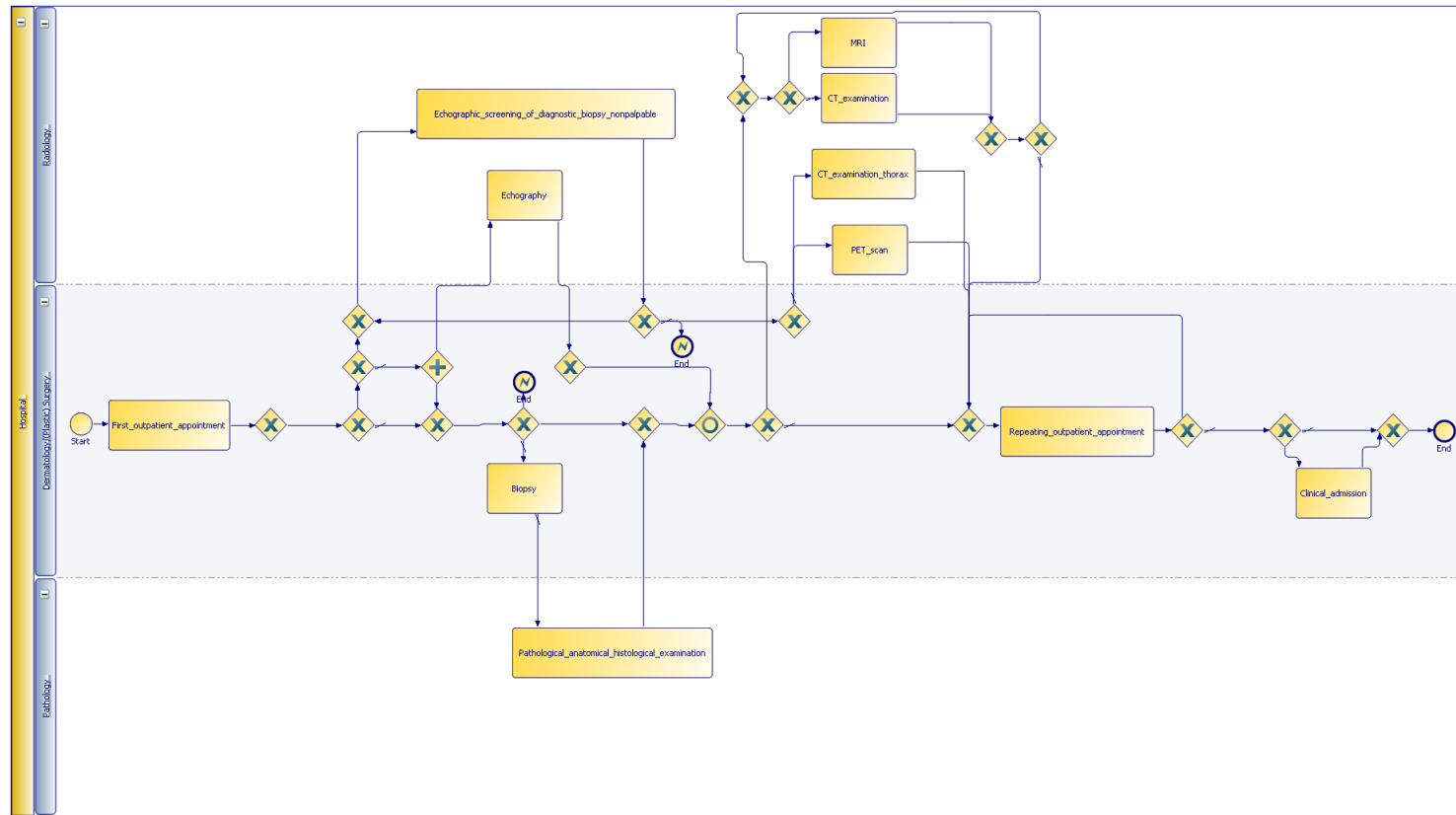


Figure A.12: Data adjusted process model for diagnostics (incl. additional diagnostics and aggregated events) SCC

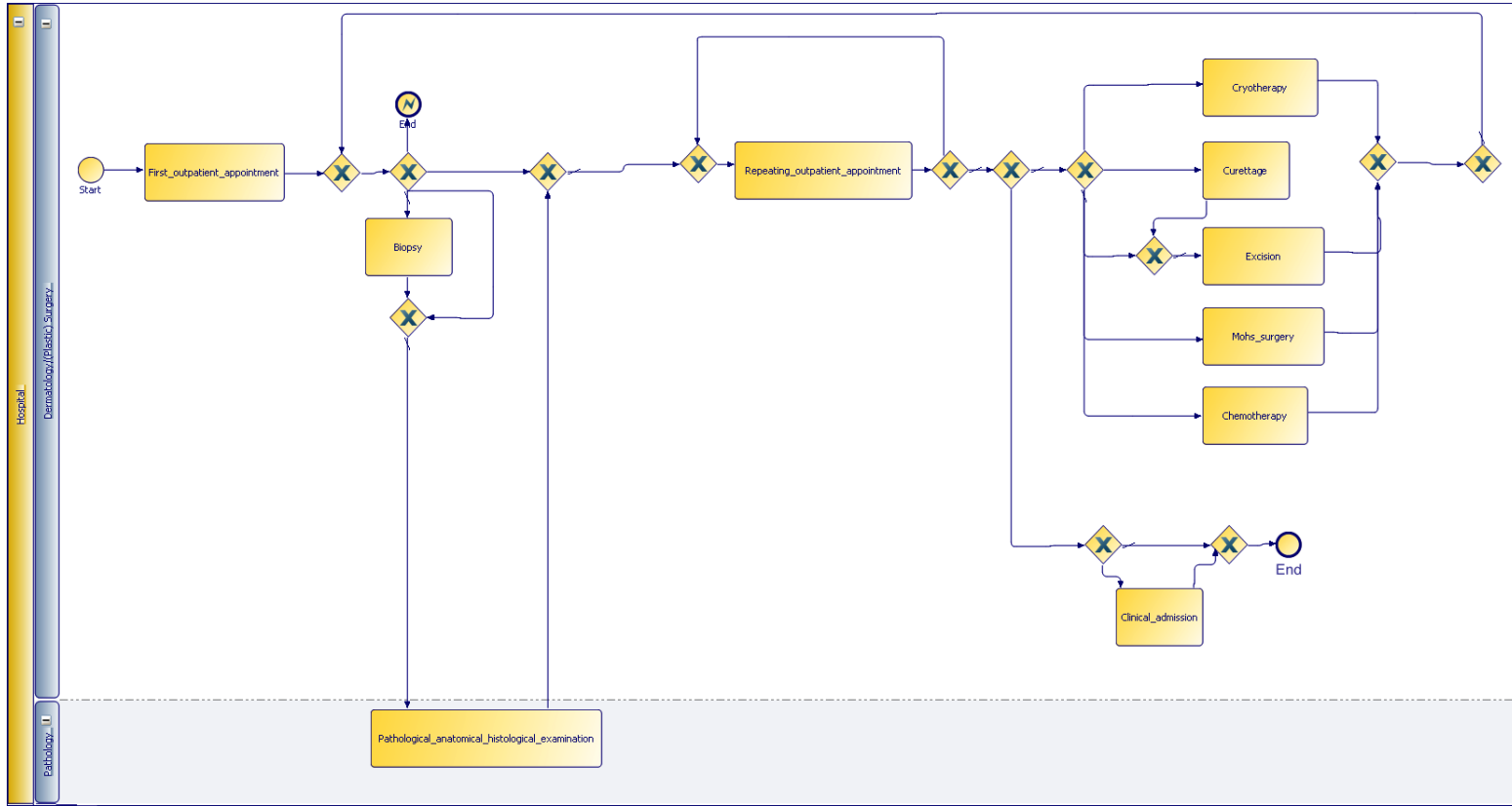


Figure A.13: Data adjusted process model for diagnostics treatment and follow-up (excl. additional diagnostics and aggregated events) SCC

## Appendix B

# Influential patient characteristics for NMSCs

Table B.1: Important patient variables for SCC (according to guidelines)

<b>Useful data in analysis</b>	
<i>Patient data (General)</i>	<i>Categories</i>
Gender	male female
Age	integer
Smoking	yes no
Skin type	I – VI
Organ transplant patient	yes no
Usage of systemic retinoids	yes no
Usage tanning device	yes no
Skin burns	yes no
Albinism	yes no
Chronic skin inflammation (Ulcera or Lichen sclerosus et atrophicans)	yes no
(Secondary) UV protection	yes no
<i>Patient data (Risks metastases)</i>	
Localization SCC	1 – 6
Diameter	$\leq 20$ mm > 20mm
Tumour depth	$\leq 4$ mm > 4mm
Level of dermal invasion	1 $\vee$ 2
Histological differentiation (Broders)	1 – 3
Immune status	1 $\vee$ 2
TMN stadium	T0M0N0 – T4M3N1
<i>Patient data (Pathology)</i>	
Surgical excision margin	... cm



# Appendix C

## Data mining

Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems [20]. This computational process is out of scope of this thesis but could be a logical next step. Therefore some background information is given in this appendix to give some guidance.

The common data mining approach used in practice is called: the CRISP-DM [51], and consists of the following stages:

**Business understanding** An essential step in every data mining project is to understand the project objectives from a business perspective. Without proper knowledge of the business, it is very difficult to find a useful solution. The steps in the *business understanding* phase are: determining the business objectives, assessing the situation, determining the data mining goals and producing the project plan.

**Data understanding** The *data understanding* phase starts with initial data collection. To detect potential data quality problems, discover insights into the data and to form hypotheses about hidden concepts in the data, it is important to increase data understanding. The steps in the *data understanding* phase are: initial data collection, description of data, exploration of data and data quality verification.

**Data preparation** The raw data collected needs to be prepared before it can be used for modelling. The *data preparation* phase covers all activities to construct this useful dataset. The steps in the *data preparation phase* are: selection of data, cleaning of data, construction of data, integration of data and the formatting of data.

**Modeling** Modelling techniques are selected and parameters (if present) are adjusted to optimal values. Association or prediction rules that follow from the model are dependent on the modelling technique selected. For example in the healthcare domain, association or prediction rules need to have face validity; physicians must accept the logic and science of the rules in order to use them in practice [9]. The steps in the *modelling* phase are: selection of modelling technique, generation of test design, creation of models, assessment of models.

**Evaluation** By evaluating the model and its construction, one can evaluate whether it properly achieves the business objectives. If certain business objectives are not sufficiently considered, the previous steps in the data mining project should be reviewed. The steps in the *evaluation* phase are: evaluate results, process review, determine next steps.

**Deployment** After evaluating the model, it is important that the client understands which actions must be taken in order to make use of the created model. The steps in the *deployment* phase are: plan deployment, plan monitoring and maintenance, production of the final report and a review of the report. An alternative approach for CRISP-DM is SEMMA (Sample, Explore, Modify, Model and Assess), is developed by SAS Institute <sup>1</sup>.

---

<sup>1</sup><http://www.sas.com/>

The approach in this thesis did already use many elements from the CRISP-DM approach in it.

## C.1 Data modeling in R

The CARET (short for classification and regression training) package that is designed for the R programming language. The primary tool in the package is the TRAIN function, which can be used to: evaluate the effect of resampling model turning parameters on model performance, choose the optimal model across the parameters and estimate the model performance based on the training set.

This can be formalized in the following pseudo code:

```

1: Define set of model values to evaluate
2: for each parameter set do
3:   for each resampling iteration do
4:     Hold-out specific samples
5:     Fit the model on the remainder
6:     Predict the hold-out samples
7:   end for
8:   Calculate the average performance across hold-out predictions
9: end for
10: Determine the optimal parameter set
11: Fit the final model to all the training data using the optimal parameter set

```

### Model building

Common steps during model building are [36]: first, the model parameters are estimated based on the training data. The parameters that cannot be directly calculated from the data should be determined afterwards. Finally the performance of the final model should be calculated based on the test data.

To find an optimal model, the available dataset is typically split into two separate sets:

- **A training set:** This dataset will be used to estimate the model parameters.
- **A test set:** This dataset will give an independent assessment of the model's efficacy and should therefore not be used during model training

If the data in the dataset is accurate, more data will give better estimates. The time spent on training and testing of data is important on the other hand. Too much time spent on training of the data could lead to over-fitting. In that case the model fits the training data very well, but is not generalizable. Too much time spent on testing of the data will not give a reliable assessment of the model parameters.

### Classification algorithms

Models can be built using a variety of algorithms in order to estimate the model parameters. These algorithms determine the estimated response and rely on characteristics of the input data.

Classification for example, is the problem of identifying to which category an observation belongs. The function to classify data is learned from labeled training data (i.e., supervised learning). First one of the most well known and robust classification algorithms is explained below (Decision tree learning). Moreover two algorithms are explained that have more challenging characteristics, but are expected to deliver promising results.

**Decision tree (CART algorithm [11])** Decision trees are formed by a collection of rules based on variables in the data set. Rules based on the values of certain variables are selected to get

the best split to differentiate observations based on the dependent variable. When the algorithm has selected the best rule to split a node into two, the same process is applied on the node from the resulting branch. Splitting stops when CART detects no further gain can be made, or some pre-set stopping rules are met. Each branch of the tree ends in a terminal node. Each observation falls into one and exactly one terminal node and each terminal node is uniquely defined by a set of rules. A very popular method for predictive analytics is Leo Breiman's Random forests, which will be discussed below. The method introduced in [11] is implemented in R<sup>2</sup> via the *tree* package.

**Rule based classification** Association rules that are classified based on a particular attribute were first proposed in [38]. The Apriori algorithm [7] is one of the most well-known algorithms for association rule mining [6], but was not yet suitable for classification problems. In [22] a simple, yet surprisingly effective method for learning decision list, based on rule learning is introduced. This algorithm is called PART and its main advantage is not its performance but simplicity. By combining two rule learning paradigms it produces good rule sets, without the need for global optimization. The method introduced in [22] is implemented in R<sup>3</sup> via the *RWeka* package. More information can be found in [29].

**Random forest** Breiman and Cutler's [10] random forest is a powerful technique for classification problems. The idea behind the Random forest algorithm is that combining models will increase the classification accuracy. The random forest algorithm uses a large collection of decorrelated decision trees. Each tree is grown as follows: The number of observations in the training set ( $N$ ) are sampled at random with replacement from the original data. This sample from the original data will be the training set. For  $M$  input variables a number  $m \ll M$  is specified.  $m$  variables are selected at random out of the original  $M$  variables the best variable out of  $m$  is used to split the node ( $m$  is held constant). Each tree is grown to the largest extent as possible (without pruning). Each individual tree in the forest with a low error rate is a strong classifier. Increasing the strength of the individual trees decreases the forest error rate. Reducing  $m$  reduces both the correlation and the strength. Increasing it increases both. In between one will find the optimal  $m$  for the most accurate classifications with a random forest. The method introduced in [10] is implemented in R<sup>4</sup> via the *randomForest* package.

### C.1.1 Predictive model

Predictive models in a clinical setting are developed with the goal of providing estimates of outcome probabilities to complement healthcare professionals. They become increasingly popular in medical research [59]. To be able to use a predictive model in practice, they should be thoroughly developed, validated and assessed on clinical impact.

The CARET package in R can be used as software to build a predictive model [35]. Tools used from this software during model building are:

- data splitting (function: *createDataPartition()*)
- model tuning using re-sampling (function: *trainControl()*), and
- variable importance estimation (function: *varImp()*)

As an example, the R code<sup>5</sup> to train and split a random forest (*rf*) model using a dataset called *Data* (dependent variable: *DV*) is shown below. For tuning the model 10 fold cross validation (*cv*) is used. The training set contains 70% of the observations from *Data*.

---

<sup>2</sup><https://cran.r-project.org/web/packages/tree/>

<sup>3</sup><https://cran.r-project.org/web/packages/RWeka/>

<sup>4</sup><https://cran.r-project.org/web/packages/randomForest/>

<sup>5</sup>For a good reference to understand the basic R code, we refer to "R by example" by Jim Albert and Maria Rizzo.

```
inTraining <- createDataPartition(y = Data$DV,  
                                  p = 0.7  
                                  )  
train <- Data[inTraining,]  
test  <- Data[-inTraining,]  
  
randomForest_model <- train(DV ~ .,  
                             data = train,  
                             method = "rf",  
                             trControl = trainControl(method = "cv", number = 10)  
                             )
```

As can be seen in the code above, without complex code a model can be trained. The `CARET` package is highly customizable to make more advanced models. Our opinion is that R or a different statistical programming language increases the understanding of the actual analysis (and data preparation).

## Appendix D

# SQL query HiX-datawarehouse

```
SELECT  rdp.PatientNr,
        rdp.Geslacht,
        rdp.GebDatum,
        dfd.DBCNummer,
        afa.AfspraakNummer,
        ofo.OperatieNummer,
        vdv.DeclCode,
        vdv.DeclOms,
        vdv.DeclAfdCode,
        vdv.DeclAfdOms,
        vdv.InvCode,
        vdv.InvOms,
        vfv.datum_key,
        rdl.VolledigeOmschrijving,
        vdb.Omschrijving,
        odh.Hoofdverrichting_key,
        ada.Naam,
        ada.Artstype,
        ada.Aanvrager_key,
        ada.HoofdspecialismeOmschrijving,
        ada.Instantienaam,
        adu.Naam,
        adu.Artstype,
        adu.Uitvoerder_key,
        adu.HoofdspecialismeOmschrijving,
        adu.Instantienaam,
        rdd.Datum,
        rdd.Dag_Naam,
        rdlg.Leeftijd,
        rdlg.LHCR

FROM    VerFeiVerrichtingen vfv
        left join DBCFeiDBC dfd on vfv.DBCFeiten_key = dfd.DBCFeiten_key
        left join RefDimPatient rdp on vfv.Patient_key = rdp.Patient_key
        left join RefDimDatum rdd on vfv.Datum_key = rdd.Datum_key
        left join VerDimVerrichtingcode vdv on vfv.VerrichtingCode_key =
            vdv.VerrichtingCode_key
        left join AgeFeiAfspraken afa on vfv.AfsprakenFeiten_key = afa.Afsprakenfeiten_key
        left join OKaFeiOperaties ofo on vfv.OperatiesFeiten_key = ofo.OperatiesFeiten_key
```

```
left join RefDimLocatie rdl on vfv.Locatie_key = rdl.Locatie_key
left join VerDimHoedanigheid vdh on vfv.Hoedanigheid_key = vdh.Hoedanigheid_key
left join OKaDimHoofdVerrichting odh on ofo.Hoofdverrichting_key =
    odh.Hoofdverrichting_key
left join VerDimBehandelsoort vdb on vfv.Behandelsoort_key = vdb.Behandelsoort_key
left join AgeDimAanvrager ada on vfv.Aanvrager_key = ada.Aanvrager_key
left join AgeDimUitvoerder adu on vfv.Uitvoerder_key = adu.Uitvoerder_key
left join RefDimLeeftijdsgroep rdlg on vfv.Leeftijdsgroep_key =
    rdlg.Leeftijdsgroep_key

WHERE rdp.PatientNr = ()
      and vfv.datum_key >= '20090101' and vfv.datum_key < '20150101'

ORDER BY vfv.datum_key asc, dfd.DBCnummer, afa.Afspraaknummer, ofo.Operatienummer,
         rdp.PatientNr
```

## Appendix E

# R script to aggregate activities for SCC

```
levels(log_SCC$Activity) [levels(log_SCC$Activity)==  
  "Herhaalpolikliniekbezoek"] <- "First_outpatient_appointment"  
levels(log_SCC$Activity) [levels(log_SCC$Activity)==  
  "Eerstepolikliniekbezoek"] <- "Repeating_outpatient_appointment"  
levels(log_SCC$Activity) [levels(log_SCC$Activity)==  
  "Pathologisch anatomisch histologisch onderzoeken overige cytodiagnostische onderzoeken"] <-  
  "Pathological_anatomical_histological_examination"  
levels(log_SCC$Activity) [levels(log_SCC$Activity)==  
  "Polikliniekbezoek niet zijnde een eerste polikliniekbezoek dat leidt tot opening van een nieuwe DBC"] <-  
  "First_outpatient_appointment"  
levels(log_SCC$Activity) [levels(log_SCC$Activity)==  
  "Cryotherapie diathermie na aldhylfrecatoreenvoudige behandeling van bijvoorbeeld naevus of wrat"] <-  
  "Cryotherapy"  
levels(log_SCC$Activity) [levels(log_SCC$Activity)==  
  "Klinische opname"] <- "Clinical_admission"  
levels(log_SCC$Activity) [levels(log_SCC$Activity)==  
  "Opverwgezwellen uitgaande van cutis subcutis en/of onderhuidsveten bindweefsel of verwijderen  
  corpora aliena inwendig metalen hechtingen apparaat volgens Tonnedm vexcisie"] <-  
  "Excision"  
levels(log_SCC$Activity) [levels(log_SCC$Activity)==  
  "Teleconsult tussen poortspecialisten patiënt exclusief screen to screen beeld contact zie 190019"] <-  
  "Repeating_outpatient_appointment"  
levels(log_SCC$Activity) [levels(log_SCC$Activity)==  
  "Proef excisie stans of mesal of niet met coagulatie methyfrecaatore exclusief het pathologisch  
  onderzoek"] <- "Biopsy"  
levels(log_SCC$Activity) [levels(log_SCC$Activity)==  
  "Echografie van het hart"] <- "Echography"  
levels(log_SCC$Activity) [levels(log_SCC$Activity)==  
  "Operatie van grote enge compliceerde gezwellen excl Mohs chirurgie"] <- "Excision"  
levels(log_SCC$Activity) [levels(log_SCC$Activity)==  
  "Echografie van de buikorganen"] <- "Echography"  
levels(log_SCC$Activity) [levels(log_SCC$Activity)==  
  "CT onderzoek van het abdomen retroperitoneum inclusief inbegrepen orale en/of rectale contraststof-  
  met of zonder toediening van een intraveneus contrastmiddel"] <-  
  "CT_examination"  
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
```

```

"CTonderzoekvandethoraxhethartengrotevateninclusiefinbrengencontrastmiddel"] <-
"CT_examination_thorax"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"CTonderzoekvandehersenenenschedelmetofzonderintraveneuscontrastmiddel"] <-
"CT_examination"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"Operatieveverwijderingvangezwellencorporaalienaendergelijkeuitgaandevanofzich-
bevindendeindieperliggendestructurendanincode038911isomschreven"] <-
"Excision"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"MRIlumbosacralewervelkolom"] <- "MRI"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"Coagulatiemaligneuitwendigetumorensmededeweinigtijdrovendecoagulatievand-
kleineretumorenvanrectumvulvaofmondholte"] <- "Coagulation"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"CTonderzoekvande wervelkolom"] <- "CT_examination"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"Bioptmatigcomplexecytologie"] <- "Biopsy"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"Verstrekkingchemotherapieperinfuusofperinjectiebijgemetastaseerdetumoren"] <-
"Chemotherapy"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"Diagnostischepunctiesvannietpalpabeleafwijkingenoforganenonderechografischecontrole"] <-
"Echography_of_diagnostic_biopsy_nonpalpable"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"PETWBwholebodyoncologie"] <- "PET_scan"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"MRIheupenondersteextremiteiten"] <- "MRI"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"Echografievandeeschildklierenofhals"] <- "Echography"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"Operatievangroteengecompliceerdegezwellen"] <- "Excision"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"Eenvoudigbiopteenvoudigecytologieexclbepalingenopdeaanwezigheidvanmicroorganismen-
zie050513of050514"] <- "Biopsy"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"MRIhersen"] <- "MRI"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"Verstrekkingchemoimmunotherapieperinfuusofperinjectie"] <- "Chemotherapy"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"MRIhersenstandaard"] <- "MRI"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"Echografievanbovensteextremiteiten"] <- "Echography"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"MRIcervicalewervelkolomenofhalsinclusiefcraniovertebraleovergang"] <- "MRI"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"Verstrekkingimmunotherapieperinfuusofperinjectieexcldesensibilisatiemiddelsimmuno-
therapiebijkinderenzie039150exclbehandelingmetmethotrexaatMTXbijkinderenzie039138"] <-
"Chemotherapy"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"Echografievanmamma"] <- "Echography"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
"Echografievanhetoogeenofbeiderzijdsinclusiefmetingvandeoogbol"] <-
"Echography"

```



```

levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Verstrekkingchemotherapieperinfuusofperinjectiebijnietgemetastaseerdetumoren"] <-
  "Chemotherapy"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "MRIachtersteschedelgroeve"] <- "MRI"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "MRIabdomen"] <- "MRI"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "MRIhersenenmetcontrast"] <- "MRI"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "MRIshoudersbovensteextremiteiten"] <- "MRI"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "CTonderzoekvandeaankezichtsschedelmetofzonderintraveneuscontrast"] <-
  "CT_examination"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "OperatieveverwijderingvangezwelendoormiddelvanMohschirurgie"] <-
  "Mohs_surgery"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Eenvoudigegroteresectiematigcomplexbioptbijzondercytologischpreparaat"] <-
  "Biopsy"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "CTonderzoekvandebovensteextremiteitenmetofzonderintraveneuscontrast"] <-
  "CT_examination"
levels(log_SCC$Activity) [levels(log_SCC$Activity)== "MRIbekken"] <-
  "MRI"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Coagulatievanintraoculaireaandoeningennietzijndeeenablatioretinaeperoog"] <-
  "Coagulation"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "CTonderzoekvandeondersteextremiteitenmetofzonderintraveneuscontrast"] <-
  "CT_examination"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "MRIthoracalewervelkolom"] <-
  "MRI"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Coagulatievanintraoculaireaandoeningenmitsnietvallendonder030896of030897"] <-
  "Coagulation"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Coagulatievanintraoculaireaandoeningendmvfocalelaserbehandeling"] <-
  "Coagulation"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Naaldbioptcomplexecytologischepunctie"] <-
  "Biopsy"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Echografievanhethartenofdethorax"] <-
  "Echography"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Curettagage"] <- "Curettagage"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "CTvanhetbekkeninclusiefinbrengeenoraleenofrectalecontraststofMetofzonder-
  toedieningvaneenintraveneuscontrastmiddel"] <-
  "CT_examination"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Coagulatievanintraoculaireaandoeningendmvpanretinalelaserbehandeling"] <-

```

```
"Coagulation"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Complexbioptmatigcomplexeresectie"] <-
  "Biopsy"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Echografievanhetbewegingsapparaat"] <-
  "Echography"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Screentoscreenbeeldcontacttervervanginvaneenfysiek-
  herhaalconsultvooranderevormenvanteleconsultzie190025"] <-
  "Repeating_outpatient_appointment"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "MRIthoraxwandmammaenmediastinum"] <- "MRI"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Echografievandeschedelnietbedoeldwordtdemidlineecho"] <-
  "Echography"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Diagnostische puncties van niet palpabele afwijkingen of organen, onder CT-controle."] <-
  "CT_examination_of_diagnostic_biopsy_nonpalpable"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Diagnostische puncties van niet palpabele afwijkingen of organen, onder rntgencontrole."] <-
  "X-ray_of_diagnostic_biopsy_nonpalpable"
levels(log_SCC$Activity) [levels(log_SCC$Activity)==
  "Lipbiopsie."] <- "Biopsy"
```