PROCESS MINING FOR BREAST CANCER PATIENTS' CLINICAL PATHWAY

A Case Study at Helsinki University Hospital

Lappeenranta–Lahti University of Technology LUT

Master's Programme in Business Analytics, Master's thesis

2022

Paula Tanni

Examiners: Associate Professor Jan Stoklasa

Post-Doctoral researcher Jyrki Savolainen

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT School of Business and Management

Business Administration

Paula Tanni

Process Mining for Breast Cancer Patients' Clinical Pathway: A Case Study at Helsinki University Hospital

Master's thesis

2022

76 pages, 17 figures, and 5 tables

Examiners: Associate Professor Jan Stoklasa, Post-Doctoral researcher Jyrki Savolainen

Keywords: Process mining, health care data, clinical pathway

Process mining is an emerging discipline that can be used for data-driven analysis of real-world processes. This master's thesis aims to recognize how process mining is currently applied in healthcare analytics, harness the process mining technology on real-world hospital data, discover what the typical processes look like for breast cancer patients at Helsinki University Hospital, and how they compare to the planned process model.

The first part of the study was conducted as a literature review and the second part as a case study with the Process Mining for Python (PM4Py) framework. The data for the case study was collected from HUS Data Lake from the years 2020 and 2021 and consisted of 855 patients' appointments at Helsinki University Hospital. The patients were split into two subcategories based on their treatment protocol and only the variants with the highest frequencies were included in the analysis.

For the given data, the process for typical clinical pathways for two different patient groups was challenging to discover and they only covered the pathway of 182 patients. The discovered typical pathways and the planned process models were very similar, and the discovered time delays between the activities were close to estimated. The results of the thesis show that process mining can be applied for improving the efficiency and patient experience of patient treatment. The use of process mining tools can give new insights into the medical processes and give feedback on how the actual process compares to the best practice guidelines.

TIIVISTELMÄ

Lappeenrannan–Lahden teknillinen yliopisto LUT

LUT-kauppakorkeakoulu

Kauppatieteet

Paula Tanni

Rintasyöpäpotilaiden hoitopolun prosessilouhinta – tapaustutkimus Helsingin Yliopistollisessa Sairaalassa

Kauppatieteiden pro gradu -tutkielma

76 sivua, 17 kuvaa ja 5 taulukkoa

Tarkastajat: Apulaisprofessori Jan Stoklasa, Tutkijatohtori Jyrki Savolainen

Avainsanat: Prosessilouhinta, terveydenhuollon data, kliiniset hoitopolut

Prosesssilouhinta on kasvava tieteenala, jota voidaan käyttää reaalimaailman prosessien datalähtöisessä analyysissä. Tämän pro gradu -tutkimuksen tavoitteena oli tunnistaa analytiikan käyttömahdollisuuksia prosessilouhinnalle terveydenhuollossa, soveltaa prosessilouhintamenetelmiä terveydenhuollossa tallennetuille tiedoille ja löytää rintasyöpäpotilaiden tyypilliset hoitopolut Helsingin Yliopistollisessa Sairaalassa, sekä vertailla löydettyjä prosessimalleja potilaille suunniteltuihin hoitoprosesseihin.

Tutkimuksen ensimmäinen osio toteutettiin kirjallisuuskatsauksena ja toinen osio tapaustutkimuksena. Työkaluna käytettiin Process Mining for Python (PM4Py) ohjelmistokirjastoa. Tapaustutkimukseen tarvittava data kerättiin HUS Tietoaltaalta vuosilta 2020–2021 ja se koostui 855 potilaan potilaskontaktitiedoista Helsingin Yliopistollisessa Sairaalassa. Potilaat jaettiin kahteen alaryhmään heidän hoitoprotokollansa mukaan ja analyysiin sisällytettiin vain ne hoitopolut, joilla oli suurin frekvenssi.

Potilaiden tyypilliset hoitopolut olivat haastavaa löytää prosessilouhintamenetelmällä datasta ja löydetyt tyypilliset hoitopolut edustivat yhteensä vain 182 potilaan hoitopolkua. Analyysillä löydetyt tyypilliset hoitopolut ja potilaille suunnitellut hoitopolut olivat hyvin samankaltaiset ja aikaviiveet aktiviteettien välillä olivat lähellä arvioituja aikaviiveitä. Tämän tutkimuksen tulokset osoittavat, että prosessilouhintamenetelmiä voidaan soveltaa potilaan hoitoprosessien tehokkuuden ja hoitokokemuksen parantamiseksi. Prosessilouhintatyökalut voivat antaa uusia näkökulmia hoitoprosesseihin ja se mahdollistaa toteutuneiden prosessien ja hoitosuositusten vertailun.

Table of contents

Figures

Tables

1. Introduction

The health and social service reform of Prime Minister Sanna Marin's Government program aims to reduce the inequalities in health, improve the availability and accessibility of services, and control the growth of costs of healthcare (Finnish Government 2021). Breast cancer is the most common cancer type in Finland for females, with approximately 5000 cases per year (Joensuu & Rosenberg-Ryhänen 2014; Mattson, Auvinen, Bärlund & Jukkola-Vuorinen 2016). The volume of breast cancer patients is significant at Helsinki University Hospital (HUS), with about 1500 breast cancer patients annually (HUS Syöpäkeskus 2018). Electronic information systems in healthcare record valuable patient event data, such as appointments and laboratory and radiology tests, which can be used for improving patient care. This data can be utilized in decision making, for a better understanding of care processes and to increase the efficiency of healthcare. (Garcia, Meincheim, Junior, Dallagassa, Sato, Carvalho, Santos & Scalabrin 2019; van der Aalst 2016, 400)

In order to find ways to improve clinical pathways and patient treatment, healthcare processes should be thoroughly examined and analyzed (Erdogan & Tarhan 2018; Martin, De Weerdt, Fernández-Llatas, Gal, Gatta, Ibáñez, Johnson, Mannhardt, Marco-Ruiz, Mertens, Munoz-Gama, Seona, Vanhienen, Wynn, Boilève, Bergs, Joosten-Melis, Schretlen & Van Acker 2020). Healthcare processes can be split into medical treatment processes and organizational business processes (Lenz & Reichert 2007; Rovani, Maggi, de Leoni & van der Aalst 2015). In healthcare, there may be differences between the best practice guidelines and the actual clinical patient care (Rovani et al. 2015). This can occur because of a lack of best practice guidelines for the process, or the guidelines represent idealized scenarios that do not materialize in practice (Rovani et al. 2015). The best practice process models cannot take into account the complexity of the process and the distribution of several sub-activities often involved (Rebuge & Ferreira 2012).

Business process models are traditionally made by hand for analyzing and gaining a better understanding of the process (Van Der Aalst 2011). Process mining is an emerging discipline that can be used for data-driven analysis of real-world processes to uncover the true behavior and performance of business operations (van der Aalst 2016, 3). It is a sub-field of data science (Berti, van Zelst & van der Aalst 2019) and a method combining computational

intelligence and data mining as well as process modeling and analysis (Daniel, Barkaoui & Dustdar 2011, 172; Van Der Aalst 2016, 31). Process mining is a research field that focuses on generating techniques that can extract insights into processes from the action execution data. This data is useful and provides previously unknown information for evidence-based process improvement. (Martin et al. 2020) For organizations, process improvements can reduce costs and response times (Buijs, van Dongen & van der Aalst 2012). Process mining provides tools that enable fact-based insights into the process and therefore support the improving processes (Van der Aalst 2016, vii).

In healthcare, process mining techniques can help to identify the differences between the ideal clinical practice and reality (Rovani et al. 2015) as well as improve health management and the quality of care by finding the best practices (Dallagassa, Garcia, Scalabrin, Ioshii & Carvalho 2021).

1.1. Purpose of the Study

This thesis aims to harness process mining technology for analyzing the clinical pathways of breast cancer patients treated with radiation therapy at Helsinki University Hospital. By executing process mining on event logs from patients' contacts at HUS, this study seeks to find the typical clinical pathways of patients.

This study will increase the knowledge of process mining tools at HUS IT Management and the results of the study will give new insights about breast cancer patients' treatment processes to the clinicians at HUS Comprehensive Cancer Center. The research is a pilot study for creating an analytical toolset for process mining at Helsinki University Hospital. The toolset could be used for reporting the outliers of patients falling off the typical clinical pathway. In the future, this reporting could provide notifications for the demand of an intervention with the patients to the clinicians of HUS and hence limit the risk of prolonging the patient's treatment.

This thesis aims to answer the two following research questions:

1. How is process mining currently applied in healthcare analytics?

This question will be answered through a literature review of process mining studies in the healthcare field.

> 2. What do process mining discovery models of breast cancer patients' look like at Helsinki University Hospital and how do they compare with the planned process model for breast cancer patients?

Research question 2 will be answered with a process discovery model, generated with the PM4Py-tool from breast cancer patients' data. The discovered model will be compared with the handmade process model for conformance.

1.2. Delimitations, Limitations, and Assumptions

The researcher can control the data pre-processing actions and interpret the results of the analysis. The data extraction can be controlled by the researcher to some extent since the researcher can decide the source systems and the size of the cohort. The filtering and transformation of the data are controlled by the researcher as well as choosing the analytical tools for process mining.

There are some limitations to this study. The researcher does not have control over the content of the data, meaning the data is as adequate as it is recorded in the information system. There is likely some incorrectly recorded or missing data that will be challenging to detect and correct. Also, even though the researcher can control the tools used for analysis, she will not do modifications to the existing tools but uses the tools as they currently are. The open-source algorithms may be further developed in the future, which may alter the results of the study if repeated later.

The researcher assumes that the data recorded in the information systems represent reality, is valid and recorded in the right way. The timestamps and activities of breast cancer treatments are correctly saved in the databases and integrations to the data lake are updating correctly. The researcher also assumes that the process mining algorithms are well functioning and give trustworthy results.

The thesis will follow the guideline for responsible conduct of research by the Finnish national board on research integrity in all stages of the study. All phases of the study will be

thoroughly explained to make them transparent and repeatable in the future. References are used for granting the previous researcher's the credit they deserve. (TENK 2012)

2. Theoretical Background

The theoretical background in this thesis concentrates on the basics of process mining approaches, how they can be applied and what kind of tools have been developed for it. Breast cancer treatment is explained on a general level.

2.1. Process Mining

Process mining aims to discover, monitor, and improve real processes with extracted action execution data from existing information systems. The information systems record gigantic amounts of details of the activities. The basic requirements for process mining data are that it contains both a unique identifier for each case and a label for each activity. (van der Aalst 2012; van der Aalst 2016, 31–32, 276 Martin et al. 2020) In this thesis, the data gathered from the actions within the breast cancer treatment process, are called event logs.

This chapter explains the key concepts of process mining. First, the process mining perspectives and types are introduced, and examples of different process mining graphical presentations are explained. This is followed by the tools, techniques, and algorithms for process mining.

2.1.1. Aspects of Process Mining

Depending on the aim of the process mining research, analysis can be conducted in four different perspectives: *control-flow perspective, organizational perspective, case perspective, and time perspective.* Control-flow perspective aims to find the ordering of the activities, expressing them in notation such as Petri net, BPMN, or UML. The organizational perspective focuses on the resources related to the event log, finding which actors, such as people or departments, are involved in the process and how they are related. The case perspective investigates how a certain case is characterized in the process and the time perspective

focuses on the frequency or timing of events. This enables analysis of service levels, predicting processing time or bottlenecks in the process. Different perspectives of process mining can be overlapping to some extent. (van der Aalst 2016, 34, 275)

In this thesis, the process discovery for breast cancer patients focuses on the control-flow perspective since the purpose is to find the common type of events and the ordering of them in patients' clinical pathways. When one is comparing the discovered model to the hand-drawn model of the clinical pathway, the emphasis is also on the time perspective as one can examine the delays between different events and identify patients that do not follow the typical pathway this way.

There are three main types of process mining: *discovery*, *conformance,* and *enhancement* (van der Aalst 2012; van der Aalst 2016, 31–32; Martin et al. 2020). The main types will be introduced in the next chapters.

2.1.2. Process discovery

The process discovery technique takes event log data as input to create a model without using any previous knowledge of the process. A process discovery model can be obtained with different algorithms that are more thoroughly explained in chapter 2.2. (van der Aalst 2016, 33; Leemans 2017). Process discovery is a starting point for another type of process mining analysis.

There are many challenges with process discovery and compromising between the different techniques and their challenges depend on the goals of the process mining analysis. Process mining aims to create a *high fitness* model, a generalized model that represents the underlying pattern of events. The discovered model should comply with the event log well. (van der Aalst 2016, 38; Leemans 2017; Leemans, Fahland & van der Aalst 2018) There is a risk of creating a model that is either *overfitting* or *underfitting*, overfitting meaning the model represents the reality in a too detailed manner and it will not represent new event log data. On the other hand, if the model is underfitting, it is too general and allows unrelated behavior. (van der Aalst 2016, 38) According to Leemans (2017), the discovered model has high *log-precision* or does not underfit the event log if the event log data includes all behavior of the discovered model. If the log-precision would be low and the model showed more behavior

than the event logs include, it could lead to concluding the behavior that is absent in the model.

However, including the missing behavior may be necessary when representing the behavior of the system. Process discovery techniques do not usually assume that the event log includes all possible behavior of the system since the cohort usually does not represent all possible traces in the system. The algorithms should generalize the behavior in the event log to reach conclusions about the behavior of the system. (Leemans 2017)

If the process mining study aims to analyze only the majority of the behavior, then the fitness guarantees are not relevant, as only the most common behavior should be included (Leemans 2017). Event logs may contain a lot of *noise* or exceptional behavior that should not be included in the model (van der Aalst 2016, 39).

Another challenge is having sound semantics of the data. Machines and software cannot interpret models that do not have well-defined semantics. The conclusions drawn from such a model are not guaranteed to be correct. Also, the conformance checking cannot be reliably applied to such data. The model itself should be *sound*, or not have any deadlocks or anomalies. (Leemans 2017)

The *rediscoverability* can also be a challenge. The process discovery model should ideally represent a similar set of traces as the system, meaning it can rediscover the language of the system the event log was extracted from. Measuring rediscoverability relies on assumptions about both the system and the event log data. (Leemans 2017; Leemans et al. 2018)

The discovered model can be represented in, for example, a *Business Process Model and Notation, Petri Net,* or *Process Tree* format. These will be introduced in Chapter 2.3.

2.1.3. Conformance Checking

Conformance checking means verifying the process model against reality. An existing process model can be used for measuring and screening how the extracted event log data or discovered process model compares to it. This knowledge can be used for identifying deviations in the logs and measuring their severity. (Buijs et al 2012; van der Aalst 2016,33;

Leemans 2017; Leemans et al. 2018) Conformance checking measures can be used to evaluate, for example, the model fitness and log-precision described in chapter 3.2.

Conformance checking can be approached from two different perspectives. When the process model aims to be *descriptive* or capture reality, conformance checking measures how the process model captures the real behavior, and it aims to repair the model. When the model is *normative* or influences reality, conformance checking concentrates on measuring how the reality deviates from the ideal model, meaning the deviations in the event log are under examination. (van der Aalst 2016, 39) This thesis' conformance checking part focuses on the latter example, as I am not trying to alter the process model but rather find the deviations in the data gathered from patient information systems.

2.1.4. Process Enhancement

The third type of process mining, enhancement, means finding ways to improve the existing process model with information from event log data. Process enhancements can be sub-categorized as *repair* and *extension* type enhancements. Process repair can be used for repairing the process model to better replicate reality. The enhancement can also be an extension, where a new perspective is included in the model by cross-correlation with the event log. The extensions can be, for example, showing frequencies, resources, bottlenecks, and throughput times. (van der Aalst 2016, 33)

2.2. Tools and techniques for Process Mining

The best techniques for healthcare process mining are the ones that can deal with a large amount of noise and enable the sorting of different behaviors making it possible for separate analysis (Rebuge & Ferreira 2012; Dallagassa et al. 2021). There are many software tools available, both open-source and commercial, and the software is often accessible through a graphical interface (Berti et al. 2019). Open-source tool ProM has been a popular tool in recent studies, and it has many plug-ins for various process mining approaches (Rebuge & Ferreira 2012; Partington et al. 2015; Rovani et al. 2015; Erdogan & Tarhan 2018; Kurniati et al. 2019; Helm, Lin, Baumgartner, Lin & Küng 2020; Pika et al. 2020). The Fuzzy Miner

algorithm, as used in a commercial tool called Disco, can be used for discovering health process models, as it provides a deeper understanding and can assist with reducing the complexity and challenges of complicated healthcare models (Erdogan & Tarhan 2018; Helm et al. 2020; Dallagassa et al. 2021).

Process mining tools have been further developed in recent years, and self-developed, case-specific approaches have been popular, often used together with tools such as ProM. *Inductive visual miner* is a built-in miner in ProM and has been most popular in recent studies. (Helm et al. 2020) In Garcia et al.'s (2019) comprehensive systematic mapping of 1278 research articles, the most popular techniques for process mining were Heuristic Miner, Alpha and variations, and evolutionary-based. Helm et al. (2020) suggested that the use of Heuristic miner algorithms may be recently decreasing since in their literature review only two out of 38 of the studies selected for review used the Heuristic miner algorithm. However, Dallagassa et al. (2021) found in their literature review that Heuristic Miner was used in 25% of their research articles indicating no decrease in its use.

2.2.1. Inductive Miner

Inductive miner (IM) is a type of algorithm for process discovery. Some of the advantages of inductive miner are that the discovered models correspond to block-structured workflow net systems and some of the challenges of process mining discovery can be well avoided. Inductive miner algorithms enable perfect fitness meaning they can generate all the behavior of the event log in the discovered model (Ghawi 2016; Leemans 2017) and also high log precision and rediscoverability (Leemans 2017).

Inductive miner algorithms are designed to perform four different functions:

1) *FindCut*: Identifying cuts in the event log data turning them into process tree operators
2) *SplitLog*: Splitting the event log data
3) *BaseCase*: Handling the individual events
4) *FallThrough*: Handling the exceptional cases

These four functions are generally considered independent, but some algorithms may rely on returned values of another function. (Leemans 2017)

The inductive miner algorithm constructs a process tree recursively from top to down (Leemans 2017) and works recursively with the divide and conquer strategy (Ghawi 2016). The discovery algorithm first detects the most significant behavior of the event log data, creating the root of the process tree. It then divides the activities of the event log into sub-logs. The sub logs are separately split into smaller sub logs. These splits of sub logs are recursed until the algorithm finds a base case, meaning that dividing the log into subsets is no longer possible. (Ghawi 2016; Leemans 2017) There is a possibility that the sub log contains empty traces before finding the base case, and a *fall through* or handling of the exception is needed, making it possible for recursion to continue. This improves the soundness of the model as there will be no deadlocks. (Leemans 2017) The discovered process tree can be transformed into a Petri net (Ghawi 2016).

2.3. Process-Modeling Notations

An important part of process mining is finding the appropriate way to present the discovered model. Process mining results can be expressed in various notations which vary in the information they can provide. (van der Aalst 2019)

2.3.1. Process Tree

A process tree is an abstract representation of a hierarchical block-structured workflow net. It is shaped like a tree with two kinds of nodes: branch nodes and leaf nodes (Figure 1). The leaves are *activity labels* or basic process steps and the branch nodes that are connecting the leaves are *operators*. (Buijs et al. 2012; Leemans, Fahland & van der Aalst 2013; Leemans et al 2018)

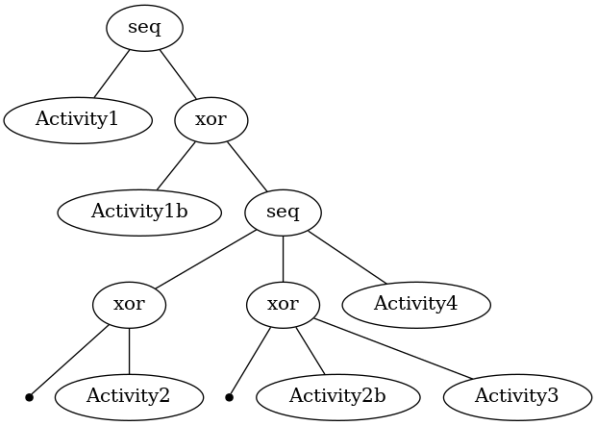

Figure 1. An example of a process tree

The operators have one or more children, and their labels describe how their children should be combined (Buijs et al. 2012; Leemans, Fahland & van der Aalst 2013; Leemans et al. 2018). One of the benefits of algorithms that produce process discovery into a process tree is that it guarantees a sound model (Leemans et al. 2018).

2.3.2. Petri Net

Petri nets are a commonly used and examined modeling language, which allows the modeling of concurrence events. Petri nets are executable and can be used for analysis with many different techniques. (van der Aalst 2016, 59–61, 65)

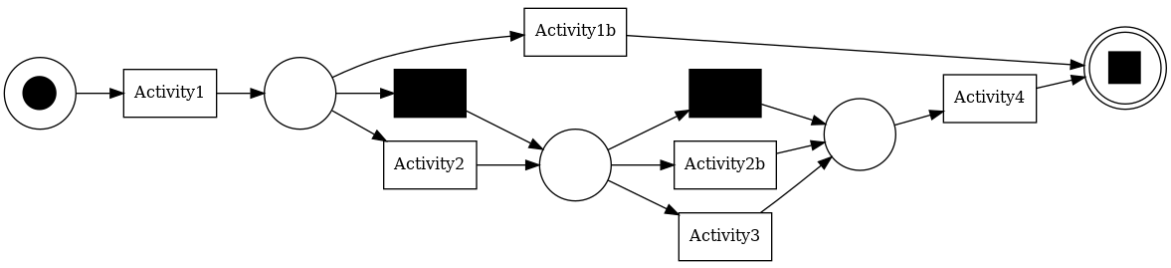

Figure 2. An example of a Petri net

The graphical notation of Petri nets is intuitive and consists of places and transitions (Figure 2), where tokens flow through the network. A WorkFlow net is a subclass of a Petri net, with a dedicated source place for process start and end. (van der Aalst 2016, 59–61, 65)

2.3.3. Directly-Follows Graph

Directly-Follows Graph (DFG) is a graph with nodes that represent activities and edges with the direction that correspond to the directly-follows relationship. The nodes and edges can also include information about the frequency and timing of the activities (Figure 3). (van der Aalst 2019)

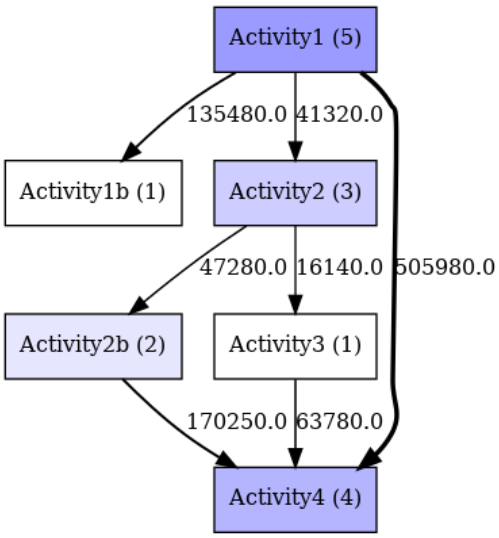

Figure 3. An example of a Directly-Follows Graph

The graphs can be filtered with a threshold on the frequency of the traces, the number of events, and the number of direct successions for each relation. The directly-follows graphs do not display the transition types between the activities like Petri nets and process tree. (van der Aalst 2019)

2.4. Breast Cancer Treatment

This chapter describes how patients are diagnosed and treated with breast cancer in Finland. Usually, the patient seeks medical help after finding a lump in her breast or by breast cancer screening (Terveyskylä 2022a). The treatment process is determined by the cancer type and patients' wishes (Vehmanen 2020).

2.4.1. Breast Cancer Surgery

Patients are diagnosed with breast cancer at the primary, occupational, or private health care unit with a biopsy or the symptoms and radiology tests indicate possible breast cancer, after which they will be referred to breast cancer surgery. After the referral, breast cancer surgery contacts the patient for instructions and books the first appointment as well as further examinations if needed. (HUS 2021) The first appointment (Figure 4) takes place usually about a week after HUS has received the referral (HUS 2020a). During the appointment, a patient meets with the surgeon, breast cancer nurse, anesthesiologist or anesthesia nurse (HUS 2021). Approximately three weeks after HUS has received the referral (HUS 2020a), the surgery is usually performed as a breast-conserving surgery where the tumor is removed with clean margins. If the tumor is too large to be removed with conserving surgery or it is not possible to treat the patient with radiation therapy after the surgery, the whole breast can be removed with mastectomy surgery. The breast can also be removed if the patient wishes it. In some breast cancer cases, the patient can also be treated with chemotherapy before the surgery. (Suomen rintasyöpäyhdistys 2021)

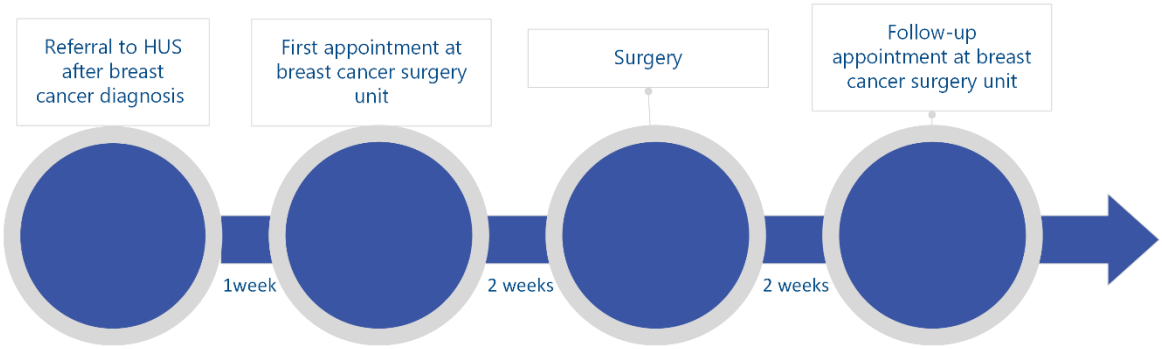

Figure 4. Process of breast cancer surgery

Approximately two weeks after the surgery there is a follow-up appointment where the patient's state of health is assessed, and the patient receives information about the results of the surgery. (HUS 2021) After this appointment, the patient's medical treatment plan will be determined in a multi-professional meeting and treatment will be customized individually, based on national and international recommendations. (HUS 2021; Terveyskylä 2022b)

2.4.2. Adjuvant Therapy

The first appointment with an oncologist is typically after two weeks from the follow-up appointment at the breast cancer surgery unit (Figure 5). This appointment is meant for planning the treatment and can be at the Chemotherapy or Radiation therapy unit, depending on which medical treatment plan the patient will follow. (HUS 2021; Terveyskylä 2022b) If drug therapy is tentatively planned or the patient will not receive radiation therapy at all, the appointment is at the chemotherapy unit. If the patient is scheduled to have radiation therapy only or radiation therapy combined with hormonal therapy, the first oncology appointment will be at the radiation therapy unit (HUS 2021). After this appointment, the treatment begins usually about a month after surgery once the wounds have fully healed (HUS 2021; Terveyskylä 2022b).

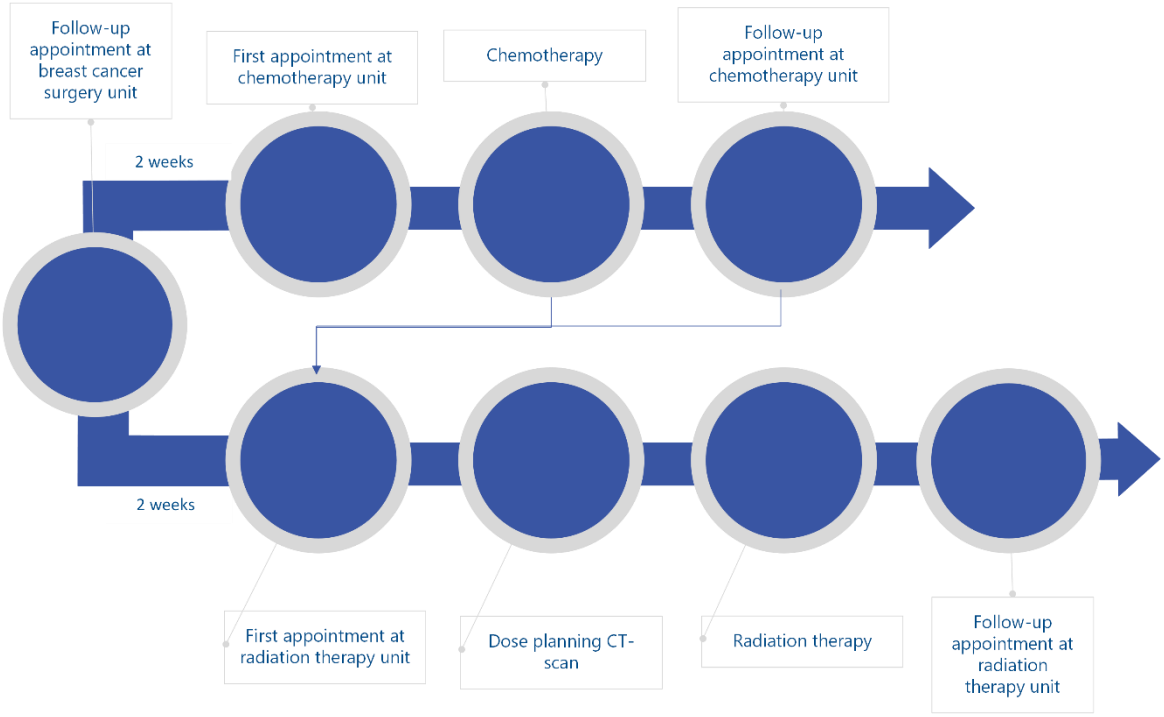

Figure 5. Process of chemotherapy and radiation therapy

Chemotherapy is usually administered 6–8 times every 3 weeks after the surgery and before the radiation therapy period. There are drug therapies that target specific breast cancer types. These medications are administered with a medication-specific protocol and can be administered during and after the radiation therapy period. (Suomen rintasyöpäyhdistys 2021) The patients can also be prescribed hormonal medication for many years after the surgery (Suomen rintasyöpäyhdistys 2021; Terveyskylä 2022c).

Radiation therapy is usually recommended for patients who had breast-conserving surgery. Radiation therapy starts with an oncologist and nurse appointments after the possible chemotherapy. After the appointment, a computer tomography scan is performed in the treatment position and the images are used for dose planning. About a week after the CT scan the radiation therapy starts, and it is administered daily 15–30 times. At the end of the radiation therapy period, the patient has another appointment where the current symptoms are evaluated, and future follow-ups are planned. (Suomen rintasyöpäyhdistys 2021; Terveyskylä 2022d)

2.4.3. Follow-ups After Breast Cancer Treatment

After the possible chemotherapy and radiation therapy, the patient usually has a follow-up appointment with an oncology nurse or an oncologist where future follow-ups are planned. Usually, the follow-up period lasts between 5 and 10 years after the treatments and contains both radiological examinations and appointments with either an oncology nurse or an oncologist. (Terveyskylä 2022e)

The follow-ups aim to find possible relapse as early as possible and to monitor the possible hormonal therapy symptoms. Process mining for the follow-up appointments and possible treatments for relapses for breast cancer patients is out of scope for this thesis.

3. Literature Review

The first part of this study was a literature review of previous research. The literature on process mining was collected in November 2021. This was conducted to find different process mining research frames and to gain a general knowledge of what kind of research has already been conducted and what kind of research is still required. The literature was searched through LUT Primo library services with the keyword “process mining” and the results were complemented through the snowball method from found articles.

The literature on process mining in healthcare was collected in December 2021. The articles were searched and obtained from the LUT Primo library services with the keywords “Process mining” and “healthcare” and sorted by relevance. The article needed to be published in a peer-reviewed journal. The articles were first chosen based on their title, then the abstract, and finally the whole text. All results were not included since the search resulted in a large number (> 30 000) of research articles. Especially literature reviews were emphasized since they summarized many individual research results.

The studies were arranged in a tabular form and basic information about each research was written down such as the main results and further study ideas. The studies were analyzed for finding different and common ways of applying process mining in healthcare.

3.1. Previous Studies on Process Mining

Process mining has been successfully applied in many industries or domains, most notable being healthcare, ICT, Manufacturing, Education, and Finance (Garcia et al. 2019). Many previous studies have been conducted on process mining in recent decades (Berti et al. 2019). Studies have focused on algorithm development (van der Aalst, Weijters & Maruster 2004; Adriansyah, Sidorova & van Dongen 2011; Leeman, Dirk & van der Aalst 2015; Berti et al. 2019), case studies for process discovery (Rebuge & Ferreira 2012; Arias, Rojas, Aguirre, Cornejo, Munoz-Gama, Sepúlveda & Capurro 2020), conformance checking (Rovani et al. 2015) and process analysis or enhancement (Baek, Cho, Kim, Hwang, Song & Yoo 2018;

Valero-Ramon, Fernandez-Llatas, Valdivieso & Traver 2020), many studies also combining these process mining types (Erdogan & Tarhan 2018; Ibanez-Sanchez, Fernandez-Llatas, Martinez-Millana, Celda, Mandingorra, Aparici-Tortajada, Valero-Ramon, Munoz-Gama, Sepúlveda, Rojas, Gálvez, Capurro & Traver 2019). Also, studies on making the process mining log data more useful have been conducted, from a data quality perspective (Kurniati, Rojas, Hogg, Hall & Johnson 2019), data security perspective (Pika, Wynn, Budiono, ter Hofstede, van der Aalst & Reijers 2020) and filtering the data (Rabbi, Lamo & MacCaull 2020; Vidgof Djurica, Bala & Mendling 2020).

Previous studies had research questions or study aims related to successful log data preparation and transformation for process mining (Rebuge & Ferreira 2012; Partington, Wynn, Suriadi, Ouyang & Karnon 2015; Djurica et al. 2020), process inspection through discovered model (control-flow) (Rebuge & Ferreira 2012; Erdogan & Tarhan 2018), process mining model performance (Rebuge & Ferreira 2012; Rovani et al. 2015; Ibanez-Sanchez et al. 2019), comparing process model through process mining and ideal situation (Rovani et al. 2015; Erdogan & Tarhan 2018), finding ways to improve the process (Erdogan & Tarhan 2018) and ways to identify characteristics that affect the process (Baek et at. 2018; Ibanez-Sanchez et al. 2019; Arias et al. 2020; Valero-Ramon et al. 2020).

3.2. Process Mining in Healthcare

Process mining has demonstrated good quality results in the healthcare domain (Garcia et al. 2019; Arias et al. 2020). The process mining studies in healthcare cover clinical pathways, patient treatment, and the primary processes of a hospital (Garcia et al. 2019). Real-world healthcare issues should be the starting point for research (Martin et al. 2020). In healthcare, the log data routinely stored in the electronic health records can be used for process mining (Arias et al. 2020). By mining the healthcare process, customer experience can be improved (Arias et al. 2020; Martin et al. 2020).

Continuous efforts by researchers and the research community are required to develop the usability and understandability of process mining in healthcare (Martin et al. 2020). Healthcare process mining is relatively new and underutilized (Rebuge & Ferreira 2012; Partington et al. 2015; Garcia et al. 2019) and can affect identifying patient flows for certain

diseases, treatment chances, and correlations, decision making, cost management, and quality of care. With process mining applications it is possible to evaluate performance, and clinical pathways, check if medical protocols and guidelines are followed in a clinical setting, visualize how resources are distributed, and find possible bottlenecks in the process. (Rovani et al. 2015; Garcia et al. 2019)

Healthcare processes are known to be unique and nontrivial compared to other domains regarding characteristics, variability and complexity, privacy, multidisciplinary activities, treatments, and procedures. (Rebuge & Ferreira 2012; Garcia et al. 2019) The process is commonly nonlinear, nonautomated, and does not follow the planned structure like moving from one completed activity to the next, but typically interacts with patient condition and response to treatment and multidisciplinary medical experts' decision on best possible care in patients' situation (Partington et al. 2015).

Healthcare processes models are challenging to represent, and log mining often generates a so-called spaghetti type of process model (van der Aalst 2016; Erdogan & Tarhan 2018; Dallagassa et al. 2021) These unstructured process models can be obtained from very heterogeneous event logs of various activities (van der Aalst 2012). These models can be created, for example, with a heuristic miner, but they can be challenging to comprehend as they contain an overwhelming amount of behavior and therefore provide little value to the process. To simplify these models, one could reduce the complexity of the model by focusing on a group of similar behavior activities or most frequent activities. (van der Aalst 2012; Vidgof et al. 2020) An example of a spaghetti process model would be to model all patients visiting a hospital and simplifying it would be to select a group of patients with a similar medical problem (van der Aalst 2012).

In healthcare, process mining usually concentrates on a specific context and requires defining parameters and clinical knowledge from the field (Garcia et al. 2019; Dallagassa et al. 2021). Due to challenges in terms of population, treatments, and activities compatibility between different settings (Partington et al. 2015), process mining applications are also usually executed in one hospital or organization, making them not global or cross-organizational. (Partington et al. 2015; Garcia et al. 2019, Martin et al. 2020) Since analyzing spaghetti processes is challenging, in this thesis, instead of analyzing all data available, the focus will be on one patient group and their event logs regarding one illness.

Process model discovery typically aims to identify the most frequent activities and use those for developing medical protocols. (Dallagassa et al. 2021) Partington et al. (2015) study found that most of the research papers focus on process discovery techniques, whereas conformance and enhancement process mining techniques are used in very few studies, suggesting these two techniques are underutilized in healthcare.

Dallagassa et al. (2021) also found that the most popular area of application of process mining in healthcare was discovery process models. The second most popular with 24.8% of papers reviewed was resource assessment. The third popular area with 17.4% was focusing on conformity checks. They also discovered many predictive analyses and outliers' studies.

Healthcare data is highly sensitive, and the privacy of the data needs to be addressed when considering process mining with patient data. Aspects of privacy such as *anonymizing*, *event log attribute values,* and *atypical process behavior* should be considered when applying process mining algorithms. What makes data privacy more challenging is, that the algorithms rely on data accuracy and representativeness, like 1) *all events belong to a particular case*, 2) *variables that represent the case identifier, and the events are accurate* and 3) *timestamps are reliable and accurate.* (Pika et al. 2020) In this study, these various aspects of pseudonymizing the data were considered since the data is extracted from Electronic Health Records through Azure Data Lake.

Process mining can be executed on different data types, and it is possible to follow one indicator with sensor equipment. This can be utilized for following chronic conditions over some time by conducting an interactive process mining technology on sensor data. This can help with understanding, measuring, and managing patient chronic conditions and reveal patient patterns within a group of people suffering from the same chronic condition. (Valero-Ramon et al. 2020)

3.3. Summary of Literature

The literature showed many different aspects and benefits of applying process mining. The results of each research article used in the literature review for health care application are summarized in Table 1. The articles show that process mining is most commonly used for process discovery. Process mining has been applied in many domains of health care being

one of the most popular. In health care, the processes have been examined and discovered in various specialties.

There is a variety of tools and techniques for applying process mining to health care data. Since health care processes are easily spaghetti-like and hold a lot of information, the best tools are the type that can handle a lot of noise and can be used together with effective filtering and aggregation functions.

Table 1. Summary of literature

Year	Writers	PM Tools	Summary of Results
2012	Rebuge, Á. & Ferreira, D.R.	Medtrix Process Mining Studio (MPMS), ProM	The proposed methodology provides insights into the case workflow. The best techniques for health care process mining are the ones that can deal with a large amount of noise and enables the sorting of different behaviors making it possible for separate analysis
2015	Partington, A., Wynn, M., Suriadi, S., Ouyang, C. & Karnon, J.	Nitro software, ProM	Literature review: data pre-processing for process mining, the process discovery techniques were well covered, conformance analysis was not commonly explained, process enhancement explained in only 1 article Case study: similarities and differences were analyzed between different organizations
2015	Rovani, M., Maggi, F.M., de Leoni, M. & van der Aalst, W.M.P. 2015.	ProM, Declare	The declared model consists of 10 activities and 13 constraints, the process mining model does not show constraints related to the activities executed before the hospital admission, so the constraints are removed from the model. Some differences between the best practice model and the process mining model at the last steps of the model
2018	Erdogan, T.G. & Tarhan, A.	ProM, Disco	Bottlenecks and deviations that were crucial for determining measures were identified to improve the efficiency of the surgery process

2018	Baek, H., Cho, M., Kim, S., Hwang, H., Song, M. & Yoo, S.		Patients with diagnoses I63.8, I63.9, and I21.9 were associated with a longer stay at the hospital. Other variables also correlated with the length of stay.
2019	Garcia, C.D.S., Meincheim, A., Junior, E.R.F., Dallagassa, M.R., Sato, D.M.V., Carvalho, D.R., Santos, E.A.P. & Scalabrin, E.E.		Most popular research topics: Process discovery algorithms, then conformance checking and architecture and tools improvements. Most popular application domains: healthcare, information and communication technology, manufacturing, education, finance, and logistics
2019	Ibanez-Sanchez, G., Fernandez-Llatas, C., Martinez-Millana, A., Celda, A., Mandingorra, J., Aparici-Tortajada, L., Valero-Ramon, Z., Munoz-Gama, J., Sepúlveda, M., Rojas, E., Gálvez ,V., Capurro, D. & Traver, V.	PMApp	Process mining can identify the processes in health services, characterizing the specificity of an illness, evaluating and measuring how changes in the organization alter the process, patient behavior differences, comparing the actual process with the gold standards, showing the chain-value of the process to the patient
2019	Kurniati, A.P., Rojas, E., Hogg, D., Hall, G. & Johnson, O.A.	ProM	MIMIC-III database can be used for process mining for a hospital's standard administrative process, investigating variations in the treatment steps, and visualizing differences and commonalities between multiple process models. Date shifting with anonymization makes it impossible to analyze process mining.

2020	Helm, E., Lin, A.M., Baumgartner, D., Lin, A.C. & Küng, J.		Most popular tools: ProM, Disco, PALIA together with other tools. Often a mix of different tools. Most popular techniques: Fuzzy miner, ProM environment was often used with self-developed and case-specific approaches, Inductive visual miner in more recent studies, Trace Clustering technique. BPMN, ANOVA, and machine learning are not used frequently. Heuristic miner was popular previously. Most popular perspectives: Control flow, Conformance, Organizational, Performance Most popular encounter environments in SNOMED CT coding: Inpatient, Outpatient, Accidents and Emergency department (AED), General Practioner or GP practice site, Pharmacy Most popular clinical specialty: Medical specialty, surgical specialty, emergency medicine. Medical specialty consists of clinical oncology, community medicine, dentistry, dietetics and nutrition, emergency medicine, general practice, gynecological oncology, medical specialty, nursing, obstetrics and gynecology, and surgical specialty.
2020	Arias, M., Rojas, E., Aguirre, S., Cornejo, F., Munoz-Gama, J., Sepúlveda, M. & Capurro, D.	Celonis platform	Process mining in healthcare helps to understand how processes' variants and touchpoints, may affect the customers' experience, and to discover the process end-to-end
2020	Pika, A., Wynn, M.T., Budiono, S., ter Hofstede, A.H.M., van der Aalst, W.M.P. & Reijers, H.A.	ProM	Privacy protection for healthcare data while preserving data utility for process mining analyses is challenging. For process discovery, the generalization of timestamps did not affect the quality of discovered models but did affect the performance analysis

2020	Rabbi, F., Lamo, Y. & MacCaull, W.		The developed model-based slicing technique includes filtering and grouping of activities and provides a higher level of abstraction of the data which can be used for process mining
2020	Valero-Ramon, Z., Fernandez-Llatas, C., Valdivieso, B. & Traver, V.	PMApp	Chronic conditions such as obesity, hypertension, and hyperglycemia can be analyzed with process mining and identify sub-populations in them
2021	Dallagassa, M.R., Garcia, C.D.S, Scalabrin, E.E., Ioshii, S.O. & Carvalho, D.R.		The discovery process model type was most frequent, followed by resource analysis and evaluation. The most popular algorithms were the Fuzzy Miner and Heuristic Miner. Other: Inductive Visual Miner in ProM version 6 onwards

In health care, process mining can give valuable information about the whole process and give new insights that can help with improving the process. The information gained from comparing the discovered patient treatment processes to the best practice or the treatment plan can be of great value to the organization.

4. Data and Methodology

The data and methodology chapter concentrates on the methodology of this study. The background for the decisions for choosing the analyzing methods and the process of data collection and preprocessing are explained in detail.

4.1. Healthcare Data from HUS

Helsinki University Hospital, HUS, provides specialist medical care services. HUS is the largest healthcare provider and the second-largest employer in Finland. (HUS 2022) HUS Comprehensive Cancer Center is a part of HUS, concentrating on the treatment of various types of cancer. Breast cancer patients are a large patient group with approximately 1500 breast cancer surgeries per year. Breast cancer surgery is the main type of treatment for breast cancer, followed by adjuvant therapy. (HUS Syöpäkeskus 2018)

Information security needs to be addressed as the healthcare data is highly sensible (Pika et al. 2020) and this should be considered throughout the research process. This research project required a research permit from HUS Comprehensive Cancer Center because it handles and analyses its patients' data. Register specification and assessment of the effects were required as well as a non-disclosure and information security agreement. The study also required a contact person from Comprehensive Cancer Center. Ethical approval was not required as this is conducted as a register study. (HUS Tutkimus 2021) The research permission was granted in January 2022.

4.2. Data Extraction

The second part of this study focuses on the process of mining the electronic health records of breast cancer patients. Information systems record the event log data in unstructured form, which means that the data is distributed between many different systems and database tables.

The event data needs some work to extract. Data extraction is considered a part of the process mining effort. (van der Aalst 2016, 32)

The data was extracted from HUS Azure Data Lake in January and February 2022, accessing the databases with Python in Databricks-tool. The patient data is pseudonymized on the Data Lake and all secondary patient identifications have been removed. Research-specific re-pseudonymization was performed before performing the analysis. The re-pseudonymized patient id was used only to link the activities of the treatment to each patient. As the researcher is an employee of the team Analytics, Data Lake, and Data Service at HUS, she had access to both the Databricks and Data Lake tools directly. The supervisor of the team permitted the use of these resources for this study. All data handling from extracting the data to the analysis was performed in a secure Databricks-notebook, avoiding any data transfers and only the final visualizations of the models were transferred outside the secure environment for the thesis report and appendices.

The source system for the data was the new patient information system Apotti which was deployed at HUS Comprehensive Cancer Center in October 2020 (HUS 2020b). Data was extracted from Apotti referrals, procedures, and outpatient and inpatient systems. The imaging procedures were extracted from an old imaging information system, as Apotti was not deployed for this module in early 2021.

The cohort was first identified with national, THL-specified procedure codes for breast-conserving surgery of all procedure of patients. The referrals to the surgery were identified with unit codes of the surgery ward. The outpatient visits were identified with unit codes and ICD-10 diagnosis code starting with C50 or the diagnose code being empty, in case the patients were simultaneously treated for other cancer types as well. From the procedures data, it was noticed that some patients were admitted to the breast cancer surgery ward as inpatients for surgery, and for those patients, the inpatient visit data were extracted using a key linking the procedures to the inpatient period. Other inpatient periods were not included in the study. Overall, the data consisted of a cohort of 855 patients that had activities in 19 different resources or unit codes from HUS.

For each patient, the dataset was filtered to only include events after the referral to breast cancer surgery. To exclude patients that still had their cancer treatment ongoing at the beginning of 2022, the data was cleaned to only contain the patients that had at least one visit

to Radiation Therapy Unit before December 2021. The visits that had a visit type of “phone call” or “letter”, as well as diagnostic imaging and physical therapy sessions, were excluded from the data as they were not present in the hand-drawn process and their event position in the patient pathway was considered non-critical.

From each data source, the timestamp for the start of the event, event type, and resource were extracted with primary and foreign keys that helped link the data sets together. Outpatient and inpatient visits were joined with the procedure data. The timestamp for the end of each event was mostly missing from the data and therefore excluded from this study.

4.3. Data Preprocessing

After data extraction, the dataframe needed some manipulation for it to work for process mining. First, the data sets from referrals, and visits together with procedures, were manipulated to the same format so that they could be concatenated to one PySpark dataframe where each row represents an event in patients’ clinical pathway.

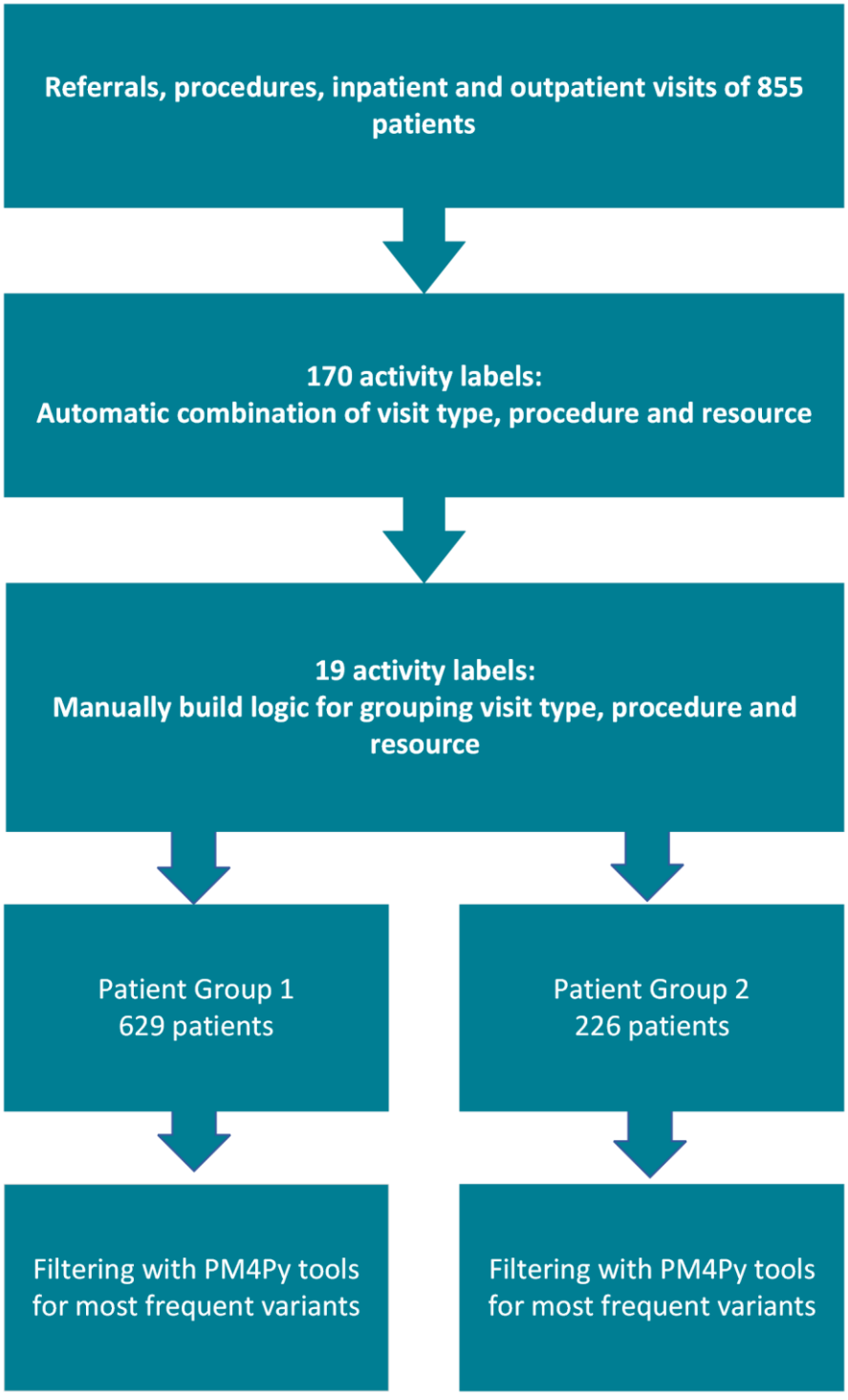

Figure 6. Data preprocessing

To create a discovery model on the data, additional variables were created based on the data. A label for each event in the patients' clinical pathway was necessary. As a first step, refer-

rals for surgery were simply labeled “referral”, and visits were first labeled with a combination of the visit type, procedure code, and the resource or unit code (Figure 6). This resulted in 170 different labels for the events. These automatically created labels soon proved to be too complicated for process discovery, since many of them represented the same kind of activity in two different resources or two different procedure codes. Exploratory analysis was conducted on this eventlog to get an impression of the data. With automatically created activity labels, where visit types “phone call” and “letter” were not excluded and the data from the whole cohort of 855 patients were included. In this test each case had a unique clinical pathway, meaning that every patient had a different order of activities in their care.

Based on the results of the initial testing, the activities were further investigated to see which ones represent the same type of activity inpatient clinical pathway. Activities were re-labeled by grouping them with manually built logical filters to match 19 unique event labels listed in Table 2.

Table 2. Activity Labels

Activity	**Abbreviation in analysis**
Ablation Surgery	Abl Surgery
Breast-Conserving Surgery	BC Surgery
Chemotherapy	Chemo
Dose Planning CT	DP CT
First Appointment at Chemotherapy Unit	1V Chemo
First Appointment at Radiation Therapy Unit	1V RT
First Appointment at Surgery Unit	1V Surgery
Last appointment at Radiation Therapy Unit	Last V RT
Palliative Visit	PVisit
Physician at Chemotherapy Unit	PChemo
Physician at Radiation Therapy Unit	PRT
Physician at Surgery Unit	PSurgery
Radiation Therapy	RT
Referral for Surgery	Referral
Repeat Surgery	RSurgery
Surgical Procedure	SProc
Visit at Chemotherapy Unit	VChemo

Visit at Radiation Therapy Unit	VRT
Visit at Surgery Unit	VSurgery

The breast cancer patients that are not treated with radiation therapy are out of scope for this study, therefore based on the options for adjuvant therapy after breast-conserving surgery (HUS 2021), discovery models were expected to create two distinct clinical pathways:

1) patients treated with surgery and radiation therapy

2) patients treated with surgery, chemotherapy, and radiation therapy.

For the process discovery purposes, the dataset was split into two subsets based on the patient's adjuvant therapy. Patient group 1 includes the data from 629 patients that were treated with breast-conserving surgery and radiation therapy. Patient group 2 was identified with at least one visit to the chemotherapy unit and includes data from 226 patients whose breast cancer treatment consisted of surgery, chemotherapy or other drug therapy, and radiation therapy.

As a final step, only a few PySpark dataframe columns were selected and renamed to fit the format that can be analyzed with the Process Mining for Python framework. These columns were:

1) *case:concept:name* , a pseudonymized patient identifier
2) *"time: timestamp"*, the starting time of the event or activity
3) *"concept: name"*, the label of the activity
4) *"org: resource"*, the label of the resource or unit code

The dataframes were ordered by the case and time. The Process mining for Python framework and further steps for data filtering will be introduced in the data analysis chapter.

4.4. Data Analysis

The analysis was conducted with Process Mining for Python (PM4P) library in Databricks-notebook using Spark Clusters. The PM4Py framework was selected because it has been

developed to provide extendibility, customized algorithms, and large-scale experiments which are not available through the software tools for process mining. The PM4Py enables algorithm development and customization in process mining analysis, wide-ranging documentation, stable algorithm through thorough testing as well as easy integration of other algorithms or Python libraries in data science. It also works in a collaborative ecosystem that is based on sharing code and results in the process mining community. (Berti et al. 2019) The use of PM4Py in Databricks notebook also enables easy automation and integration to existing information systems.

Process mining for the Python framework consists of different packages of algorithms that allow analyzing a given process through process discovery, conformance checking, and enhancement (Berti et al. 2019). For process discovery, it uses various algorithms, such as the Alpha(+)-algorithm developed by van der Aalst (van der Aalst et al. 2004) and the Inductive Miner (Leemans, Fahland & van der Aalst 2015). The Alpha-algorithm takes event logs as input and creates a Petri net which describes the behavior of the event log data (van der Aalst 2016, 33). The inductive miner was introduced in detail in Chapter 2.2.

Filtering of the event log is possible based on the timeframe, case performance, endpoints of the trace, trace variants, attributes, or paths. Python visualization libraries such as GraphViz and NetworkX can be used with PM4Py. (Berti et al. 2019)

Discovery models can be conducted with more than one algorithm that PM4Py offers. For this study, the process discovery was the first part of the analysis, and discovery model was created with inductive miner first since it is widely used and tested. (Helm et al. 2020; PM4PY 2021) The alpha-miner was also tested on the data, and the directly-follows graphs were generated.

As a first step of the analysis, the eventlog containing the activities for patient group 1 was tested with PM4Py statistics. The variants or orders of the activities were sorted to display the ten most frequent ones. The most frequent order of the activities had 104 observations and the second most frequent 31 observations. However, even this event log revealed the additional need for filtering the data since testing process discovery with the inductive miner with all variants, the process discovery resulted in a spaghetti-like process, as the literature suggested might happen (van Der Aalst 2016, 399–400). This study aimed to identify the typical clinical pathway for patients going through breast cancer treatment. Therefore, it was

necessary to exclude the variants or orders of the activities from the data that had a low frequency, meaning they were unique or almost unique to each patient and could not be considered as a typical treatment path.

The data was then filtered with the start activity of "Referral for Surgery". Since including all variants would not represent the typical pathway for patients, only the four most frequent variants were selected and the frequencies of variants after them decreased dramatically. These variants represented 163 cases or patients out of the 629 in patient group 1. The process models were created with an inductive miner and directly-follows graphs. These models represented the typical clinical pathway for patients that were treated with breast-conserving surgery and radiation therapy and were not treated with chemotherapy or other drug therapy.

For finding a typical process model for patient group 2, the patients that were also treated with chemotherapy or other drug therapy, the most frequent variants were challenging to find since most of the patients had a unique clinical pathway. For process discovery, PM4Py auto filtering of the less frequent activities was applied and only the activities after the activity "Referral for Surgery" were considered. The variants were sorted through frequencies of them and only the top seven most frequent variants were selected. The rest of the variants were present in the data only once and hence excluded since they could not be considered typical orders of the activities. The eventlog was created and consisted of 19 cases or patients. However, the frequencies of the variants included in the analysis were still relatively small, ranging between 2 and 4, which means that the typical pathway only represented a minority of patients. This could affect the reliability of the study negatively. The models were created using the inductive miner algorithm and directly-follows graph for these seven variants.

The discovered inductive miner models were visualized as process trees and Petri nets and the directly-follows graphs with frequencies and activity performance. The inductive miner models were evaluated through measurement of fitness, precision, generalization, and simplicity.

For conformance checking, PM4Py enables token-based replay and alignments. Token-based replay was performed on the inductive miner Petri net to test the executed transitions and to identify possible remaining or missing tokens. (Adriansyah et al. 2011)

PM4Py framework offers tools and statistical tests for analyzing the process model further. The temporal profile calculates the average time and standard deviation between the start of two activities in seconds. This gives a reference point for identifying the patients who spend extended time on their pathway. (PM4PY 2021) The temporal profile was performed on models for both patient groups 1 and 2, and the output was calculated in days since patients had typically only one activity per day.

The comparison between the hand-drawn process model and the discovered model was conducted visually for the order of the activities. The time delays compared to the planned process and the actual process was conducted with the temporal profile and directly-follows graphs. The results of these analyses will be presented next.

5. Results

This chapter discusses the results of the analysis for process discovery and conformance checking. The patients were divided into two groups as described previously, based on their treatment path for breast cancer. Patient group 1 was used in process discovery for finding the typical clinical pathway for patients treated with breast-conserving surgery and radiation therapy. Patient group 2 was used for finding the typical breast cancer pathway for patients treated with breast-conserving surgery, radiation therapy, and chemotherapy, or another type of cancer medication.

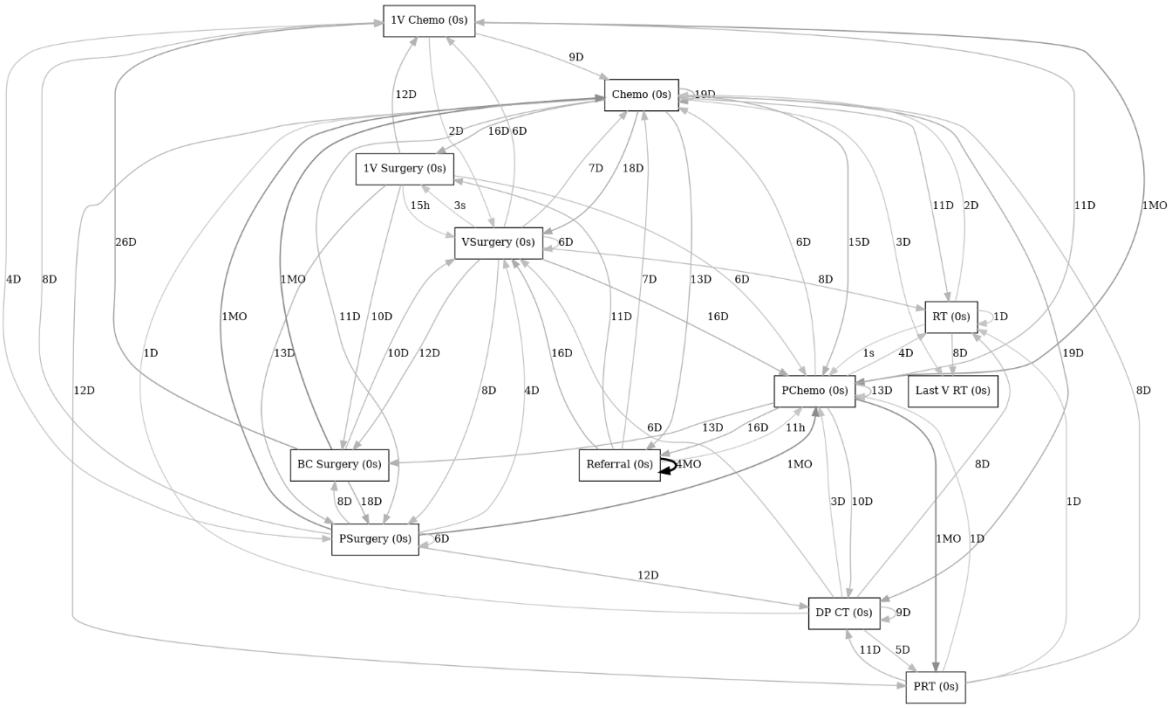

Figure 7. The discovered spaghetti-type process model for patient group 2 without filtering the infrequent variants

Identifying the typical process models for patient groups 1 and 2 with process discovery, without filtering the infrequent variants, resulted in a spaghetti-type of processes (Figure 7). The analysis of the processes containing this much heterogeneous behavior was not possible

and therefore only the most frequent variants or orders of the activities were selected for both patient groups.

5.1. The Number of Activities and the Length of Treatment for the Cohort

All patients had on average 27.7 (median 24) activities in their breast cancer treatment pathway, ranging between 6 and 84. These activities only included the ones that were labeled in this study. The average length of treatment period from referral to surgery until the end of their cancer treatment was 167.8 days (median 108). The length of the treatment period ranged between 36 and 649 days.

The distinctions in the number of activities and the length of treatment between the different patient groups are presented in Table 3. As expected, the patients in the typical pathway for patient group 2 had on average a higher number of activities, 32.0 compared to the patients in the typical pathway for patient group 1 with 22.5 activities. Similarly, the length of treatment was on average 222.5 days for patients in the typical pathway in patient group 2 compared to patients in the typical pathway in patient group 1 with an average of 97.3 days.

Table 3. The number of activities and length of treatment for patient groups

	Typical Pathway for Patient Group 1	**Typical Pathway for Patient Group 2**	**All Patients in Group 1**	**All Patients in Group 2**	**All patients**
Screening / Description	The most frequent variants for patients treated with surgery and radiation therapy	The most frequent variants for patients treated with surgery, chemotherapy, and radiation therapy	All patients treated with surgery and radiation therapy	All patients treated with surgery, chemotherapy, and radiation therapy	Combination of Patient Groups 1 and 2
Number of observations	163	19	629	226	855
Percentage of the cohort	5.24	2.22	73.57	26.43	100
Average (median) number of activities	22.5 (23)	32.0 (32)	23.6 (24)	39.0 (38)	27.7 (24)
Min number of activities	21	30	6	16	6
Max number of activities	23	35	54	84	84
Average (median) length of treatment in days	97.3 (95)	222.5 (223)	125.3 (99)	285.9 (249)	167.8 (108)
Min length of treatment in days	73	204	36	162	36
Max length of treatment in days	144	252	519	649	649

The results of the process discovery and the conformance checking are covered in the following chapters. The process discovery is represented separately for the two patient groups.

5.2. Process Discovery for Breast Cancer Patients

5.2.1. Patients Treated with Surgery and Radiation Therapy (group 1)

The process discovery analysis was conducted with the inductive miner for the four most frequent variants within the patient group 1. The discovered models represent the typical clinical pathway for patient group 1 with 163 cases, which is 25,9% patients out of the total number of 629 patients treated with breast-conserving surgery and radiation therapy.

The Directly-Follows Graphs visualize the time delays and frequencies between the two consecutive activities. In Figure 8 the time delays are displayed in rounded in hours, days, or months. This graph could also visualize the length of each activity, but since the end time for the activities was mostly not available, they are not visualized.

Figure 8 shows that the activities at the beginning and the end of a patient's clinical pathway are similar for all patients in the analysis for patient group 1, starting with the *Referral for Surgery* and *First Appointment at Surgery Unit* and ending with multiple visits with *Radiation Therapy* and *Last appointment at Radiation Therapy Unit.* Interestingly, some differences can be discovered in the patients' treatment as the activities *Visit at Surgery Unit* and *Dose Planning CT* do not appear to be on all patients' care paths.

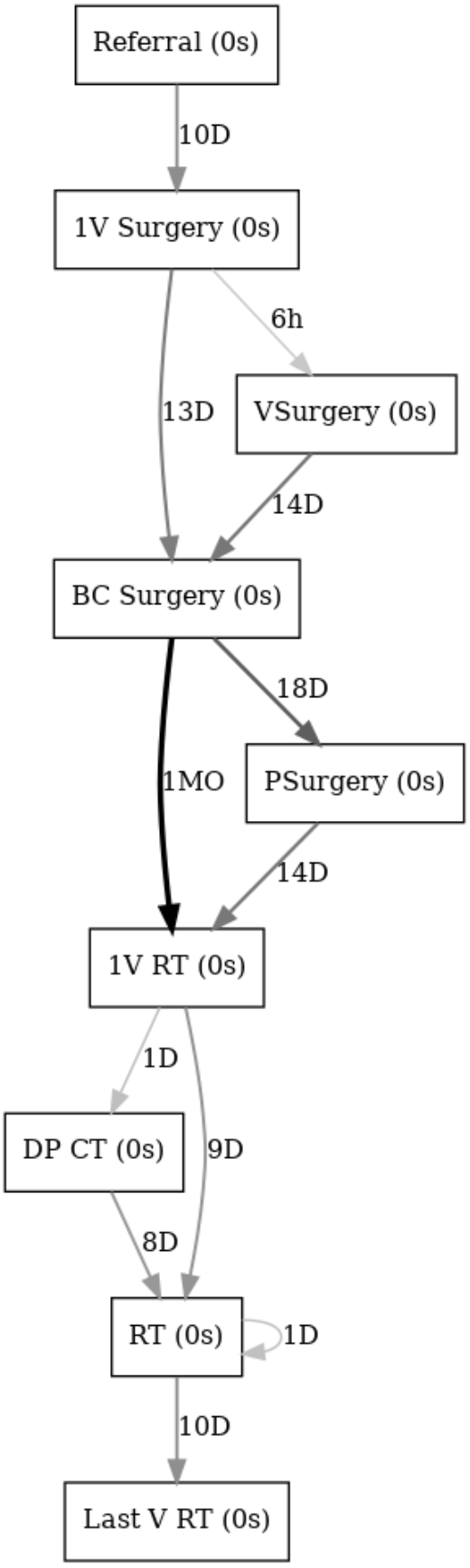

Figure 8. Directly-Follows Graph with delays for Patient Group 1

The calculated average delays in Figure 8 show that the patients' treatment start with the *First Appointment at Surgery Unit* on average 10 days after the *Referral for Surgery* is received at the HUS Comprehensive Cancer Center. The figure reveals that there seem to be small differences in the time delays depending on whether the patient had a *Visit at Surgery Unit* before the surgery or a *Physician at Surgery Unit* before the start of radiation therapy. These time-delay differences were more closely examined and presented in Figure 9.

Figure 9 visualizes the frequencies of the events and the average delays between different activities in seconds. The displayed frequencies for each activity help to understand how many patients follow each path. The activity of *Visit at Surgery Unit* is present for 150 out of 163 patients. After the *Breast-Conserving Surgery*, activity *Physician at Surgery Unit* is occurring for 148 patients whereas the rest of the patients do not have this activity in their pathway.

The accurate time delays in seconds reveal a small difference in the time delays between the different care paths of the patients. There are only 13 patients that do not have a *Visit at Surgery Unit* and for those patients, the delay between *First Appointment at Surgery Unit* and *Breast-Conserving Surgery* seems to be on average 1.3 days less. On average, there is a 2.3 bigger delay in days if the patient has a *Physician at Surgery Unit* activity between the activities *Breast-Conserving Surgery* and *First Appointment at Radiation Therapy Unit*. However, the frequencies in these activities are quite small and the difference may not be considered significant.

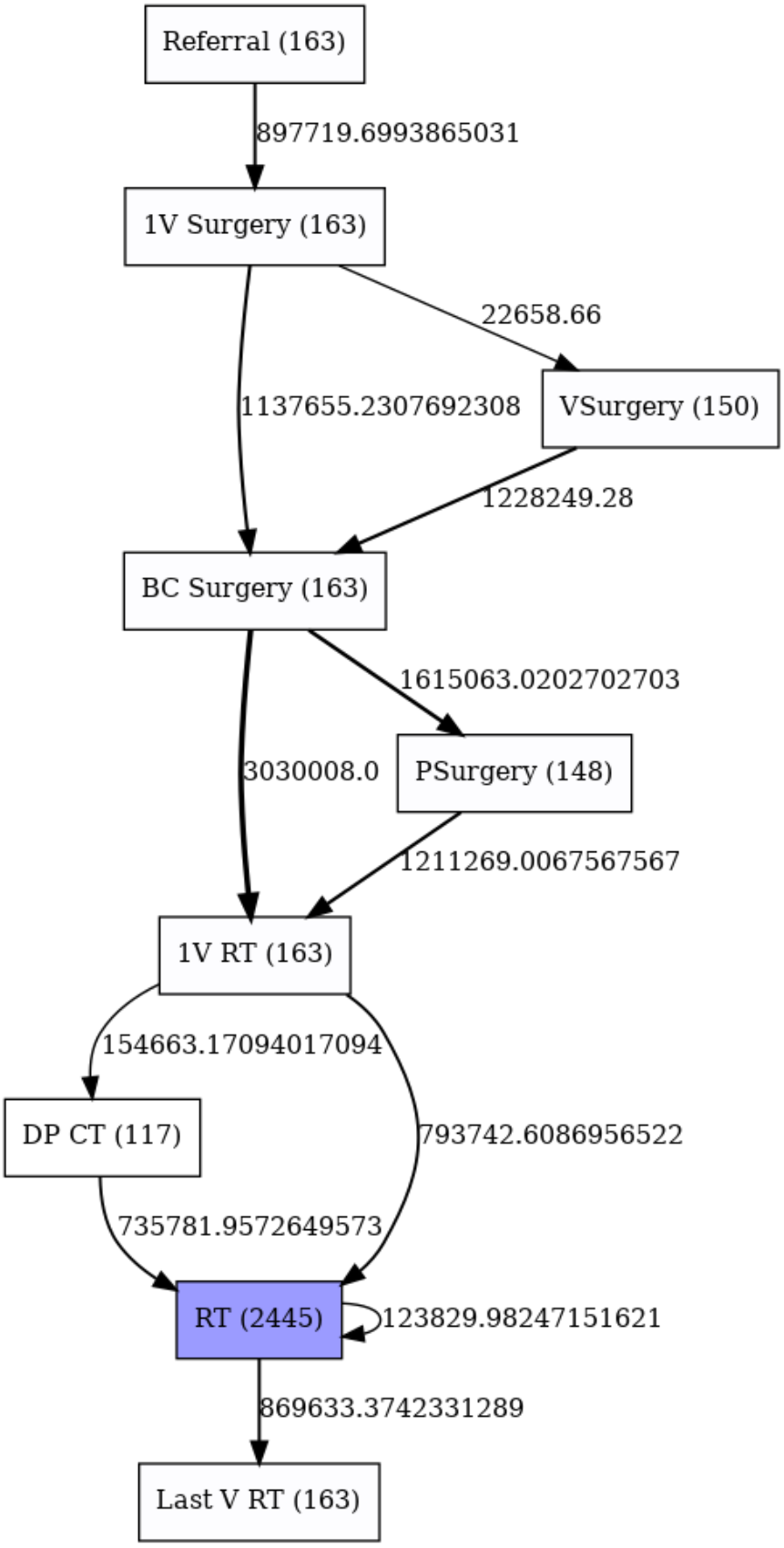

Figure 9. Directly-Follows Graph for Patient Group 1 displaying the frequency of each activity and time delays in seconds.

After the *First Appointment at Radiation Therapy Unit*, based on the directly-follows graph, some of the patients are moved directly to *Radiation Therapy*, whereas 117 patients go through a *Dose Planning CT* first. Again, there is a small difference in the time delays, and the analysis would suggest that for patients that do not go through a *Dose Planning CT*, their radiation therapy would start about 1.1 days earlier. Since the radiation therapy cannot technically be performed without the dose planning CT (Suomen rintasyöpäyhdistys 2021) the CT scan is perhaps performed on these patients on the first visit to the radiation therapy unit and hence recorded in a different way, meaning the activity labeling may not have worked perfectly on this part or the data is missing of these records. The frequency of 2445 for Radiation Therapy means that the patients were treated on average 15 times. It should be noted that the rounded time delays between different activities in Figure 8 do not indicate the delay in patients' care path that was present in Figure 9 for patients that did not have a *Dose Planning CT* between their *First Appointment at Radiation Therapy Unit* and *Radiation Therapy*.

The inductive miner algorithm creates a process tree of the discovered model. The process tree (Figure 10) visualizes the process into 9 different sequential branches of which each holds only one activity. The xor and xor loop branches visualize the different types of transitions in different patients' treatment paths.

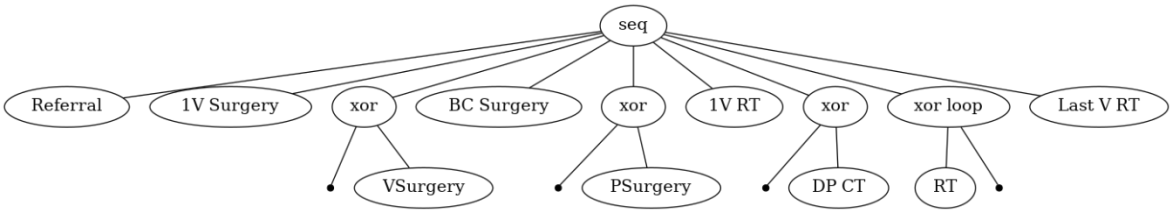

Figure 10. The Process Tree of the Patient Group 1 Inductive Miner displaying the transition types for each activity

The process tree model was turned into a Petri net (Figure 11), and it displays the model in a chain of events. The different trails reveal where the patients had deviations in their clinical pathway. The Petri net also visualizes the hidden transitions between the different activities as black boxes. These transitions can be used to evaluate the model performance.

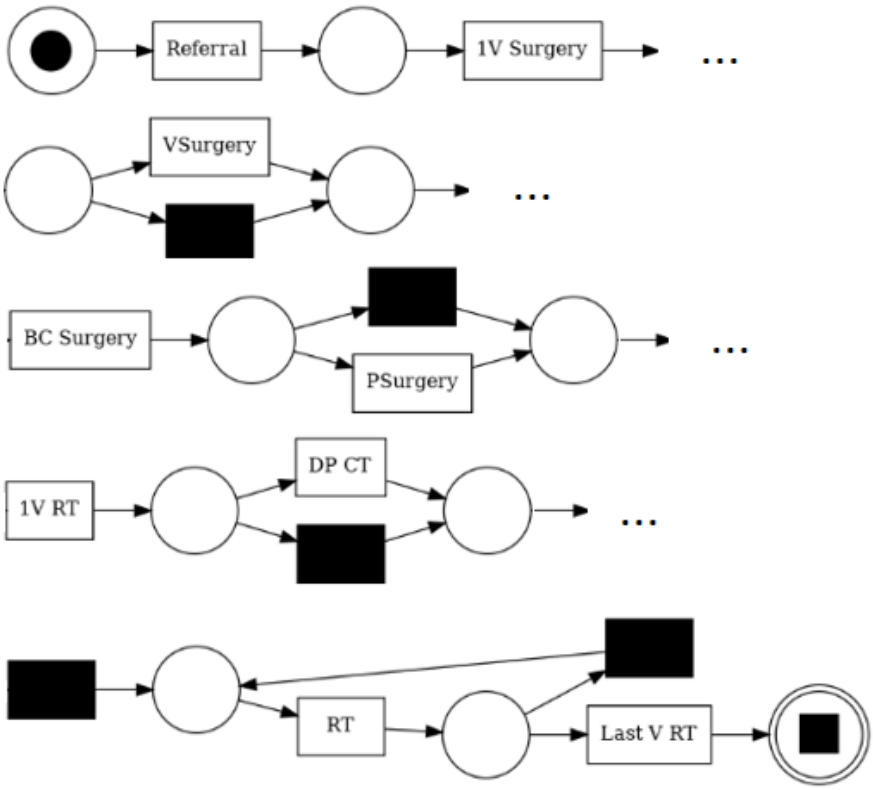

Figure 11. Petri net of the Patient Group 1 Inductive Miner displaying the process in a chain of events with the hidden transitions

Token-based and alignment-based replay revealed that all transitions executed as expected and no missing or remaining tokens were identified with a fitness of 100%. The precision of the model was 0.624 and the generalization was 0.896. The simplicity of the model was evaluated to be 0.806.

5.2.2. Patients Treated with Surgery, Chemotherapy, and Radiation Therapy (group 2)

The process discovery for the typical clinical pathway within the patient group 2 represents 19 cases, which is 8.4% patients out of the total number of 226 patients treated with breast-conserving surgery, chemotherapy, or another type of cancer medication and radiation therapy. The eventlog consists of 7 distinct variants or orders of the activities with frequencies

between 2 and 4. The results of the process discovery cover the processes for these 19 patients. The rest of the variants in patient group 2 were excluded since they were all unique to one patient. The process discovery was conducted with the inductive miner and the directly-follows graphs. The directly-follows graphs gave a clear visualization of the typical process for the 19 cases with information about the time delays (Figure 12) and the frequencies (Figure 13) between different activities.

The directly-follows graph (Figure 12) gives a clear insight into the typical process and its deviations for patients in patient group 2. The beginning of the treatment and its deviations are very similar to the typical pathway of patient group 1 with some patients not going through the activities *Visit at Surgery Unit* and *Physician at Surgery Unit.*

In the next steps in the patient's pathway, there seems to be more variation. The adjuvant therapy starts with the *First Appointment at Chemotherapy Unit* and the first *Chemotherapy* session, but the physician meetings seem to take place either at the chemotherapy unit or at the radiation therapy unit. Nevertheless, the patient treatment seems to be quite identical to all patients in the typical pathway for patient group 2 after the *Dose Planning CT*.

In Figure 12 the time delays between different activities for patient group 2 are presented in rounded hours, days, or months. Interestingly, the deviations in the clinical pathway also seem to affect the delays in the activities. The delays between the different types of visits at the chemotherapy unit seem too hard to interpret with a directly-followed graph since the transitions between the activities are not presented in the graph. However, one can see that the *Chemotherapy* is repeated on average every 20 days, and the *Dose Planning CT* was usually performed 18 days after the last *Chemotherapy.*

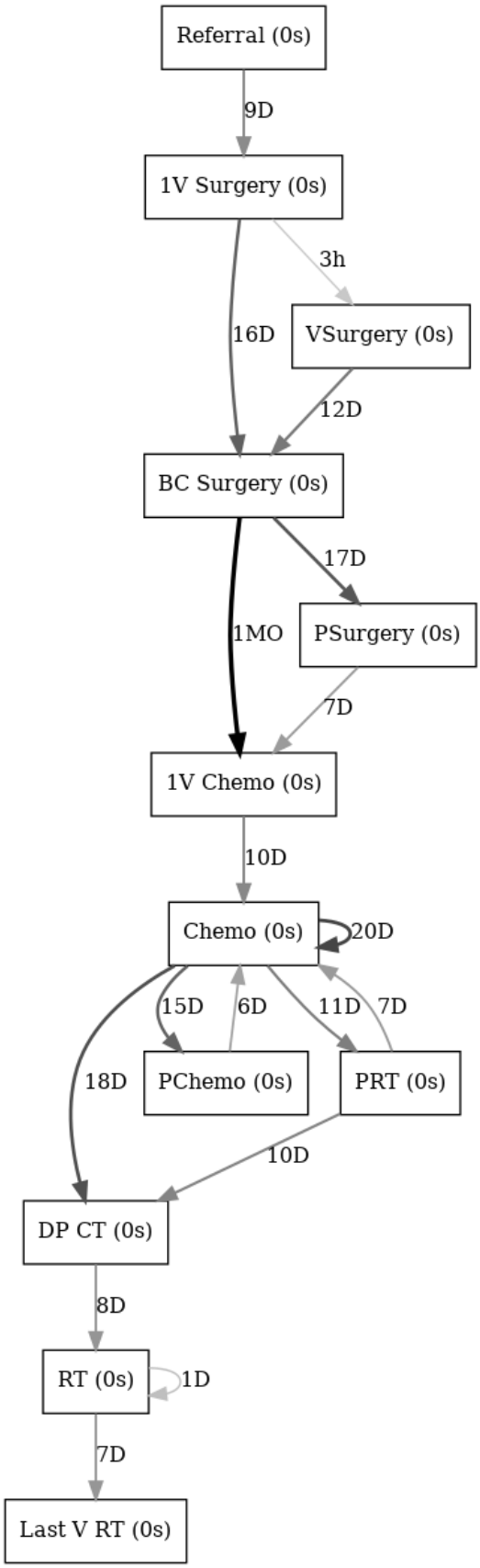

Figure 12. Directly-Follows Graph with delays for Patient Group 2.

Figure 13 visualizes the distinctions in the frequencies of different activities. The frequencies help to see how many patients actually go through the different paths in the model. 16 out of the 19 patients go through the activities *Visit at Surgery Unit* and *Physician at Surgery Unit* as their third and fifth activity in their pathway. Interestingly, those three patients that do not have these activities have a bigger time delay of 3.5 days between activities *First Appointment at Surgery Unit* and *Breast-Conserving Surgery*, and 6.3 days between activities *Breast-Conserving Surgery* and *First Appointment at Chemotherapy Unit*.

Patients had on average two *Physician at Chemotherapy Unit* activities which may explain the variation in the order of chemotherapy activities. Nine patients had a *Physician at Radiation Therapy Unit* before the *Dose Planning CT* which was performed for all 19 patients. Activity *Radiation Therapy* was repeated on average 15.8 times for patients.

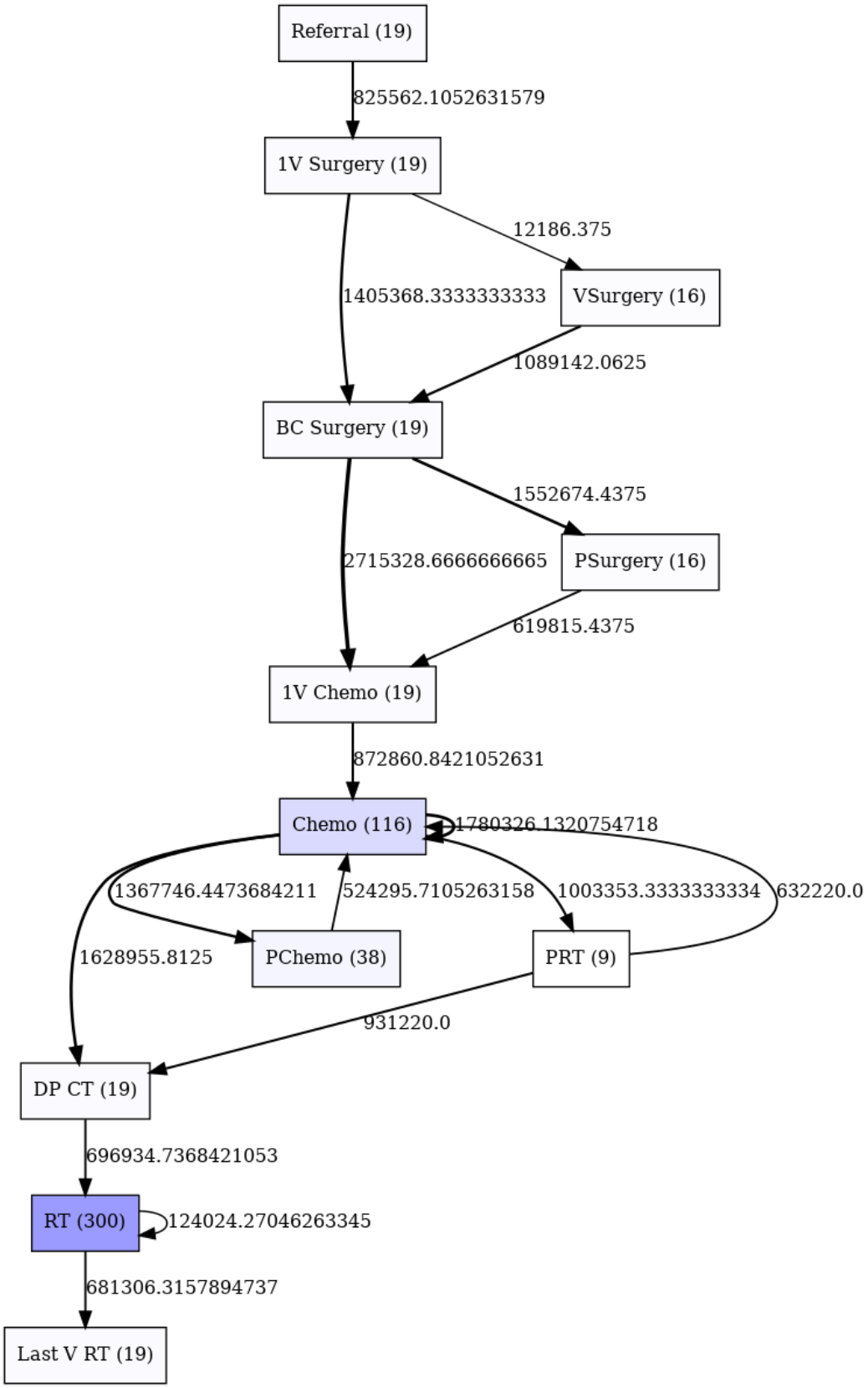

Figure 13. Directly-Follows Graph for Patient Group 2 displaying the frequency of each activity and time delays in seconds.

The inductive miner algorithm created a process tree (Figure 14) which visualizes the transition types between the different activities. The sequential events take place from left to right, but in the "xor" branches only one or the other occurs. The graph shows how even the typical clinical pathway with only 19 patients has a lot of deviation in the order of the activities between the *First Appointment at Chemotherapy Unit* and *Dose Planning CT* with several sub-branches. The transition types between the activities are valuable in patient group 2 since the directly-follows graphs were challenging to interpret the chemotherapy part of the patient's clinical pathway. In Figure 14 it is clear that all the patients have both an appointment with the physician and the chemotherapy.

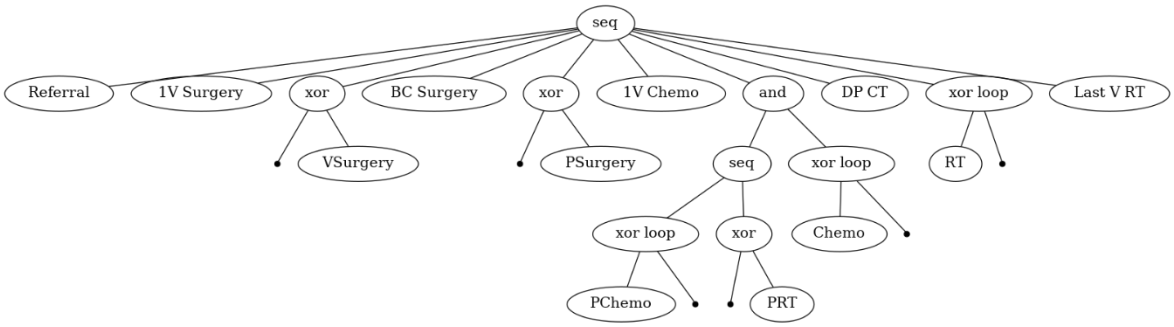

Figure 14. The Process Tree of the Patient Group 2 Inductive Miner displaying the transition types for each activity

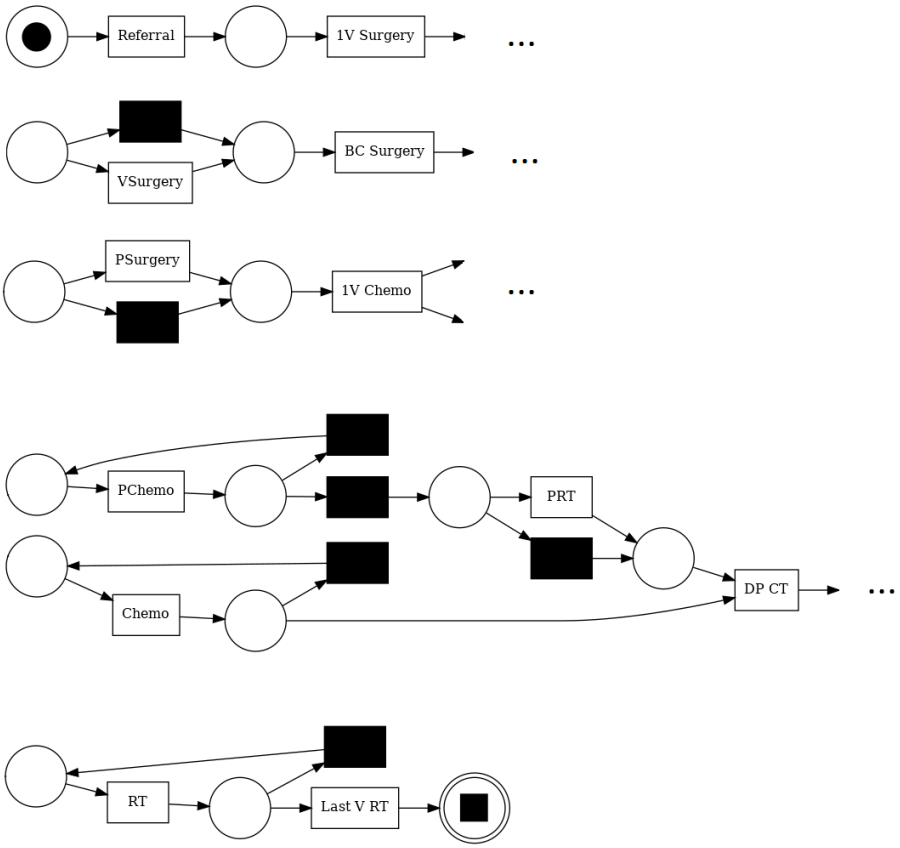

Figure 15. Petri net of the Patient Group 2 Inductive Miner displaying the process in a chain of events with the hidden transitions

In Figure 15 the process tree has been transformed into a Petri net. Visually it can be observed that the start and the end of the patients' pathways were similar in both patient groups 1 and 2. The order of Chemotherapy and physician appointments varied within the group of 19 patients significantly. The inductive miner model's performance can be evaluated in a Petri net format with token-based and alignment-based replay. The performance check revealed that all transitions were executed as expected and no missing or remaining tokens were identified with fitness of 100%. The precision of the model was 0.529 and the generalization was 0.757. The simplicity of the model was evaluated to be 0.739.

5.3. Conformance Checking for Breast Cancer Patients

For conformance checking, the planned process model of breast cancer patients' clinical pathway was compared with the discovered model. The planned process was described in chapter 2.3. The process discovery resulted in a process model that could easily be compared to the planned, hand-drawn process model and identified most of the same activities in patients' clinical pathways. Overall, the discovered process was similar to the hand-drawn model and the activities were mostly performed in the same order as planned. Due to a limited number of visit types in the patient information system, the data labeling was not able to identify all distinct activities in the patient's pathway, and thus some of the activities were named differently than in the hand-drawn process. For easier comparison, the hand-drawn model was transformed into a directly-follows graph with similar activity labels as used for process discovery (Figure 16 and Figure 17).

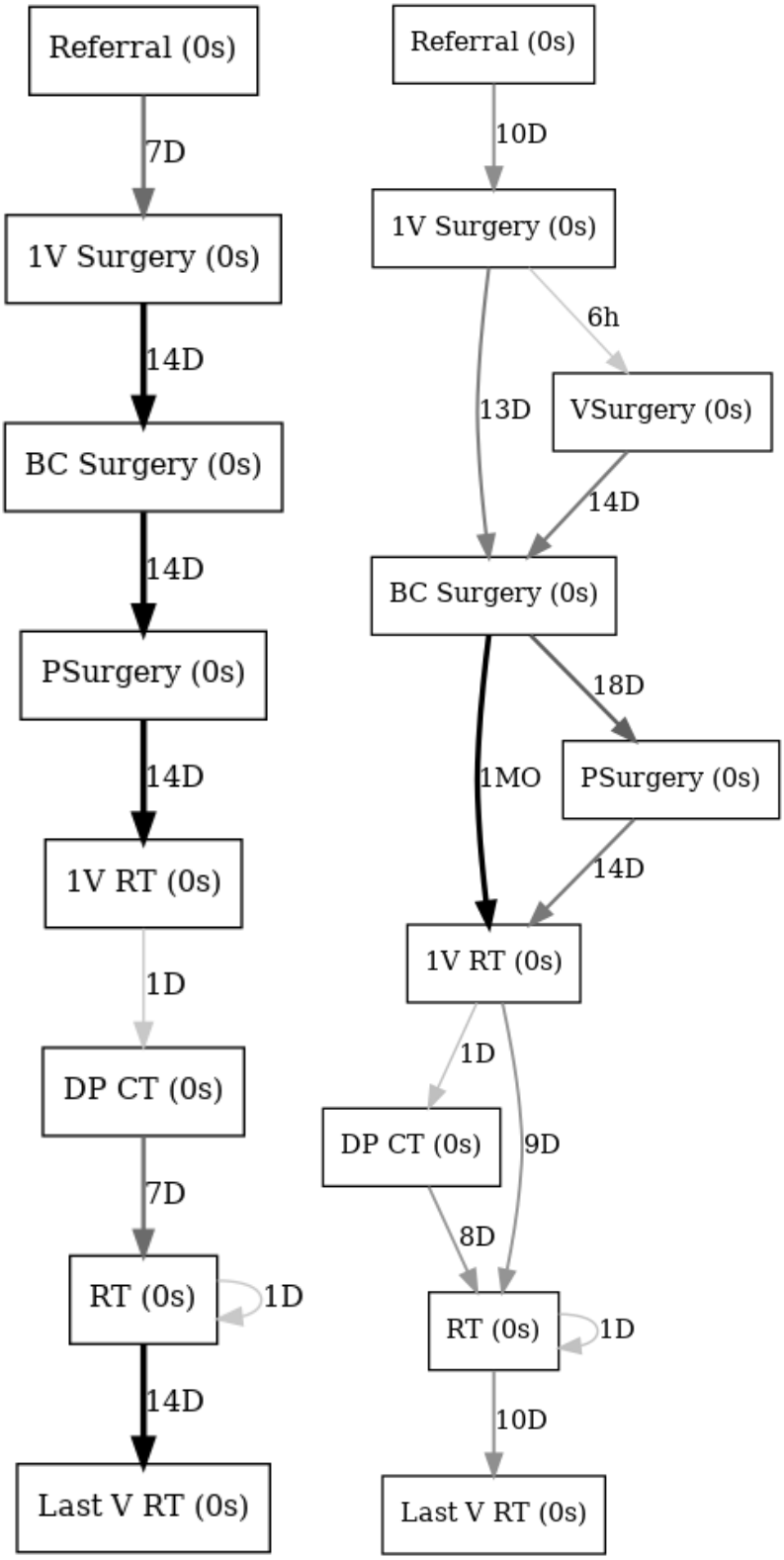

Figure 16. The planned process model (left) versus the discovered model (right) for the patient group 1.

The breast cancer surgery process was very similar to discovered model and the hand-drawn model. Most patients in patient groups 1 and 2 had an additional appointment in the breast cancer surgery unit on the same day as the first appointment. According to the data, all patients did not have a follow-up appointment after surgery.

For patient group 2, the typical pathway represented only 19 patients and there was more variation present in the patient processes. The follow-up appointment at the chemotherapy unit was identified at various points in the patient's chemotherapy process and not just after the last chemotherapy session. On average, the patients had two follow-up appointments. For some patients in patient group 2, the first appointment at the radiation therapy unit was missing and the treatment was continued directly with dose planning CT after chemotherapy.

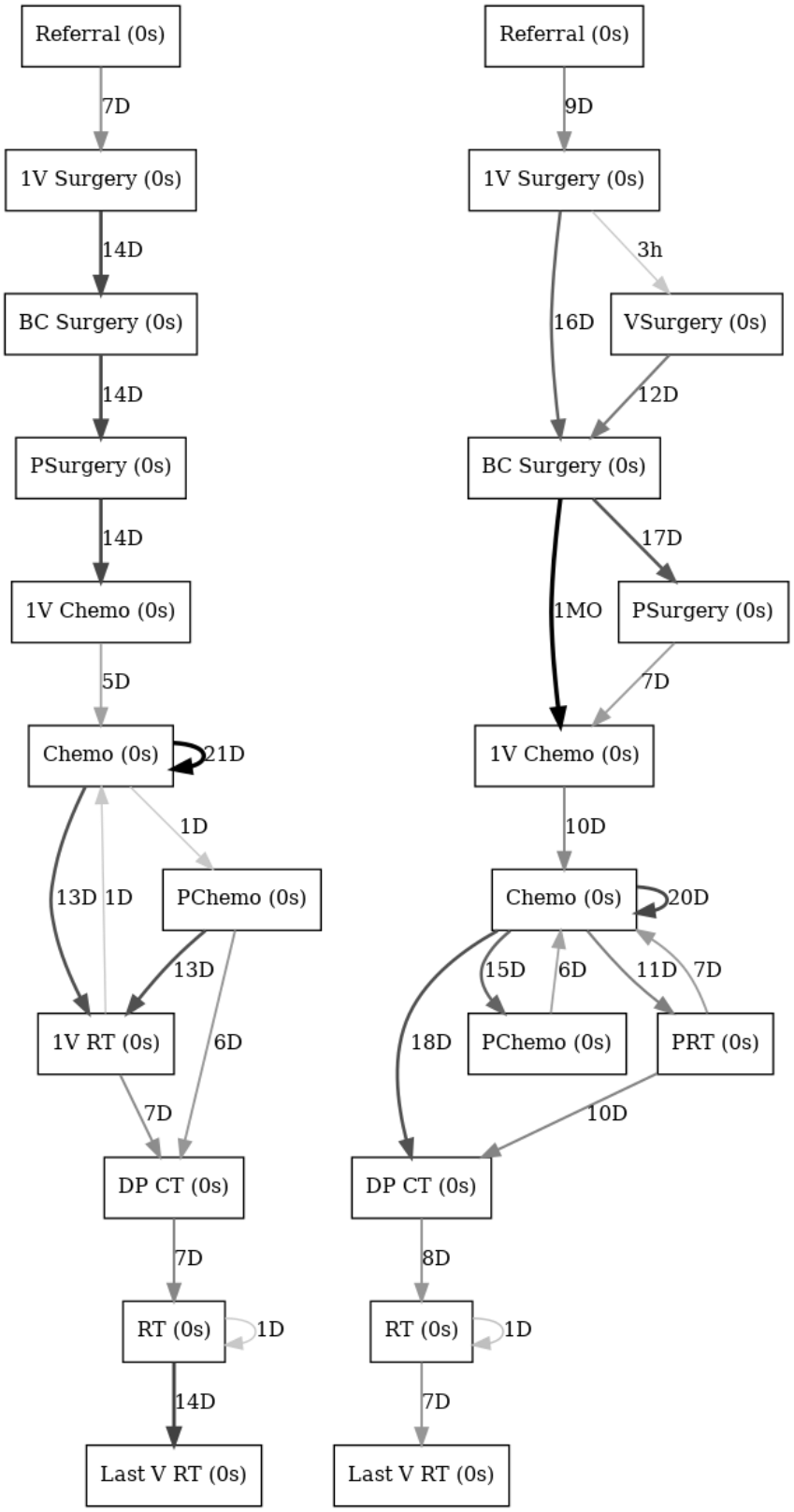

Figure 17. The planned process model (left) versus the discovered model (right) for patient group 2.

The radiation therapy part of the patient's cancer treatment followed the planned pathway for both patient groups 1 and 2. For patient group 1, some patients did not appear to go through a *Dose Planning CT*.

The time delays between different activities compared to the ones published on patient information websites are presented in Table 4. The *First Appointment at Surgery Unit* is about a week after receiving the *Referral*, whereas in this data the delay was 10 days for Patient Group 1 and 9 days for Patient Group 2, suggesting a slightly longer delay.

Table 4. Comparison of the estimated and actual time delays between different activities

Activity	**Activity**	**Estimated delay**	**Delay for Patient Group 1**	**Delay for Patient Group 2**
Referral	First Appointment at Surgery Unit	7 days	10 days	9 days
Referral	Breast-Conserving Surgery	21 days	24 days	21 days
Breast-Conserving Surgery	Physician at Surgery Unit	14 days	18 days	17 days
Breast-Conserving Surgery	Adjuvant therapy	1 month	1 month	1 month
Chemotherapy	Chemotherapy	21 days	NA	20 days
Dose Planning CT	Radiation Therapy	7 days	8 days	8 days
Radiation Therapy	Radiation Therapy	1 day	1 day	1 day

In this study, the breast-conserving surgery was performed on average 14 days after the *First Appointment at Surgery Unit* for patient group 1 and 12 days after for patient group 2 and the delay between the *Referral* and *Breast-Conserving Surgery* was 24 and 21 days respectively. This is compatible with the estimated three-week delay between these two events (HUS 2020a). The follow-up appointment at the surgery unit after the *Breast-Conserving*

Surgery took place on average 18 days after the surgery for patient group 1 and 17 days after for patient group 2, instead of the planned two weeks (HUS 2021).

The treatment was planned to continue with either the *First Appointment at Chemotherapy Unit* for patient group 2 or the *First Appointment at Radiation Therapy Unit* for patient group 1 in a month after the *Breast-Conserving Surgery* and the data showed that this estimate is accurate. The first *Chemotherapy* for patient group 2 was administered on average 10 days after the first appointment and repeated every 20 days. This conforms with the estimated cycle of three weeks. (Suomen rintasyöpäyhdistys 2021)

Radiation Therapy started on average 8 days after the dose planning CT for both patient groups, which is close to the estimated one-week delay. Also, the *Radiation Therapy* was administered daily as expected and repeated 15 times for patient group 1 and an average of 15.8 times for patient group 2, which is between the estimated 15–30 times. (Suomen rintasyöpäyhdistys 2021; Terveyskylä 2022d)

In summary, the discovered process model and the planned hand-drawn model are very similar for both patient groups. There are slightly longer delays between some activities than expected, must most activities take place in the patient's clinical pathway in the same order and on the same schedule as expected.

5.4. Result Analysis

The results of this thesis consisted of the process discovery and conformance checking for breast cancer patients. The analysis was conducted on two different subgroups of the data:

1) the patients treated with breast-conserving surgery and radiation therapy and
2) the patients treated with breast-conserving surgery, chemotherapy, or other cancer drug therapy and radiation therapy

Process discovery for including all patients in groups of 1 and 2 resulted in a spaghetti-type process model which made it impossible to interpret with too many variants or orders of the activities for patients. Since the aim of the study was to identify the typical clinical pathway, the variants or orders of the activities that had a low frequency were excluded from the analysis. In patient group 1 the top four of the variants had a higher frequency, after which the

order of the patient activities was only similar for a couple of patients in each variant. In patient group 2 identifying the most frequent variants was more challenging and even the most frequent variants had only frequencies between 2 and 4. All the rest of the patients in patient group 2 had a unique order of treatment activities and were excluded from the analysis. Since only the most frequent variants were selected for analysis, they represented only a small portion of the patients, 25.9% of the patients in patient group 1 and only 8.4% of the patients in patient group 2. These are considered the typical clinical pathway for breast cancer treatment in this study. The percentage of patients following these typical pathways was expected to be considerably higher. However, selecting only the high frequency variants in the data may result in low percentage of cases. In Vidgof's et al. (2020) study, the most frequent variant only corresponded to 3% of all traces in the eventlog and most variants were unique. Arias et al. (2020) were able to cover 28% of their patient's pathways in their study with most frequent variants. Compared to these studies, the results of this study represented well the eventlog data especially in patient group 1.

The process discovery models were created with directly-follows graphs and the inductive miner. The directly-follows graphs visualized the frequencies and the delays between the different activities. The inductive miner visualized the processes with a process tree that showed the order and the structure of different activities and as a Petri net that showed the transitions between the activities and could be used for testing the model performance. The model performance (Table 5) shows that both the models performed well. The inductive miner for patient group 2 resulted in slightly lower performance compared to the model for patient group 1 which could be explained through more deviations and longer pathways in patient group 2. Overall, the models represent the eventlog well and there is no indication of over- or underfitting.

Table 5. Inductive Miner model evaluation

	Inductive miner for the typical pathway for Patient Group 1	**Inductive miner for the typical pathway for Patient Group 2**
Fitness	100%	100%
Precision	0.624	0.529
Generalization	0.896	0.757
Simplicity	0.806	0.739

Similar to previous research, this study demonstrated the complexity of patient treatment and procedures in healthcare, resulting in unique and nontrivial processes (Rebuge & Ferreira 2012; Garcia et al. 2019). The typical clinical pathway was challenging to find for a large, heterogeneous group of patients. As previous studies have proved, including all patients' data in the analysis resulted in a spaghetti-type of process model which did not follow the planned structure and included an overwhelming amount of behavior. (Partington et al. 2015; van der Aalst 2016; Erdogan & Tarhan 2018; Dallagassa et al. 2021)

Process mining can help to understand the variants and touchpoints of the process that affect the patient's experience (Arias et al 2020). Process mining gave good insights into the data of patients treated for breast cancer and proved that the process mining technique can be used for this type of health data at HUS. Process discovery algorithms were able to discover the clinical pathway for the patients and the log data recorded in electronic health records proved to be satisfactory for this type of analysis, reinforcing the previous knowledge on process mining in the healthcare domain. (Rovani er al. 2015; Garcia et al. 2019; Ibanez-Sanchez et al. 2019; Arias et al. 2020; Martin et al. 2020).

This study strengthens the research in process discovery and conformance checking of the process models in the field of healthcare of which the latter may be underutilized according to Partington et al. (2015). In the conformance checking part of this study, one was able to find some deviations between the planned process and the actual process, similar to results by Rovani et al. (2015), Erdogan & Tarhan (2018), and Ibanez-Sanchez et al. (2019). The time delays between the activities in this thesis were very close to the estimated or planned time delays within the focus groups included in the analysis. Process mining can also be useful in visualizing the differences and commonalities between multiple process models as Kurniati et al. (2019) proved in their study. Although in this study, most of the surgery and radiation therapy parts of the process were identical between the typical pathway of the two patient groups, some differences were also identified. The typical pathway of patient group 1 had on average 15 radiation therapy sessions, whereas patient group 2 had 15.8. For some patients in patient group 1, the activity of *Dose Planning CT* was missing.

Identical to other process mining applications in healthcare, this study concentrated on a very specific context. The parameters used for analysis, such as labeling the activities in the patient's clinical pathway, required some clinical knowledge by the researcher and the application is not repeatable in other settings, hospitals, or patient groups. (Partington et al. 2015; Garcia et al. 2019; Martin et al. 2020; Dallagassa et al. 2021)

Reducing the complexity of the data and the noise of the models proved to be important in this study, reinforcing previous knowledge on the process mining health care data (van der Aalst 2012; Rebuge & Ferreira 2012; van der Aalst 2016, 39; Leemans 2017; Vidgof et al. 2020; Dallagassa et al. 2021) The data showed that it is challenging to find the correct activity labels for events since the same visit types are used for multiple outpatient visits. Complex rules for automatic data labeling were built manually for this study. Grouping of activities and a higher-level abstraction of the events could be useful for this type of study, which is similar to Rabbi's et al. (2020) findings.

Pika et al. (2020) concluded that privacy protection of health care data while preserving data utility for process mining can be challenging. The different aspects of privacy with healthcare data proved to be important in this study. The eventlog attribute values could possibly reveal sensitive information, and that is why the report of the thesis does not hold information about the patient identifier, timestamps, or even the resources. Since the activities are labeled on multiple conditions, it is not possible to map a single visit type directly to the activity label presented in this thesis without accessing the securely saved Databricks notebook. Only the activity labels and the average delays between the different activities are presented in the graphs. Also, the atypical behavior, which could compromise the identity of some patients, is not present in this study since the aim was to find the typical clinical pathway and a single variant cannot be followed from the graphs. The typical pathways that were discovered in this study were quite comparable with the planned process model also presented on the patient information websites. (Pika et al. 2020)

This study did not concentrate on the long-term goal of reporting the outliers of the typical breast cancer treatment. Since the typical clinical pathway only represented a minority of the patients, this approach needs more investigation. This study succeeded in developing basic tools for calculating the delays between the different activities in the treatment and this could be useful for designing the reporting of outliers. The delays between the activities could also

be utilized for recognizing the bottlenecks and improving the efficiency of the treatment process in the future, similar to the study conducted by Erdogan & Tarhan (2018).

6. Conclusions and Discussion

This thesis aimed to find the typical clinical pathway for breast cancer patients at Helsinki University Hospital with PM4Py process mining tools. The process mining analysis consisted of process discovery and conformance checking with the existing, hand-drawn process model. The event log data was collected from HUS Data Lake for the years 2020 and 2021, preprocessed, and split into two subsets based on the treatment of the patient: if the patient was treated with breast-conserving surgery and radiation therapy they were categorized into patient group 1 and if they were also treated with chemotherapy or other drug therapy, they were categorized into patient group 2. The typical pathway was discovered with 4 most frequent variants for patients in patient group 1 representing 163 patients and 7 most frequent variants in patient group 2 representing 19 patients. The conformance checking showed that the discovered typical process was very similar to the hand-drawn process models and the time delays between the different activities were close to the estimated delays.

6.1. Conclusions

The results show that the clinical data from HUS is useful for this type of analysis, and it can give new insights into the data and the process. Handling and filtering the eventlog data can be challenging and give feedback on how users are recording the data in patient information systems.

The analysis of this study showed that the typical clinical pathway was challenging to find among heterogeneous eventlog data (van der Aalst 2016, 39; Leemans 2017; Vidgof et al. 2020). In order to avoid the spaghetti-type of a process model, the cohort should be carefully chosen to only include one treatment protocol for one patient group. Identifying the typical pathway is perhaps easier for processes that have only a limited number of activities since increasing the number of activities also increases the possible combinations of orders for the activities. Data preprocessing is critical for this type of analysis and requires medical knowledge from the field.

The results of this thesis show that the PM4Py framework's tools performed well for this type of analysis where the goal was to identify the typical process among many variants. As a practical implication of this study, the average time delays between the different activities are useful if the process mining approach would be used for monitoring the patients going through the treatment, notifying the clinicians when the delay is greater than expected.

6.2. Answering the Research Questions

This thesis aimed to answer the two following research questions:

1. How is process mining currently applied in healthcare analytics?

2. What do process mining discovery models of breast cancer patients' look like at Helsinki University Hospital and how do they compare with the planned process model for breast cancer patients?

The first research question was answered through a literature review of process mining studies in the healthcare field. The studies showed that process mining is applied in healthcare in numerous ways (Partington et al. 2015; Garcia et al. 2019; Ibanez-Sanchez et al. 2019; Helm et al. 2020; Dallagassa et al. 2021). The most popular approach is the process discovery of different healthcare processes (Partington et al. 2015; Garcia et al. 2019; Helm et al. 2020; Dallagassa et al. 2021). Conformance checking and process enhancement are less popular approaches but give valuable information about the patient experience, bottlenecks in the process, the distribution of resources, and suggestions for improving the efficiency of the processes (Rovani et al. 2015; Erdogan & Tarhan 2018; Ibanez-Sanchez et al. 2019; Arias et. al 2020).

Process mining analysis has also been used for identifying subgroups of patients within the processes and for predicting the patient behavior based on patient characteristics (Baek et al. 2018; Ibanez-Sanchez et al. 2019; Valero-Ramon et al. 2020). Most studies focus on one process within one organization (Partington et al. 2015). The processes in healthcare are often complex and unique, and the process mining techniques should be able to handle a large amount of noise (Rebuge & Ferreira 2012; Garcia et al. 2019).

Research question 2 was answered with a process discovery model, generated with the PM4Py-tool from breast cancer patients' data. The discovered model was compared with the handmade process model for conformance. The process discovery was conducted on the most frequent variants of two different eventlog datasets: patient groups 1 and 2. The directly-follows graphs (Figure 8 and Figure 12) that visualized the processes with rounded delays between the activities were easily understandable even without a prior understanding of the process or process models. The typical processes start with the referral to surgery, followed by the first meeting at the surgery unit, the breast-conserving surgery, and a follow-up meeting after the surgery. The treatment of patient group 1 is continued with radiation therapy, with the first appointment with the physician at the radiation therapy, followed by dose planning computer tomography and then repeated visits of radiation therapy and a final appointment at the radiation therapy unit. The typical treatment for patient group 2 had the chemo or drug therapy between the surgery and radiation therapy, where the drug therapy started with the first appointment at the chemotherapy unit and was followed by repeated chemotherapy sessions and a follow-up meeting at the unit. The radiation therapy was followed after the drug therapy similar to patient group 1.

The conformance checking showed that the planned process for breast cancer treatment and the discovered models were very similar but some minor differences between the two could be identified. According to the process discovery models, most of the patients seem to have two appointments on the day of the first appointment at the surgery unit. Also, the follow-up appointment after the surgery appears to be absent for some patients. In patient group 1, some of the patients do not go through a dose planning CT and in patient group 2, the first appointment at radiation therapy is missing compared to the planned process model. The time delays between the activities were also very similar in the planned process models and the discovered models. Some delays were slightly longer than expected. The delay between the referral and the first appointment at the surgery unit was estimated to be seven days but according to discovered typical models, the delay was 10 days for patient group 1 and 9 days for patient group 2. The delay between the breast-conserving surgery and the follow-up appointment was estimated to be 14 days but in this study was 18 and 17 days for the two patient groups, respectively.

6.3. Validity and Reliability of the Study

The validity and reliability are important for measuring the trustworthiness of the study. Reliability of the study measures how well the results are repeatable in future studies and validity means evaluating how well the methods of the studies measure the right issue. (Hirsjärvi, Remes & Sajavaara 2009, 231)

The data pre-processing was an important part of the study and had a significant role in the validity and reliability of the results. Since it was not possible to go through each of the patients and their activities and manually label them, the rules of the activity labeling were based on clinical knowledge of the visit types which most of the cases followed. This could have led to some mislabeling of some patients if the visit types were recorded in a dissimilar way.

Filtering of the activities also had a significant impact on the results. Since all the visit types "letter" and "phone call" were excluded, it is possible that some of the relevant activities were lost. Since the data was from 2020 and 2021 when Covid-19 was a big health concern, especially for cancer patients, some of the hospital visits could have been replaced with phone calls. The length of treatment and number of activities seemed to have a lot of variation and it could be because the patients were not excluded if they had two different breast cancer treatments or if their treatment were on hold for some reason.

The reliability of the results for patient group 2 may not be very high since the typical clinical pathway only represented 19 patients out of 226 which is 8.4% of the patients within this group. The variants that were included in the analysis had frequencies between 2 and 4 compared to the excluded variants which were all unique. The results for this patient group indicate that there is a lot of variation in the activities for these patients and most patients have a unique clinical pathway. The pathways are also longer with more activities which may increase the risk of also having more possible combinations in the order of activities.

The data accuracy and representativeness are important in this study and the reliability of the results. The data was able to fulfill the suggested requirements stated by Pika et al. (2020): *all events belong to a particular case* and *timestamps are reliable and accurate.* The requirement *variables that represent the case identifier, and the events, are accurate* and may not be fulfilled completely, since the events are based on the data labeling rules and

may not cover every patient's visits completely and the rules are done by the researcher with her best knowledge.

There was also a lot of variety in the ways of booking the patient activities, especially in the radiation therapy appointments which resulted in challenges in the activity labeling. Since the patient information system Apotti had been just introduced to the HUS Comprehensive Cancer Center units when this data was collected, the booking of the appointments was considerably changed and some decrease in the operations occurred for employees to have more time to learn the new system. The researcher's limited experience with the new database could have also affected the data labeling.

If this study would be repeated in the coming years, the activity labeling should be reconsidered because the rules for booking the visits and the system itself have probably gone through some changes. The continuous development of cancer treatments should also be considered. Especially the drug therapies are versatile and could explain the challenges in finding the typical pathway for patient group 2.

According to this study, there were slightly longer delays between some activities than expected. The longer delays were present between the referral and the first appointment at the surgery unit and the surgery and the follow-up appointment after it. The typical clinical pathways represented only a minor number of the patients in patient groups 1 and 2. Therefore the time delays calculated for these patients do not represent all patients treated for breast cancer and should not be used for drawing conclusions about the treatment of all patients. Since the delays were only 2-3 days longer than expected the differences in the delays could be caused by weekends, public holidays, or simply by the schedule of the patient and physician appointments.

A thorough explanation of all steps towards the analysis and how the results were drawn from the data should increase the validity and reliability of this study. Increasing the patient cohort for example through collecting the data from previous years could have increased the validity of this study, but the data preprocessing would have been even more complex with several patient information systems and the patient treatment protocols could also have been developed, making the results even more complicated.

6.4. Future Research

The Process mining for Python framework provides many tools that were not tested in this study but that could give even more insights into the process in the future. The analysis of possible bottlenecks and simulation of processes could be useful approaches for health care data and could possibly help with increasing the efficiency of the processes.

Since this study was limited to only finding the typical pathway for breast cancer patients, it would be useful to examine the reason why the processes vary so significantly for the rest of the patients. It would be interesting to find which factors determine the variability in the patient clinical pathway and if it could be predicted from patient characteristics. This information could be harnessed for identifying patients that are at risk for prolonged treatment and to find sub-groups within the breast cancer treatments.

This study concentrated on one patient illness in one hospital. Repeating the study in multiple settings would give more reliable results of the treatment itself and would give an insight into how the processes differ between different organizations.

The data preprocessing with the activity labeling had an impact on the results of this study and was very time-consuming. It would be useful to investigate possible ways to automate the activity labeling. Creating ways to aggregate the activities reliably would be very beneficial. The aggregated data activity groups could be used for process mining and reduce the complexity of the models.

7. References

van der Aalst, W.M.P., Weijters, T. & Maruster, L. 2004. Workflow Mining: Discovering Process Models from Event Logs. IEEE transactions on knowledge and data engineering, 16 (9), 1128-1142.

van der Aalst, W.M.P 2012. Process mining: Overview and opportunities. ACM Trans. Manage. Inf. Syst. 3,2 (7).

Van der Aalst, W.M.P 2016. Process mining: Data Science in Action. Springer. Available: https://link.springer.com/content/pdf/10.1007%2F978-3-662-49851-4.pdf

van der Aalst, W.M.P 2019. A practitioner's guide to process mining: Limitations of the directly-follows graph. Procedia Computer Science 164, 321–328.

Adriansyah, A., Sidorova, N. & van Dongen, B.F. 2011. Cost-based Fitness in Conformance Checking. 2011 Eleventh International Conference on Application of Concurrency to System Design, 06, 57-66.

Arias, M., Rojas, E., Aguirre, S., Cornejo, F., Munoz-Gama, J., Sepúlveda, M. & Capurro, D. 2020. Mapping the Patient's Journey in Healthcare through Process Mining. International Journal of Environmental Research and Public Health 17, 6586.

Baek, H., Cho, M., Kim, S., Hwang, H., Song, M. & Yoo, S. 2018. Analysis of length of hospital stays using electronic health records: A statistical and data mining approach. PloS one, 13 (4), p.e0195901-e0195901.

Berti, A., van Zelst, S.J. & van der Aalst, W.M.P. 2019. Process Mining for Python (PM4Py): Bridging the Gap Between Process- and Data Science. ArXiv.

Buijs, J.C.A.M., van Dongen, B.F., van der Aalst, W.M.P. 2012. A genetic algorithm for discovering process trees. IEEE Congress on Evolutionary Computation, 1–8.

Daniel, F., Barkaoui, K. & Dustdar, S. 2011. Business Process Management Workshops. BPM 2011 International Workshops. Springer. Available: https ://doi.org/10.1145/22402 36.22402 57

Dallagassa, M.R., Garcia, C.D.S, Scalabrin, E.E., Ioshii, S.O. & Carvalho, D.R. 2021. Opportunities and challenges for applying process mining in healthcare: a systematic mapping study. Journal of ambient intelligence and humanized computing, 2021-02-08.

Erdogan, T.G. & Tarhan, A. 2018. A Goal-Driven Evaluation Method Based On Process Mining for Healthcare Processes. Appl. Sci. 8, 894.

Finnish Government 2021. 3.6.1 Restructuring of health and social services. Accessed 30.12.2021. Available: https://valtioneuvosto.fi/en/marin/government-programme/restructuring-of-health-and-social-services

Garcia, C.D.S., Meincheim, A., Junior, E.R.F., Dallagassa, M.R., Sato, D.M.V., Carvalho, D.R., Santos, E.A.P. & Scalabrin, E.E. 2019. Process mining techniques and applications - A systematic mapping study. Expert Systems With Applications 133, 260-295.

Ghawi R. 2016. Process Discovery using Inductive Miner and Decomposition. A Submission to the Process Discovery Contest @ BPM2016. Technical Report. Accessed 18.2.2022. Available: https://arxiv.org/pdf/1610.07989.pdf

Helm, E., Lin, A.M., Baumgartner, D., Lin, A.C. & Küng, J. 2020. Towards the Use of Standardized Terms in Clinical Case Studies for Process Mining in Healthcare. International Journal of Environmental Research and Public Health 17, 1348.

Hirsjärvi S, Remes P & Sajavaara P 2009. Tutki ja kirjoita. Kustannusosakeyhtiö Tammi, Helsinki.

HUS 2020a. Rintarauhaskirurgia. Accessed 12.2.2022. Available: https://www.hus.fi/hoidot-ja-tutkimukset/rintarauhaskirurgia

HUS 2020b. Apotti-potilastietojärjestelmän käyttöönotto etenee HUSissa. Accessed 26.2.2022. Available: https://www.hus.fi/ajankohtaista/apotti-potilastietojarjestelman-kayttoonotto-etenee-husissa

HUS 2021. Rintasyöpäpotilaan hoitopolku. Accessed 28.12.2021. Available: https://www.hus.fi/hoidot-ja-tutkimukset/rintasyopapotilaan-hoitopolku

HUS Syöpäkeskus 2018. Toimintakertomus. Accessed 21.12.2021. Available: https://www.hus.fi/tietoa-meista/potilashoito-laatu-ja-potilasturvallisuus/syopakeskus

HUS Tutkimus 2021. Tutkimuslupa, opinnäytetyön tutkimuslupa ja tietolupa. Accessed 29.12.2021. Available: https://www.hus.fi/tutkimus-ja-opetus/tutkijan-ohjeet/tutkimuslupa-opinnaytetyon-tutkimuslupa-ja-tietolupa

HUS 2022. About us. Accessed 29.1.2022. Available: https://www.hus.fi/en/about-us

Ibanez-Sanchez, G., Fernandez-Llatas, C., Martinez-Millana, A., Celda, A., Mandingorra, J., Aparici-Tortajada, L., Valero-Ramon, Z., Munoz-Gama, J., Sepúlveda, M., Rojas, E., Gálvez ,V., Capurro, D. & Traver, V. 2019. Toward Value-Based Healthcare through Interactive Process Mining in Emergency Rooms: The Stroke Case. International Journal of Environmental Research and Public Health 16, 1783.

Joensuu, H. & Rosenberg-Ryhänen L. 2014. Rintasyöpäpotilaan opas. Accessed 21.12.2021. Available: https://syopa-alueelliset.s3.eu-west-1.amazonaws.com/sites/271/2016/10/18170636/RintasyopapotilaanOpas.pdf

Kurniati, A.P., Rojas, E., Hogg, D., Hall, G. & Johnson, O.A. 2019. The assessment of data quality issues for process mining in healthcare using Medical Information Mart for Intensive Care III, a freely available e-health record database. Health informatics journal, 25(4), 1878-1893.

Leemans, S.J.J., Fahland, D.& van der Aalst, W.M.P. 2013. Discovering Block-Structured Process Models from Event Logs – A Constructive Approach. In: Application and Theory of Petri Nets and Concurrency, 34th International Conference, PETRI NETS 2013, 311–329.

Leemans, S.J.J., Fahland, D.& van der Aalst, W.M.P. 2015. Scalable Process Discovery with Guarantees. In: Enterprise, Business-Process and Information Systems Modeling 16th International Conference, BPMDS 2015, 20th International Conference, 85–101.

Leemans, S.J.J., Fahland, D.& van der Aalst, W.M.P. 2018. Scalable process discovery and conformance checking. Software and Systems Modeling, 17(2), 599–631.

Lenz, R. & Reichert, M. 2007. IT support for healthcare processes – premises, challenges, perspectives. Data & Knowledge Engineering 61(1), 39–58.

Martin, N., De Weerdt, J., Fernández-Llatas, C., Gal, A., Gatta, R., Ibáñez, G., Johnson, O., Mannhardt, F., Marco-Ruiz, L., Mertens, S., Munoz-Gama, J., Seona, F., Vanhienen, J., Wynn, M.T., Boilève, D.B., Bergs, J., Joosten-Melis, M., Schretlen, S. & Van Acker, B.

2020. Recommendations for enhancing the usability and understandability of process mining in healthcare. Artificial Intelligence In Medicine 109 101962.

Mattson, J., Auvinen, P., Bärlund, M. & Jukkola-Vuorinen, A. 2016. Rintasyöpäpotilaan seuranta. Duodecim 132:2317–23.

OMG 2011. Business Process Model and Notation (BPMN). Version 2.0. Accessed 18.2.2022. Available: https://www.omg.org/spec/BPMN/2.0/PDF

Partington, A., Wynn, M., Suriadi, S., Ouyang, C. & Karnon, J. 2015. Process Mining for Clinical Processes: A Comparative Analysis of Four Australian Hospitals. ACM Transactions on Management Information System 5(4), Article 19.

Pika, A., Wynn, M.T., Budiono, S., ter Hofstede, A.H.M., van der Aalst, W.M.P. & Reijers, H.A. 2020. Privacy-Preserving Process Mining in Healthcare. International Journal of Environmental Research and Public Health 17, 1612.

PM4PY 2021. Documentation for version PM4Py 2.2.17.1. Accessed 29.12.2021. Available: https://pm4py.fit.fraunhofer.de/documentation#conformance

Rabbi, F., Lamo, Y. & MacCaull, W. 2020. A Model Based Slicing Technique for Process Mining Healthcare Information. International Conference on Systems Modelling and Management 73–81.

Rebuge, Á. & Ferreira, D.R. 2012. Business process analysis in healthcare environments: A methodology based on process mining. Information Systems 37, 99–116.

Rovani, M., Maggi, F.M., de Leoni, M. & van der Aalst, W.M.P. 2015. Declarative process mining in healthcare. Expert Systems With Applications 42, 9236–9251.

Suomen rintasyöpäyhdistys 2021. Rintasyövän valtakunnallinen diagnostiikka- ja hoitosuositus 2021. Accessed 5.2.2022. Available: https://1587667.167.directo.fi/@Bin/d9fd258ac00bf479c0d139e8148deb50/1644050325/application/pdf/190430/Suomen%20Rintasy%c3%b6p%c3%a4ryhm%c3%a4n%20hoitosuositus%202021.pdf

TENK 2012. Hyvä tieteellinen käytäntö ja sen loukkausepäilyjen käsitteleminen Suomessa. Accessed 2.12.2021. Available: https://tenk.fi/sites/tenk.fi/files/HTK_ohje_2012.pdf

Terveyskylä 2022a. Rintasyövän oireet ja toteaminen. Accessed 19.3.2022. Available: https://www.terveyskyla.fi/syopatalo/sy%C3%B6p%C3%A4tau-dit/rintasy%C3%B6p%C3%A4/rintasy%C3%B6v%C3%A4n-oireet-ja-toteaminen

Terveyskylä 2022b. Rintasyövän hoito leikkauksen jälkeen. Accessed 5.2.2022. Available: https://www.terveyskyla.fi/syopatalo/sy%C3%B6p%C3%A4tau-dit/rintasy%C3%B6p%C3%A4/rintasy%C3%B6v%C3%A4n-hoito-leikkauksen-j%C3%A4lkeen

Terveyskylä 2022c. Rintasyövän hormonaalinen hoito. Accessed 5.2.2022. Available: https://www.terveyskyla.fi/syopatalo/sy%C3%B6p%C3%A4tau-dit/rintasy%C3%B6p%C3%A4/rintasy%C3%B6v%C3%A4n-hoito-leikkauksen-j%C3%A4lkeen/rintasy%C3%B6v%C3%A4n-hormonaalinen-hoito

Terveyskylä 2022d. Rinnan sädehoito. Accessed 5.2.2022. Available: https://www.terveyskyla.fi/syopatalo/sy%C3%B6p%C3%A4tau-dit/rintasy%C3%B6p%C3%A4/rintasy%C3%B6v%C3%A4n-hoito-leikkauksen-j%C3%A4lkeen/rinnan-s%C3%A4dehoito

Terveyskylä 2022e. Rintasyövän seurantakäynnit hoitojen jälkeen. Accessed 12.2.2022. Available: https://www.terveyskyla.fi/syopatalo/sy%C3%B6p%C3%A4tau-dit/rintasy%C3%B6p%C3%A4/rintasy%C3%B6v%C3%A4n-seurantak%C3%A4ynnit-hoitojen-j%C3%A4lkeen

Valero-Ramon, Z., Fernandez-Llatas, C., Valdivieso, B. & Traver, V. 2020. Dynamic Models Supporting Personalised Chronic Disease Management through Healthcare Sensors with Interactive Process Mining. Sensors 2020, 20, 5330.

Vehmanen, L. 2020. Rintasyövän hoito. Lääkärikirja Duodecim 16.9.2020. Accessed 19.3.2022. Available: https://www.terveyskirjasto.fi/dlk00468.

Vidgof, M., Djurica, D., Bala, S.& Mendling J. 2020. Cherry-Picking from Spaghetti: Multi-range Filtering of Event Logs. Enterprise, Business-Process and Information Systems Modeling. BPMDS 2020, EMMSAD 2020. Lecture Notes in Business Information Processing, vol 387. Springer.