Process Mining in Oncology
A Case Study at the Netherlands Comprehensive Cancer Organisation

Himalini Aklecha
Student Number: 1032670
Department of Mathematics and Computer Science
Eindhoven University of Technology
January 2018
## Contents

1 Introduction ........................................... 4  
1.1 Division of Work ........................................ 5

2 About the NCR data and the Guidelines ............... 6  
2.1 Overview of the Data .................................... 6  
2.2 Overview of the guidelines on Oncoguide ............ 7

3 Project Aim ............................................ 9

4 Understanding NCR Data in a Process Mining Context 10

5 Methodology and Plan .................................. 12  
5.1 Step 1: Data Preparation .............................. 12  
5.1.1 Data Cleaning .................................. 12  
5.1.2 Converting Data to XES Event Log ............... 15  
5.2 Step 2: Re-creating Guidelines from Oncoguide .... 15  
5.3 Step 3: Mapping Event Log to the DPN for Conformance Checking 18

6 Results ............................................... 20  
6.1 Question 1: To what extent are the clinical guidelines followed in practice, according to the recorded data in the NCR? .......................... 20  
6.1.1 The Initial Process Model .......................... 20  
6.1.2 Revision of the Initial Process Model ............ 23  
6.2 Question 2: Can the process model explain extra behaviour about the events taking place at the hospital? ............................ 26  
6.2.1 Deviations in Diagnostic Tests ..................... 27  
6.2.2 Deviations in Surgery Phase ....................... 28  
6.2.3 Deviations in Chemotherapy Phase ............... 28  
6.3 Question 3: Can we use process mining to obtain insights about the time between the diagnosis and treatment of cancer and do they adhere with recommended time? 29  
6.4 Question 4: Can we observe significant deviations in the care-path based on cases in different types of hospitals, i.e. the academic, training and non-training hospitals? 31

7 Conclusion and Future Work .......................... 36  
7.1 Conclusion ........................................... 36  
7.2 Limitations and Future Work ......................... 37

A Extra Images ........................................ 39
Abstract

Process mining is an entity of data science that provides process-oriented solutions and aims to provide complete business process transparency. Multiple processes happen within an organization such as internal logistics within the business, customer services through the organization, and financial processes between businesses and this technology provides a real time analysis of these vast spaghetti-like workflows.

In healthcare, process mining has been an emerging area since process transparency is important for analysis of efficiency and efficacy treatment, and diagnosis of medical conditions. The healthcare system, in the Netherlands, follow guidelines that have been carefully established by medical experts in order to treat patients in the best possible way. In this report, we analyze the care-path process for the colon cancer patients recorded in the Netherlands Cancer Registry. Process mining techniques were used to decode these complicated processes taking place in the Dutch hospitals. Furthermore, we analyzed how the data adheres to the guidelines promoted and established by Netherlands Comprehensive Cancer Organization.
Acknowledgment

I would like to thank Dr. Joos Buijs for the amazing internship opportunity at Netherlands Comprehensive Cancer Organization and guiding me through my passion for process mining. Apart from being my supervisor and helping me with the technical aspects, he has also acted as a mentor to help me with the ups and downs of the project.

Dr. Gijs Geleijnse, a Clinical Informatics and Data Scientist at IKNL, has been a major part of my internship experience. He has contributed largely to the smooth flow of the internship and has also been my mentor guiding me through the confidence required and the obstacles faced during the internship.

A big thanks to Pauline Vissers, Rob Verhoeven, and Felice van Erning for the domain knowledge required during the entire lifecycle of the project. The project is heavily data-driven and the domain knowledge provided by them was key.

Mark Vroling, masters student at the Eindhoven University of Technology, has contributed to many sections of the internship and has been a supporting colleague.

I would like to thank Netherlands Comprehensive Cancer Organization for providing the opportunity to carry out the internship and the special thanks to the data managers for providing the NCR data.

Last but not the least, thank you to Eindhoven University of Technology for providing the opportunity, openness, and enthusiasm towards the internship as a part of the learning program.
Chapter 1

Introduction

Hospitals deal with a large number of patients on an every-day basis and these patients are addressed by different departments at different stages of treatments. These hospitals have complicated processes and sub-processes taking place at a very fast pace and there is a need to validate such processes for the efficient treatment of patients. Patients diagnosed with cancer, in specific, go through complicated procedures, and the decisions made about the treatments is key due to the adverse nature of cancer.

Data about the patients, their medical status, flow of the patients from one hospital department to another department or another hospital, all over the Netherlands, is recorded by the Netherlands Cancer Registry (NCR). The Netherlands Comprehensive Cancer Organization (IKNL) stores and maintains this data. Also, it has established guidelines for an ideal care-path for these patients. The data and the guidelines exist in isolation. Therefore is a need to analyze the compliance of the recorded data with such guidelines.

Techniques of “Process Mining” [1] are used to map the recorded NCR data with the guidelines, established by IKNL. The mapping between the data and the guidelines is used to discover anomalies and deviations taking place in the registry data. The information from such deviations can help to enhance the existing processes and is used for analyzing the degree to which the guidelines are being followed.

The process mining methodology has been thus applied on the NCR data and mapped onto the guidelines. The following project involves a high degree of domain knowledge, about the NCR data as well as process mining algorithms, in order to make sure that the algorithms used brings about the best information with respect to the guidelines and data.

This information can be used to review the guidelines and analyze patterns in the deviations observed. Our final aim is to discover insights that may improve quality of cancer care.
1.1 Division of Work

I, Himalini Aklecha, have been supported by Mark Vroling, masters student at the Eindhoven University of Technology, during the course of the project as we both were working on the same domain and some aspects of the project were executed in teamwork.

The following content was either executed or heavily discussed in a team work with Mark Vroling:

- Understanding Data in a Process Mining Context - Section 4
- Data Cleaning - Section 5.1.1

For Section 6.4, the analysis makes use of the tool that Mark has been developing for his graduation project. Some other sections were discussed with Mark due to the similar technical expertise shared between us.
Chapter 2

About the NCR data and the Guidelines

2.1 Overview of the Data

The Netherlands Cancer Registry (NCR) has been recording cancer data that is being provided by many hospitals, all over the Netherlands, which includes nine academic hospitals, forty five training hospitals, and some general hospitals. This data is stored, maintained, and analyzed by the Netherlands Comprehensive Cancer Organization (IKNL).

For this project we consider the care-path for the non-metastatic “colon cancer”-diagnosed patients. This is due to the fact that the care-path for the non-metastatic cancer, i.e. Stage 1, Stage 2 (and sub-stages of 2), and Stage 3 (and sub-stages of 3), is highly structured compared to that of Stage 4. Also, we consider only patients that were diagnosed in the year 2015 as the data shows high degree of completeness.

The NCR data, at IKNL, is stored in the following structure:

1. The **kern** table is the core table that consists of some basic information about the patients such as the unique patient ID, date of birth, clinical/pathological staging of the cancer/tumor, and the location of the cancer.

   A unique patient is identified with the combination of the keys \{RN, ZID, EID\} given in the **kern** table.

2. The table **uitslag** is the results table which contains all the events associated with the diagnostic results, clinical and pathological staging results.

3. The table **behandeling** has the treatment events which include activities such as different imaging methods (eg: MRI, CT, X-ray), different kinds of surgeries/resections (eg: primary tumor resection, ileostomy, colostomy), chemotherapy (eg: CAPOX, FOLFOX, 5-FU) and several other events intended towards detection and treatment of the cancer.
The important details in the two tables, uitslag and behandeling, are:

- Unique patient identification number (\{RN, ZID, EID\}).
- Unique event identification number (GID). This identification number is unique at the patient level.
- High-level event codes (gbs_gebeurtenis_code) regarding the results, detection and treatment of the patients. These event codes are not unique at the patient level, rather they are unique at the event name level and used by IKNL for data registration purpose.
- Timestamps of the events
  - Event start date (gbs_begin_dtm)
  - Event end date (gbs_eind_dtm)

4. The th_gebeurtenis table contains details about the high-level events taking place in the hospital. The high-level event codes are identified by the variable gbs_gebeurtenis_code.

5. The gebeurtenis_gegeven contains the GIDs and one of the key details contained in this table is the ggn_tumor_specifiek_item_code that is used to identify low-level events taking place in the hospital. The low-level event is hierarchically present within the high-level events and contain deeper level of information.

6. The th_specifiek_item table that contains information about low-level events, known as items, and their descriptions (omschrijving).

7. The st_specifiek_item_waarde that contains information about the values, for each item in (that are hierarchically present within the items), as well as the description (omschrijving) of the values.

The events happening at the hospitals are hierarchically arranged, where each level has a different level of abstract information, and can be seen in Figure 2.1.

![Hierarchically Relationship](image)

**Figure 2.1: Hierarchically Relationship**

### 2.2 Overview of the guidelines on Oncoguide

At IKNL colon cancer has been an interesting area of research. They have been carrying out oncological and palliative research that aims at improving the detection and treatment of cancer
as well as improving the quality of life for patients detected with incurable cancer. The extensive research has led to the establishment of comprehensive guidelines that show optimal care paths for the detection and treatment of colon cancer.

Oncoguide [8] is a repository for these guidelines, set up by IKNL, that show the multi-disciplinary care-path guidelines for diagnosis and treatment of colon cancer. The guidelines are in the form of a tree structure containing events and decisions that are important for the choice of diagnosis/treatment operations.

An example of the guideline tree structure, present in the Oncoguide, is given in Figure 2.2. The following figure gives only a glimpse of the first diagnostic test and the decision that influences the test performed. On expanding the tree structure, the multidisciplinary care-path contains the “diagnostic test” phase followed by the “surgery/resection” phase and ending with the “adjuvant chemotherapy” phase. The entire lifecycle containing the three main steps is known as the “diagnosis episode”.

![Figure 2.2: Example of the Guidelines in Oncoguide](image-url)
Chapter 3

Project Aim

As we have seen in the earlier section, the NCR data is an extensive collection of cancer data and IKNL stores and maintains this data. Apart from storing data, IKNL has also established some guidelines that should be followed for efficient diagnosis and treatment of colon cancer diagnosed patients.

In order to check how well the recorded data complies or follows the guidelines we have used process mining techniques to map the data onto the guidelines as well as to decode these complicated processes to answer questions given below.

- To what extent are the clinical guidelines followed in practice, according to the recorded data in the NCR?
- Can the process model explain extra behaviour about the events taking place at the hospital?
- Can we use process mining to obtain insights about the time between the diagnosis and treatment of cancer and do they adhere with recommended time?
- Can we observe significant deviations in the care-path based on cases in different types of hospitals, i.e. the academic, training and non-training hospitals?

These questions are answered by using multiple process mining techniques and a strong domain knowledge about colon cancer, the recorded data, and the structure of the guidelines established by IKNL.
Chapter 4

Understanding NCR Data in a Process Mining Context

Event logs are the basic building blocks needed for process mining and they typically comprise of:

1. Cases/traces that describe individual process instances.
2. Activities/events that describe steps of a process flow.
3. Start and end timestamps of each activity in the process.
4. Other data attributes that are features or properties that describe individual traces or individual events. These attributes are needed to obtain meaningful insights about the data by filtering the process flow. As a result, the process flow can be broken down into focused/detailed models that allow for deeper analysis.

With respect to the NCR data these process mining concepts can be visualized in the following way:

1. Patient identification number can be used to describe unique cases, since the care-path for each patient is an individual process instances.
2. A particular patient goes through multiple events at a hospital such as medical tests, surgeries, and consultations. Apart from the process flow of the patient, the hospital has its own internal logistics that is also in the form of a process flow, like creating pathology reports after a surgery, or X-ray reports after a X-ray scan (which can be in parallel or interleaving with the sub-process of the patient).

These process events form process mining activities, and the start and end dates recorded for these events are considered as the start and end timestamps, respectively. However, the end timestamps show higher percentage of incompleteness and the end timestamps are not always stored in the desirable way. For example, “resection” has an end date that corresponds to the end of the entire lifecycle of a surgery phase and not the end of the actual surgery/resection event. Thus we consider only the start date timestamps for the further analysis.
3. Few of the important data attributes present in the NCR data are listed below:

(a) Stage of the cancer
(b) Age of the patient
(c) Medical status of the patient
(d) Clinical stage of the cancer/tumor that provides information about the possible next steps needed to be taken for the patient
(e) Details where the patients were first diagnosed

These data attributes characterize traces (patients) as well as individual events.
Chapter 5

Methodology and Plan

• Step 1: Understanding how we can transform the data in order to apply process mining techniques.

• Step 2: Re-creating the guidelines mentioned in Oncoguide, established by IKNL, such that the data and guidelines talk the same language.

• Step 3: Map the event log to the re-created guidelines for conformance checking.

5.1 Step 1: Data Preparation

The steps performed in this section has been executed in Alteryx [11] (shown in Figure A.1) and few data manipulation steps have been carried out in SAS [12].

5.1.1 Data Cleaning

The first step was to use the process aspects of the data for the manipulation and cleaning of the data. Some of the main characteristics of the NCR data are i) it is distributed over multiple database tables, ii) the recorded hospital events are recorded in multiple abstract levels (Figure 2.1), and iii) multiple events contain missing start and end timestamps.

First, multiple joins were performed over database tables in order to retrieve meaningful information about the patients, their medical conditions, and the events that took place for the patients along with event/activity timestamps.

The steps taken for database manipulation are described next along with the database architecture overview which can be see in Figure 5.1:
1. Since we consider only the data and care-path for colon cancer, the **kern** table was filtered on two cancer IDs 205310 (colon cancer) and 205320 (appendix cancer), specified in the column **tTumorsoort** present in the **kern** table.

The data was further filtered with the “year of diagnosis” of 2015 present in the **tIncdat** column.

2. The unique patient ID {**RN, EID, ZID**} was used as a primary key to access the two tables, **uitslag** and **behandeling**.

A left inner join was performed on the **kern** and the **uitslag**, and separately on the **kern** and the **behandeling**. The above two joins were then appended and the resulting table contains process information for all the patients that were diagnosed with colon cancer in the year 2015.

3. The **gbs_gebeurtenis_code** is used as a primary key to access the high-level event descriptions present in **th_gebeurtenis** table.

4. The next step is to access the low-level items and their values and this is done by first accessing the table **gebeurtenis_gegeven** using the **GID** as a primary key. Then from the above join the item code (**ggn_tumor_specifiek_item_code**) is used as a primary key and a left inner join was performed with the table **th_specifiek_item** to obtain event description
and with table \texttt{th\_specifiek\_item\_waarde} to obtain item values and value description. The two tables were then appended.

5. The appended data was then grouped based on different hospital types (\texttt{tEerstezkh}) i.e. academic hospitals, training hospitals, non-training hospitals and other hospitals.

Without disrupting the confidentiality of the data, the Table 5.1 shows the basic structure of the details captured from the previous joins performed on the NCR data.

<table>
<thead>
<tr>
<th>RN</th>
<th>ZID</th>
<th>EID</th>
<th>eventDescription</th>
<th>eventCode</th>
<th>itemDescription</th>
<th>itemCode</th>
<th>waardeDescription</th>
<th>waardeCode</th>
<th>event start date</th>
<th>event event date</th>
<th>other data attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>xxx1</td>
<td>xxx2</td>
<td>xxx3</td>
<td>hemicolectomy left</td>
<td>120/152</td>
<td>apendicease</td>
<td>18D4</td>
<td>unknown</td>
<td>9</td>
<td>4444</td>
<td>4444</td>
<td>...</td>
</tr>
<tr>
<td>xxx4</td>
<td>xxx5</td>
<td>xxx6</td>
<td>CAFOX</td>
<td>45000</td>
<td>number of courses</td>
<td>99D16</td>
<td>18 courses</td>
<td>13</td>
<td>4444</td>
<td>4444</td>
<td>...</td>
</tr>
<tr>
<td>xxx7</td>
<td>xxx8</td>
<td>xxx9</td>
<td>pathology</td>
<td>998</td>
<td>distance to cutting plane</td>
<td>20D4</td>
<td>5mm</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>...</td>
</tr>
<tr>
<td>xxx10</td>
<td>xxx11</td>
<td>xxx12</td>
<td>molecular diagnostics</td>
<td>8040</td>
<td>MSI status</td>
<td>39D0</td>
<td>stable</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 5.1: Example of the Manipulated NCR Data

As we see in the above table, some high-level events have missing dates and every high-level event (eventCode) has low-level events (itemCode) under them and every low-level event contains a value (waardeCode) within it.

Second, The high-level events, that are associated with the results from a diagnostic test, were not taken as a process mining activities but were considered as events containing information (contained in the item and waarde) that were converted to case/trace attributes. Since they showed a high level of missing timestamps, the percentage of missing timestamps were also reduced drastically after the above consideration.

The high-level events that were considered for this step are:

1. Pathology
2. Molecular Diagnostics
3. Laboratory Research
4. Physical Examinations

The event “No therapy” and “end of clinical follow up” were removed which further reduced the percentage of missing activity timestamps since these events had a contextual meaning but can not be considered a workflow steps. The high-level activities that are associated with surgeries were re-named as “resection” for further simplicity.

Lastly, the high-level activity “active surveilance” was given a start date as the start date of the next activity or a start date as the end date of the previous activity after considering if the “active surveilance” is the first activity of a trace or not, respectively.
5.1.2 Converting Data to XES Event Log

In order to use the data along with process mining techniques, the data was converted into an appropriate XES event log with the help of the process mining tool ProM 6.7 [9].

The following steps were performed to obtain the right event log for conformance checking:

1. The data (CSV format) was imported into the process mining tool ProM and converted to an XES event log using the plugin “Convert CSV to XES”, created by F.Mannhardt, N.Tax, and D.M.M.Schunselaar.

   The following parameters are used: “CaseID” as the case column, “main_event_name” as the event column, and “Start.date.treatment” as the start time. The error handling is considered as “Omit Trace on Error”.

2. The decision points, mentioned in Table 5.2, were moved from event level to trace/case level using the plugin “Copy attribute to trace level (In place)”, created by F.Mannhardt. All the variables are re-named in this plugin by using the post-fix 2 (example: Emergency_surgery -> Emergency_surgery_2) so that we can visualize the trace attributes and the event attributes separately.

3. An artificial “START” activity was added using the plugin “Add Artificial Events”, created by J.Claes.

4. The eight trace attributes were then given to the “START” activities (in order to be able to read the values right from the start of each process instance) using the plugin “Copy trace attributes to the 1st event (In place)”, created by F.Mannhardt.

5.2 Step 2 : Re-creating Guidelines from Oncoguide

The decision points used in the Oncoguide colon cancer guidelines that are available in the data are shown below:

<table>
<thead>
<tr>
<th>S.no</th>
<th>Oncoguide Decisions</th>
<th>Decision as in the Data</th>
<th>High-level Event</th>
<th>waardeCode/value</th>
<th>Re-named Decisions</th>
<th>Decision Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Acute Presentation</td>
<td>spoedoperatie</td>
<td>Resection</td>
<td>1, 2, and 3</td>
<td>Emergency_surgery</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>Localization Obstruction</td>
<td>tTopog</td>
<td>Localization_obstruction</td>
<td>C180, C181, C182, C183, C184 and C185</td>
<td>left</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Lymph Angioinvasion</td>
<td>angiogeneseon</td>
<td>Pathology</td>
<td>1 and 2</td>
<td>Lymph_angioinvasion</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>Metastatic Stability</td>
<td>MSI status</td>
<td>Molecular Diagnostics</td>
<td>0</td>
<td>Metastatic_stability</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>Differentiation Degree</td>
<td>Diffgr</td>
<td>Molecular Diagnostics</td>
<td>1</td>
<td>Microsatelite_stability</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>Clinical Tumor (cT)</td>
<td>sCT</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>Cancer Stage</td>
<td>Stadium</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>Hospital_type</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5.2: Case/Patient Attributes

The last trace attribute, i.e. “Hospital_type”, in the Table 5.2 was not used as a decision point with respect to the guidelines but were used as a trace decision for some further analysis.
Re-creation of the guideline was done using the Data Aware Petri-Net (DPN) methodology, where process flows can be designed, in the form of a petri-nets, and information about decisions made during the process flow, process resources, and time perspectives can be added to the petri-net thus making it data-aware.

First, the petri-net structure was created without making the petri-net data aware. This step was done using the tool “WoPeD” (Workflow Petri-net Designer) [10] which is used for modelling, simulating, and analyzing process workflow models. The petri-net (re-created from the guidelines) is shown in **Figure 5.2**.

The process model was then checked for various properties such as “Soundness” and the “Workflow net property” with the help of the “Semantical Analysis” feature provided by WoPeD. Soundness is defined by, if the process is: (i) safe (only one token at a place at a time), (ii) can be properly completed, (iii) has an option for completion, and (iv) has absence of dead parts. The complete semantical analysis is shown in the Appendix (**Figure A.2**).

![Figure 5.2: Petri-net obtained from WoPeD](image)

The petri-net in **Figure 5.2** was then imported into ProM and edited using the plugin “Edit DPN (Text language based)”, created by F. Mannhardt.

The transitions and places were given meaningful names and the petri-net was made data aware by giving the transitions the required guard conditions as per the decision conditions in the Oncoguide (described in **Table 5.2**). Some of the aspects about the process that was created in WoPeD was also further edited in ProM.

The first section of the process workflow (diagnosis episode) deals with the diagnosis of cancer and when the patient does not need an emergency surgery, more diagnostic events are performed. The primary diagnosis gives information about the clinical stage of the tumors. After the primary diagnosis, resections/surgeries are performed followed by adjuvant chemotherapy.

Since we are using cancer data majorly collected in 2015-2016, the recurrence phase is not analyzed.
The resulting DPN can be seen in Figure 5.3.

Figure 5.3: Data Aware Petri-net
5.3 Step 3: Mapping Event Log to the DPN for Conformance Checking

In order to understand how well the data collected for the colon cancer fits the guidelines, specified in the Oncoguide, a conformance checking was performed. Conformance checking allows us to know how well the data fits with the pre-defined process model as well as other performance parameters such as the precision and mapping statistics. For this purpose we use the plugin available in ProM known as the “Multi-perspective Process Explorer” [2] [3], created by F. Mannhardt.

Using the “Multi-perspective Process Explorer” plugin, we obtained the following model shown in Figure 5.4 on using the “Data Discovery Mode”.

The “Multi-Perspective Process Explorer” uses the conformance checking and fitness calculation methodology from the “Balanced multi-perspective checking of process conformance” [3]. This method uses the concept of conformance checking and fitness and extends it to DPNs.

The “Multi-perspective Process Explorer” plugin calculates fitness by minimizing the cost that occurs due to i) move in log (default cost = 2), where events/activities took place in the data but did not match with any event in the process flow of the model, ii) move in model (default cost = 3), where events/activities took place in the model but did not take place during the process flow of the data, iii) move in both together but having different data perspective values (default cost = 1), and iv) “move on silent transition” (default cost = 0). In general, we can define fitness as the measure of how well the data represents the model when replayed over the model. The “move in model” and “move in log” together can be seen as “control-flow” move.

In Figure 5.4 we can see the initial model obtained from using the “Data Discovery” mode showing frequency perspective where each path shows the frequency of cases following through that path. This model was obtained after simply using the plugin on the event log, obtained in Section 5.1.2 and the data-aware petri net, obtained in Section 5.2.

In the next section we will discuss the performance results of the initial model and how we used the results to further improvise and restrict the process model such that the model strongly represents the guidelines.
Figure 5.4: Data Discovery Mode with Frequency Perspective
Chapter 6

Results

In this section, we will answer the main questions of the project aim:

- Question 1: To what extent are the clinical guidelines followed in practice, according to the recorded data in the NCR?
- Question 2: Can the process model explain extra behaviour about the events taking place at the hospital?
- Question 3: Can we use process mining to obtain insights about the time between the diagnosis and treatment of cancer and do they adhere with recommended time?
- Question 4: Can we observe significant deviations in the care-path based on cases in different types of hospitals, i.e. the academic, training and non-training hospitals?

6.1 Question 1: To what extent are the clinical guidelines followed in practice, according to the recorded data in the NCR?

6.1.1 The Initial Process Model

The performance results of the model discovered in Figure 5.4 is given in the table below.
<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Cases</td>
<td>8,722</td>
</tr>
<tr>
<td>Fitness</td>
<td>75.3%</td>
</tr>
<tr>
<td>Correct Events</td>
<td>31,513</td>
</tr>
<tr>
<td>Wrong Events</td>
<td>15,744</td>
</tr>
<tr>
<td>Missing Events</td>
<td>7,549</td>
</tr>
<tr>
<td>Data Violations</td>
<td>4.8%</td>
</tr>
<tr>
<td>Event Violations</td>
<td>42.5%</td>
</tr>
</tbody>
</table>

Table 6.1: Performance Measures

From Table 6.1 we see that there are total of 8,722 cases/patients and the fitness of the overall model is 75.3%.

Correct events are defined as the events/activities (non-unique) that happened in the model as well as in the data and can be considered as accurate matches. Wrong events are “move in log” events. On the other hand, missing events are “move in model” events.

We see that the majority of the events from the data could be mapped to the guideline model. However, there were more number of “move in log” than “move in model” and this can be explained by the fact that the data recorded in the hospitals have larger number of unique event/activity names that are not mentioned in the guidelines.

We then used the “Fitness Mode” of the “Multi-Perspective Process Explorer” plugin which is shown in Figure 6.1 to check the performance of the model at the event level. The color coding in the fitness model can be explained by Figure A.3 shown in Appendix A. The darker the color, the lower the fitness of the event/activity in the model.
Figure 6.1: Fitness Model
6.1.2 Revision of the Initial Process Model

On observing the traces provided by “Toggle Trace View” it was evident that in many cases there is a cost on the fitness due to the presence of “resection” as the first event followed by a diagnostic test. Since, we considered only the diagnosis episode for the colon cancer patients, this kind of behaviour seemed uncanny and was analyzed.

On analysis of the data, it was observed that there are cases where a resection and a diagnostics test were performed on the same day. Using domain knowledge it was concluded that in such cases the two events should be manually reversed in order, such that ProM sees a diagnostic test first and then a resection. On performing the inversion of events the number of such cases were reduced.

It was also observed that due to the default setting of the plugin, where “move in both together but having different data perspective values” had a cost of 1, there was high presence of guard violations throughout the model. Since our main goal was to analyze the accuracy of the data with respect to the guidelines, the cost value was increased to reduce the guard violations. On experimental basis, it was concluded that a cost of 5 on the “data-perspective” violations (guard violations) and a cost of 1 on the control-flow moves was the most appropriate in terms of reducing guard violations without over-fitting the model.

The reason for changing the control-flow moves to cost 1 can be comprehended in the following way: The nature of the data and the guidelines, which is that the data has more recorded unique activities compared to that of the guidelines, would facilitate multiple control-flow moves which can over-power the guard violation. Thus by setting the cost value for the control-flow moves to 1 against cost of 5 for guard violations, the model was tuned to follow the guards (decisions) more strongly than compared to the control-flow.

Apart from tuning performance parameters, it is also evident from the fitness model that event “X thorax” could not be found in the data and had a transition cost of 0% and by using domain knowledge it was concluded that event “X thorax” has the same contextual meaning and application as “Chest X-ray”. Therefore “X thorax” can be mapped to “Chest X-ray” while tuning the performance parameters.

Below we have the table showing the default parameter values and new parameter values:

<table>
<thead>
<tr>
<th></th>
<th>Default Values</th>
<th>New Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Move in Log</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Move in Model</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Move in both together but having different data perspective value</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Move on Silent Transition</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.2: Parameter values for fitness calculations
The data discovery mode of the new model can be seen below:

Figure 6.2: Data Discovery Model for Revised Model with Frequency Perspective
The fitness mode of the new model can be seen below:

Figure 6.3: Fitness Model for Revised Model
The performance results of the revised model is shown below in Table 6.3.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Cases</td>
<td>8,722</td>
</tr>
<tr>
<td>Model Fitness</td>
<td>94.7%</td>
</tr>
<tr>
<td>Correct Events</td>
<td>35,776</td>
</tr>
<tr>
<td>Wrong Events</td>
<td>11,481</td>
</tr>
<tr>
<td>Missing Events</td>
<td>7,740</td>
</tr>
<tr>
<td>Data Violations</td>
<td>0.3%</td>
</tr>
<tr>
<td>Event Violations</td>
<td>34.9%</td>
</tr>
</tbody>
</table>

Table 6.3: Performance Measures of the Revised Model

The over fitness of the model has increased to 94.7% showing that the data fits the improvised model better but at the same time is more representative of the guidelines as the data follows the decision points first over the control flow. This is also confirmed from the fact that the data violations have dropped from 4.8% to 0.3%.

Thus we can say that that recorded data does fit the overall structure of the guidelines by 94.7% while still following the decisions/conditions up to 99.7%.

6.2 Question 2: Can the process model explain extra behaviour about the events taking place at the hospital?

To understand the Figure 6.3 we observe activities that show low fitness values (in terms of the colors specified in Figure A.3 in the Appendix). The table given below specifies the events with a cost in fitness values. The fitness is calculated by the number of correct events versus the number of total events that pass through the transition.

<table>
<thead>
<tr>
<th>Event Name</th>
<th>Fitness</th>
<th>Guard</th>
<th>Guard Violations</th>
<th>Correct Events</th>
<th>Move in Model</th>
<th>Move in Log</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT abdomen</td>
<td>86.1%</td>
<td>Emergency_surgery = yes</td>
<td>0.4%</td>
<td>444</td>
<td>7%</td>
<td>0</td>
</tr>
<tr>
<td>reaction</td>
<td>94.2%</td>
<td>Localization_obstruction = left</td>
<td>0%</td>
<td>424</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>reaction</td>
<td>95.6%</td>
<td>Localization_obstruction = right</td>
<td>0%</td>
<td>320</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>CT colonography</td>
<td>91.1%</td>
<td>(upper level guard) Emergency_surgery = no</td>
<td>0%</td>
<td>440</td>
<td>2.138</td>
<td>0</td>
</tr>
<tr>
<td>CT abdomen</td>
<td>80.9%</td>
<td>(upper level guard) Emergency_surgery = no</td>
<td>0%</td>
<td>6,498</td>
<td>1.936</td>
<td>0</td>
</tr>
<tr>
<td>X thorax</td>
<td>94.3%</td>
<td>sCT = 0 or sCT = IS or sCT = X</td>
<td>0%</td>
<td>4,364</td>
<td>30.92</td>
<td>0</td>
</tr>
<tr>
<td>active surveillance</td>
<td>0%</td>
<td>aCT = 0 or aCT = 1 or aCT = X</td>
<td>2%</td>
<td>2</td>
<td>97</td>
<td>0</td>
</tr>
<tr>
<td>reaction</td>
<td>93.9%</td>
<td>aCT = 4 or aSt = 1 or aCT = 1 or aSt = 1 or aCT = 3 or aSt = 3 or aCT = X</td>
<td>0%</td>
<td>4,137</td>
<td>44</td>
<td>0</td>
</tr>
<tr>
<td>reaction</td>
<td>98.6%</td>
<td>aCT = 2 or aSt = 2 or aCT = 1 or aSt = 1 or aCT = X</td>
<td>0%</td>
<td>2,438</td>
<td>73</td>
<td>0</td>
</tr>
<tr>
<td>nodule transition</td>
<td>100%</td>
<td>aCT = 1 or aSt = 1 or aCT = X</td>
<td>1.1%</td>
<td>1,214</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CAPOX</td>
<td>94.3%</td>
<td>(upper level guard) sStadium 3 or sStadium = 3A or sStadium = 3B or sStadium = 3B or sStadium = 3B</td>
<td>0%</td>
<td>1,104</td>
<td>66</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.4: Events that show presence of Non-Perfect Fitness
6.2.1 Deviations in Diagnostic Tests

Cases with No Acute Presentation:

From Table 6.4 and Figure 6.3, the most evident deviations are the events “CT colonography” and “active surveillance”.

From the “Trace View” and the event log, it was seen that there are multiple events, such as “sigmoidscopy” and “echo”, that were performed in the place of “colonoscopy” and “CT colonography”. Due to which there are control-flow moves which causes the decrease in transition fitness values. However, the event “colonoscopy” seems to show a 100% transition fitness and the event “CT colonography” has a fitness of 17.1%. This ambiguity was cleared after understanding the algorithm and it was discovered that, in the absence of no guards, when there is a move in the model and none of the events are seen in the log, the algorithm gives the first event, specified in the DPN text, with a perfect fitness and gives the second event, specified in the DPN text, with the “move in log” cost value thus reducing fitness. In practice the fitness and the cost should be divided between these two activities.

There are 2,138 missing events for the “CT colonography”, which as discussed above could be due to missing “colonoscopy” or “CT colonography”. Thus about 27% of the cases do not go through any of the two mentioned events and thus deviate from the guidelines.

The next event observed was “X thorax” which was mapped to “Chest X-ray” and this event has a transition fitness of 54.3%. The parallel event to this event is “CT abdomen” having a fitness of 80.9%. The transition fitness cost can be answered using the same reasoning that the data records other diagnostic imaging tests, such as “CT thorax” and “MRI”, that happened either in the place of “CT abdomen” or “X thorax” or between the first diagnostic test and the second diagnostic tests. In some cases there is only one diagnostic test followed by a resection or chemotherapy.

There were approximately 19% cases (1,536 of 8,034) where “CT abdomen” was not performed and about 46% (3,670 of 8,034) cases where no “X thorax” was performed. There are some cases overlapping between the two mentioned events where none of the two events were performed.

It was further observed, by using filtering query \( \{sCT = 0 \text{ or } sCT = IS \text{ or } sCT = X\} \), that there were only two cases that performed the activity “active surveillance” and the two cases where when \( \{sCT = X\} \). The cases that have \( \{sCT = 0 \text{ or } sCT = IS\} \) go through the silent transition (can be interpreted since only that transition has a guard violation on “sCT”). This means that the data for \( sCT = 0 \text{ or } sCT = IS \) is deviating completely and does not follow the guidelines.

None of the patients with a clinical tumor stage of 0 or “IS” (in situ) passed through the “active surveillance” stage. The high cost on guard violations were removed, temporarily, to observe the behaviour for these patients. It was observed that most 52% of the patients with \( \{sCT = 0 \text{ or } sCT = IS\} \), who showed no acute presentation, went through a resection and this behavior deviates from the guideline.
Cases with Acute Presentation:

When there is a need of an emergency surgery there were about 11% cases where “CT abdomen” was not performed. This is deviating from the guidelines, however, this is a smaller number than other model deviations. In these cases a resection was performed directly without a second diagnostic test.

6.2.2 Deviations in Surgery Phase

The deviations that were seen in this phase was not very high since most of the surgeries and resections were grouped as “resection” (based on some domain background). There were thus less number of different activities that could replace the event.

Also, in most cases resections were recorded and performed as assumed and thus did not show significant problems.

6.2.3 Deviations in Chemotherapy Phase

The 5.7% fitness loss for “CAPOX” can be explained due to the fact that, first, the patients having Stage 3 cancer (3, 3A, 3B, 3C) for whom no chemotherapy was given, either can take the silent transition or go through the chemotherapy event transitions in the model. However, the silent transition is guarded by the \{Microsatelite_stability = no or Microsatelite_stability = unknown\} and a violation of this guard would cost higher than “move on model”. Therefore the cases would pass through the guard-less transitions, if \{Microsatelite_stability = yes\}. As explained previously, in practice the cost should be shared between “CAPOX” and “FOLFOX”.

Some of the other chemotherapy drugs given to patients are “oxaliplatin”, “bevacizumab”, and “systemic chemotherapy” which are present in the data but absent in the guidelines. These activities could be observed from the “Trace View” and the presence of such events contribute to the cost for the adjuvant chemotherapy phase.

Another point to mention is that there are some activities, in the model, that have a presence of multiple occurrences in the data, i.e. a diagnostic tests, resection, or/and a adjuvant chemotherapy was conducted more than one time. However, the guidelines do not specify this kind of a repeated event for the diagnosis episode. Thus this behavior/deviation can be considered expected as diagnosis and treatments can be conducted repeatedly if the doctor is not satisfied with the results. The presence of such multiple occurrences also can add to the “move in log” and thus reducing the transition fitness as well as overall model fitness.

One of the important decisions used in the guidelines for this phase is the “contra-indicatie oxaliplatin”. This decision is not recorded in the data as the decision is strongly based on the doctor’s experience. Due to the absence of this decision point, there is a silent transition in Stage 3 chemotherapy phase. This is however a very generalized structure with respect to the guidelines.

To restrict the model further, the silent transition for Stage 3 patients was removed to create a sensitive model. This strict model gave an overall fitness of 95.4% which is higher than the overall
fitness of the original model (94.7%), even though the fitness of the event “CAPOX” dropped from 94.3% to 65.7%.

The drop in the fitness for event “CAPOX” can be explained by the fact that cases/patients that were initially going through the silent transition now go through one of the activities “CAPOX” or “FOLFOX”. The reason these patients take one of the two activities and not “capecitabine” or “5-FU” is due to the fact that the events “CAPOX” and “FOLFOX” do not have a guard condition. As discussed before, the guards are given a higher cost value and the algorithm tries to take the path with the least cost. However, in practice the fitness and the cost should be divided between the two activities “CAPOX” and “FOLFOX”.

Due to the missing decision point and limitation of the algorithm, this ambiguous structure cannot be explained. Future work may be required to use some prediction techniques to strengthen this.

6.3 Question 3: Can we use process mining to obtain insights about the time between the diagnosis and treatment of cancer and do they adhere with recommended time?

Another dimension of the process model is shown by using the “Performance Mode” and using the “average time” as model measure. As we have considered only the start date for each activity, the average time is the average throughput time from the start of one activity to the start of the next activity.

The performance model is shown in Figure 6.4 from which we observe that it takes 4.6 days, 4.2 days, and 6.3 days (on average) for the diagnostic tests “CT abdomen”, “colonoscopy”, and “CT colonography”, respectively, from the start. Since the “START” is an artificial event and is given the timestamp as the start date timestamp of the first activity of the trace, having a throughput time for these three first trace events is abnormal. This is possible only if there are more events, in the log, that occur before the first three events.

By observing the “Trace View” it was evident that is is indeed the case due to which we see the abnormal behavior. However, a point to note is that the first phase of the process model is majorly the diagnosis phase and thus we can conclude that it takes less than 10 days, on average, form one diagnostics tests to another.

The next phase is the surgery/resection phase and it takes 24 to 26 days from the start of a diagnostics test to the start of a resection. The activity “active surveilance” is contextually not an event but similar to a “wait and see” procedure and 87.5 days, on an average, can be expected.

According to a research conducted in 2015 [6], the European Society for Medical Oncology (ESMO) recommend adjuvant chemotherapy starting from 4 weeks after surgery up to a maximum of 8–12 weeks. Since surgeries/resections do not normally take more than a day, the throughput time between “resection” and “adjuvant chemotherapy” can be considered as approximately the time between end of surgery/resection and start of adjuvant chemotherapy.
It takes approximately 4 weeks to 9 weeks, on an average, for the start of adjuvant chemotherapy. This follows the recommended number. Another point to be noted is that the chemotherapy “FOLFOX” and “5-FU” are started only after 8 weeks from resection, on an average. However, chemotherapy “CAPOX” and “Capecitabine” were started within 6 weeks from resection.

Figure 6.4: Performance Model showing the Average Throughput Time
6.4 Question 4: Can we observe significant deviations in the care-path based on cases in different types of hospitals, i.e. the academic, training and non-training hospitals?

As mentioned in the section “Overview of the Data”, there are nine academic hospitals, forty five training hospitals, and some general hospitals from which data is being recorded.

Training hospitals are the hospitals associated with the universities where training is given to the students. On the other hand, academic hospitals conduct clinical trials and more research and thus we expect to see more deviations in such hospitals. These kind of hospitals are more exploratory in nature when compared to training hospitals.

Other hospitals include both non-training hospitals and other hospitals that either not recorded in the Netherlands or involve a general practitioner.

Analysis of how well the data recorded from these hospital groups comply with the guidelines is interesting. We have used knowledge about the previous model obtained and have discover new models grouped on different hospital types. Then a statistical analysis was done to check the significance of the fitness values for patients diagnosed in each of the hospital groups.

The performance characteristics, based on the new parameters from Table 6.2, of the different hospital types are given below:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Academic Hospital Value</th>
<th>Training Hospital Values</th>
<th>Non-training Hospital Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Cases</td>
<td>584</td>
<td>5,174</td>
<td>2,963</td>
</tr>
<tr>
<td>Model Fitness</td>
<td>92.8%</td>
<td>94.8%</td>
<td>94.7%</td>
</tr>
<tr>
<td>Correct Events</td>
<td>2,071</td>
<td>21,374</td>
<td>12,329</td>
</tr>
<tr>
<td>Wrong Events</td>
<td>930</td>
<td>6,900</td>
<td>4,064</td>
</tr>
<tr>
<td>Missing Events</td>
<td>853</td>
<td>4,394</td>
<td>2,490</td>
</tr>
<tr>
<td>Data Violations</td>
<td>0.2%</td>
<td>0.3%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Event Violations</td>
<td>46.3%</td>
<td>34.6%</td>
<td>34.7%</td>
</tr>
</tbody>
</table>

Table 6.5: Performance Measures of the Model Grouped on Hospital Types

From the Table 6.5 we see that cancer data recorded from academic hospitals seem to have the least adherence with the guidelines. This is expected as these kind of hospitals conduct more exploratory research and clinical trials. Training and non-training hospitals are quiet comparable in the overall fitness values.

Other hospitals were not included. This is because we consider only hospital where the cancer was diagnosed and therefore this category of hospital group was not found based on diagnosis hospitals.

The fitness models of the three different hospital groups are given in Figure 6.5, Figure 6.6, and Figure 6.7.
**Academic Hospitals:** Since, the academic hospitals show lowest adherence between data and guidelines, we observe individual event fitness values. “CT colonography” has a fitness of 7.7% and this is lower than that of the original model which has a fitness of 17.1%. There were 33% cases, in this hospital, that did not go through a “colonoscopy” nor a “CT colonography”.

There were approximately 43% cases where “CT abdomen” was not performed and about 75% cases where no “X thorax” was performed. There are some cases overlapping between the two mentioned events where none of the two events were performed. This deviation is quiet high when compared to the original model. Other events follow adherence patterns similar to that of the original model in [Section 6.2](#).

![Figure 6.5: Fitness Model of the Academic Hospital](image)
**Training Hospitals:** There were 24.5% cases, in the training hospitals, that do not go through a “colonoscopy” nor a “CT colonography”. This deviation is lower than the deviation of this transition fitness in the original model obtained in **Section 6.2**.

There were approximately 18.6% cases where “CT abdomen” was not performed and about 44% cases where no “X thorax” was performed. There are some cases overlapping between the two mentioned events where none of the two events were performed. This deviation is comparable to the original model. Other events also follow patterns similar to that of the original model.

![Fitness Model of the Training Hospital](image-url)
Non-training Hospitals: There were 29% cases, in the non-training hospitals, that do not go through a “colonoscopy” nor a “CT colonography”. This deviation is lower than the deviation of this transition fitness in the original model obtained in Section 6.2.

There were approximately 15% cases where “CT abdomen” was not performed and about 42% cases where no “X thorax” was performed. There are some cases overlapping between the two mentioned events where none of the two events were performed. In this hospital group, “CT abdomen” is being performed more frequently than in academic hospitals or training hospitals. However, the overall guideline adherence for training and non-training hospitals are comparable.
From Table 6.5, we can see difference in fitness values, however, statistical tests were conducted to see if this difference was statistically significant. The tests were performed using SAS.

Using the thesis research that Mark Vroling is working on, the fitness values for each patient was extracted. This was done by using the same conformance checking methodology as used by plugin “Multi-perspective Process Explorer” and further enhancing was done in order to extract fitness scores for each case/patient.

In order to check the difference in fitness scores between the three hospital groups, an ANOVA test was done. Since one of the basic assumptions of ANOVA is that the data being compared should have a normal distribution across each group, first the fitness scores were checked for normality. The fitness scores of the patients, within each hospital group, rejected the test for normality. However, since the number of samples per group is large, we can consider the central limit theorem and thus avoid the normality condition.

The three hospital groups did not have equal sample sizes and thus a Leven’s test was done to check for homogeneity of sample variance within each group. This test was rejected and a non-parametric test (Kruskal Wallis test) was then done using the SAS procedure NPAR1WAY. The chi-squared p-value is less than 0.05 and thus rejecting the null hypothesis. Therefore we can conclude that there is a significant difference of fitness between the three hospital groups.
Chapter 7

Conclusion and Future Work

7.1 Conclusion

The data recorded by the Netherlands Cancer Registry (NCR), for the year of diagnosis 2015, was mapped to the guidelines set up by the Netherlands Comprehensive Cancer Organization. The cancer data was extracted only for non-metastatic colon cancer diagnosed patients and the guidelines were analyzed only for the diagnosis episode.

On mapping the extracted data with the model from the guidelines it was found that the overall fitness of the model was relatively good and using the information from the model and the data, the model was further improved such that the data strictly follows the flow of the guidelines. Thus the guidelines behaved as the baseline model against which the data was mapped to. The fitness of the revised model was considered to be stronger which comprehends to the fact that the NCR data complies with the overall guideline structure.

To analyze the deviations in the model, the transition fitness of various events along with the log traces were explored. It was discovered that events corresponding to the diagnostic tests faced the highest degree of cost on fitness. Event “CT abdomen” showed the highest adherence to the diagnostic guidelines and “X thorax” showed the least adherence. Not all patients were diagnosed with cancer using the same diagnostic tests as mentioned in the guidelines.

Another point to be noted was that resection/surgery was also performed on patients that had a clinical tumor stage of 0 or “is” (in situ) even after no acute presentation (emergency for surgery). According to the guidelines this was not a choice to be performed. Lastly, chemotherapy given to patients adhere to the guidelines, however, there were more chemotherapy drugs given to patients in some cases that were not mentioned in the guidelines.

One of the important perspectives of process mining is the time perspective that was used to discover the average time between the different phases of the diagnosis episode. The average duration between the two phases, resection and adjuvant chemotherapy, was confirmed with existing research and recommendation. The methodology was also able to go one step ahead by discovering the average time between resection stage and the four individual chemotherapy drugs, rather than just the two phase. A sensitivity analysis was also performed for the chemotherapy phase due to
discrepancies in the guidelines for this phase.

In the Netherlands, hospitals are grouped into different types based on factors such as “does the hospital perform clinical trails”, “does the hospital conduct research”, “does the hospital follow routine procedures”, “is the hospital associated with universities”, etc. Based on such factors the hospitals, where first diagnosis was done, were grouped into three main groups: academic hospitals, training hospitals, and non-training hospitals. It was observed that the cancer treatment care-path in the academic hospitals deviates the most from the guidelines. Training and non-training hospitals follow the guidelines with fewer deviations. Furthermore, there was a significant difference in the fitness scores of the patients between the different hospital groups.

7.2 Limitations and Future Work

In this report we validate the guidelines with only patients that have been diagnosed in the year 2015 with non-metastatic colon cancer. However, the validation can be extended with patients diagnosed after 2015. IKNL has been revising and extending the colon guidelines and the model can be further extended with the revisions made. The model can also be extended for Stage 4, however, the data does not record many key decision points which are strongly required for Stage 4 cancer patients.

The decision points are important in order to design a strong baseline model and this is either missing or have a high level of unknowns in the Stages 1, 2, and 3 and due to this there still exists a gap between the guidelines and our baseline model. An example of this is in the chemotherapy phase where only an adjuvant chemotherapy (chemotherapy after resection based on pathology staging) is performed in the model, however, in the guidelines there is a care-path where the option of a neo-adjuvant chemotherapy (chemotherapy before resection based on clinical staging) may also be given under the circumstance where no surgery can be performed. This decision is either missing in the data or unknown for many cases and thus not implemented in our model. Some of the other missing decision points are “snijvlak”, “contra-indicatie oxaliplatin”, and “Perforatie bij presentatie”.

As mentioned earlier we only consider the diagnosis episode and thus do not include the recurrence phase, where cancer re-occurs after complete treatment lifecycle. This phase can added to the model and re-checked with patients that were diagnosed in 2015. This, of course, can be implemented only after a few years from diagnosis as the recurrence of cancer needs to be recorded. In this way all the discrepancies that occurred due to the reverse order of the different phases can also be justified and corrected for. On the same line, multiple occurrences of events present in the data can also be fixed either in the model re-created from the guidelines or by understanding re-occurring hospital events and applying some domain knowledge.

On the Oncoguide, colon and rectum cancer have been grouped together to represent colorectal cancer guidelines and this is because the data and care-path for colon and rectum cancer are similar in many aspects. Extending the re-created colon guidelines for rectum cancer can help to validate the care-path for the entire colorectal cancer care-path.
Bibliography


[8] https://www.oncoguide.nl


[12] https://www.sas.com/
Appendix A

Extra Images

Figure A.1: Alteryx Workflow
Figure A.2: Semantic Analysis for Petri-Net obtained from WoPeD

Figure A.3: Fitness Legend