UNIVERSITY OF TARTU
Institute of Computer Science
Innovation and Technology Management Curriculum

Musa Salamov

Process mining on Estonian healthcare data

Master's Thesis (20 ECTS)

Supervisor(s): Fredrik Milani
Sulev Reisberg

Tartu 2023

Process mining on Estonian healthcare data

Abstract:

Nowadays, healthcare information is recorded and stored electronically in most countries. Data-driven methods, such as process mining, can be used for analyzing healthcare processes by utilizing electronic health records. However, despite its potential, process mining has not been applied to healthcare settings in Estonia. Therefore, this thesis aims to assess the feasibility of process mining for Estonian healthcare data. Additionally, it researches what are the limitations and suggests improvement ideas. To achieve this objective, cervical cancer screening-related health records were analyzed using Apromore. The main results of this thesis are the list of applicable process mining use cases within the context, their limitations and a set of improvement suggestions. Thus, the contribution of this thesis is beneficial for experts and researchers who are involved in the application of process mining in the healthcare domain.

Keywords:

process mining, electronic healthcare records, clinical pathway, process analysis

CERCS: P170 Computer science, numerical analysis, systems, control

Protsessikaeve Eesti terviseandmetel

Lühikokkuvõte:

Tervishoiuinfo tekib ja seda säilitatakse tänapäeval valdavalt elektroonilisel kujul. Sellise info analüüsimiseks saab kasutada mitmesuguseid andmepõhiseid meetodeid, näiteks äriprotsesside kaevandamist. Hoolimata selle meetodi potentsiaalist pole protsessikaevandamist Eesti terviseinfol teadaolevalt varem rakendatud. Käesolev töö analüüsib protsesside kaevandamise otstarbekust Eesti tervishoiuandmetest, selle piiranguid ja parendusvõimalusi. Täpsemalt analüüsiti Apromore'i abil emakakaelavähi sõeluuringuga seotud protsesse Eesti kolmest andmekogust loodud koondandmestikul. Selle lõputöö tulemuseks on protsesside kaevandamise kasutusjuhtude loetelu, nende piirangud ja parendusettepanekud.

Võtmesõnad:

protsessikaeve, elektroonilised terviseandmed, haigustrajektoor, protsesside analüüs

CERCS: P170 Arvutiteadus, arvanalüüs, süsteemid, kontroll

Table of Contents

1 Introduction

The efficiency of healthcare services has become a vastly important aspect that needs to be considered alongside the quality [1] because of the rapidly changing nature of the healthcare industry. Taking into consideration that it is necessary to maintain high quality while possessing limited resources in terms of labor, tools, managing the process for more efficient outcomes becomes a vital part of the service. Consequently, process management and especially process mining have gained more popularity than ever in hospitals since the emergence of COVID-19 [2].

Nowadays, information systems possess an excessive amount of information which usually also includes the records of events. Essentially, it is possible to get an overview of how the processes work in a real-world context, detect if there is any difference between a given process and real data of the process, and, if possible, enrich a process with process mining techniques. Process mining is a data-driven approach for mainly discovering, monitoring and improving the process by getting details from event logs [3]. Because of the nature of the approach, initially, it purely relies on real data. As a result, it is not affected by the narrative of a specialist in the domain.

In the last decades, process mining has been applied to the healthcare domain and several systematic literature review studies have been conducted. In the healthcare domain, processes can be inspected for two different purposes [4]. First, from a clinical point of view, any unwanted or extraordinary paths of tasks can be discovered. Second, from a managerial point of view, any causes that affect the efficiency of the process either positively or negatively can be revealed. The main applications of process mining in healthcare are process discovery and conformance checking, but it has to be noted that the additional assistance of healthcare professionals is usually needed in terms of clinal insights [4], [5].

The application of innovations to medical fields has not often been very welcomed [6]. However, the requirements of society have been impacted that frame of mind positively, resulting more openness toward innovations. Rojas et al. reveal that there have been case studies for several medical fields, and process mining techniques are potentially applicable to all medical fields in one way or another [7]. Additionally, oncology and surgery are the medical fields that are the most case studies are researched. However, healthcare in each country differs from others.

In light of this, process mining is a promising technique and potentially have useful use cases that could be applied in the healthcare domain. Although its potential within the context, its applicability has never been researched in Estonian healthcare settings. The application of process mining use cases and its limitations on the data from Estonian national health databases have not been covered so far by academic research. Thus, the main aim of this research is to identify how and to what extent are feasible process mining use cases to be applied in the given context. There are three research questions that are being addressed in this thesis:

RQ1: *What process mining use cases can be applied to Estonian healthcare data?*

RQ2: *What are the limitations that restrict the implication of process mining to Estonian healthcare data?*

RQ3: *What changes would be required to improve the feasibility of process mining on Estonian healthcare data?*

The main findings of this study provide usability evaluation for process mining on Estonian healthcare data and help to improve applicability of process mining on health data in the future. The contribution of this thesis is, therefore, two-fold. The first contribution is an overview of applicability of process mining use cases on Estonian healthcare data. The second is a set of suggestions that, if applied to the Estonian health data, would increase process mining applicability. Thus, it can be utilized by domain experts who are engaged in analyzing clinical pathways and researchers who aim to eliminate the limitations with Estonian healthcare data. First, feasibility and limitations of process mining use cases will be identified analyzing the data using a process mining tool. Then it is followed by investigation of possible improvement opportunities by implementing possible solutions. For these purposes, a case will be selected from the healthcare data obtained in Estonia, and data analysis will be conducted on that case.

The remainder of this thesis is organized in the following manner. In Section 2, key concepts such as business process management, process mining, healthcare data, which outline the background, are explained. Section 3 details current academic literature regarding related works to this thesis. In Section 4, research process is described in a detailed way. Section 5 outlines the results of conducted data analysis. It is followed by the discussion of the results and limitations of the study in Section 6. Section 7 is a conclusion of the thesis, which summarizes all findings and outlines future implications.

2 Background

This section serves to introduce and explain the key concepts that form the foundation of the thesis. The first concept presented is that of business process management, which is followed by an examination of process mining and its lifecycle. This section also highlights healthcare data in Estonia.

2.1 Business Process Management

As the name implies, Business Process Management (BPM) refers to the field that deals with business processes. According to a definition by Hammer and Champy [8], a business process is "a collection of activities that takes one or more kind of input and creates an output that is of value to the customer". At its essence, a business process should have an input for creating a value as an output, which are forming the main features of a business process alongside a logical flow and a set of activities [9]. Marlon et al. [10] define a business process more explicitly as "a collection of inter-related events, activities, and decision points that involve a number of actors and objects, which collectively lead to an outcome that is of value to at least one customer". The quality and efficiency of services which are provided by an organization to its customers are dependent on the design and performance of processes. Therefore, it is clear that managing the processes can give an organization a competitive edge over its competitors while they provide similar services. Moreover, this applies not only to processes that interact directly with customers but also internal processes [10].

BPM is "a body of methods, techniques and tools to discover, analyze, redesign, execute and monitor business processes" [10]. Based on this definition, it is clear that BPM has different phases and it indicates that business processes are the center of interest of the discipline. Marlon et al. visualized all phases of the BPM lifecycle (See Figure 1) and considered it as a continuous cycle. The main idea behind treating it as a continuous cycle is the vital need of continuous improvements to keep the quality and efficiency of services at certain levels.

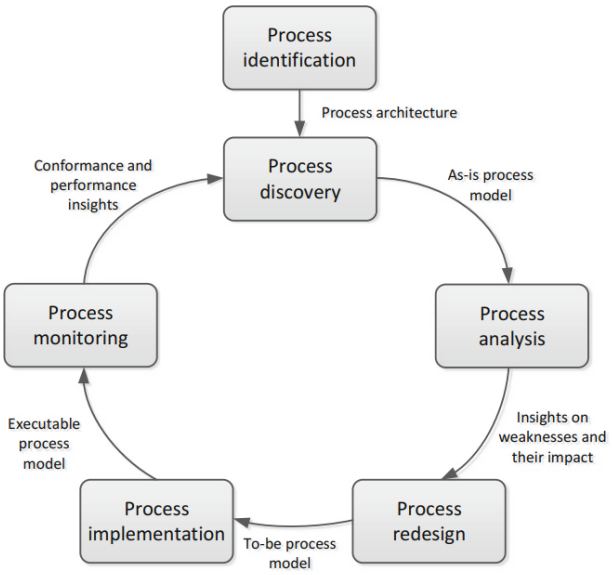


Figure 1. The BPM Lifecycle [10]

The first phase is process identification, which includes posing a business problem and an organization identifies the processes that require re-engineering, automation or improvement to address the business problem. It is followed by the process discovery phase, where an organization analyzes the current processes and produce one or more "as-is" process models. Process analysis, which is the third step, involves the analysis of identified process models to identify the areas that need improvement. After identifying the needs, bottlenecks or inefficiencies, an organization design new process models that overcome the problems of current "as-is" models and the typical outcome of process redesign phase is a "to-be" process model. The next step is implementation of the changes from "as-is" to "to-be" models, which implies that the new processes are implemented and integrated into the organization. While process monitoring may be the last phase, it is by no means any less significant than others. An organization needs to monitor the redesigned process by gathering relevant data and analyze it to track the performance of the new process. Furthermore, as there may be new issues in the process, a continues cycle is required to maintain the efficiency and quality[10].

Despite the fact that BPM techniques are very useful to manage the business process of an organization and provide insight regarding performance, they lack any connection to actual real-world data [11]. Process mining techniques eliminate that problem and are able to provide better results as it directly involves real-world event data.

2.2 Process Mining

In the past, legacy information systems were not process-aware, meaning that they did not store the details of business process activities [12]. Today, however, information systems are mostly able to generate event logs to trace business process activities. Modern enterprise information systems are usually capable of recording business events and also storing relevant events in a structured form [13]. In Process Mining Manifesto [3], it is clearly stated that "the idea of process mining is to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs readily available in today's (information) systems". It is worth mentioning that as process mining techniques rely on real event data, it enables catching the details for business activities that are not initially assumed or predicted. As in the real world, everything does not always happen as it is planned, the idea of utilizing readily available knowledge from information systems is promising. As van der Aalst et al. claim, process mining facilitates BPM based on evidence [14], instead of models made manually [15].

Dumas et al. [10] explain an event log as a collection of timestamped event records and state the usual minimum requirement for each event in a log. Three main attributes are mandatory for most process mining techniques. First, there has to be an identifier that indicates in which case the event occurred. Incidentally, a case and an event may be easily misinterpreted; therefore, it should be noted that a case is an end-to-end instance of the process (e.g. order-to-cash) and an event is the occurrence of activity or task within the process. Second, the task or activity that the event refers to, and third, a timestamp indicating when the event occurred. Fundamentally, it is possible to use process mining techniques with only these three attributes; however, additional attributes can provide better insight. Hence, extra

information, which may also be industry-specific, including but not limited to the resource and the timestamps initiation and conclusion of an event, can be of significant value. The classification of process mining techniques can possibly be based on different concepts. According to var der Aalst [16], there are three main types of process mining: process discovery, conformance checking and model enhancement. Dumas et al. [10] introduces variant analysis in addition to these three types (see Figure 2). Performance mining and model enhancement are used interchangeably.

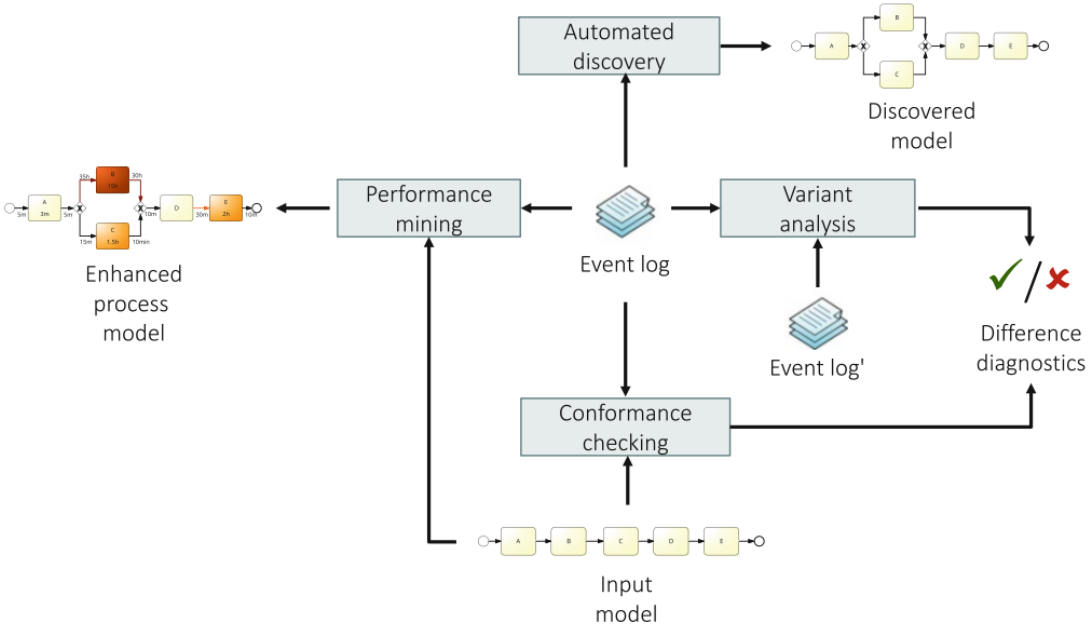


Figure 2. Four categories of process mining, required inputs and outputs [10]

Process discovery, as its name implies, is basically discovering the process from the event logs, and its output is a process model that corresponds the flow of activities in the event log. There is no need for a pre-defined business process model in process discovery; however, while it is not always necessary, it is often required for conformance checking. In essence, conformance checking is a comparison of an existing process model and the model discovered from the event logs. Business rules or policies can be used as an input instead of a process model to identify the difference. While conformance checking evaluates the alignment between model and reality, the goal of performance mining is to change the process model for improvements [3]. Last but not least, variant analysis is the comparison of two variants of the same event log. It is a great way to analyze positive and negative outcomes of the process, find the differences and reveal why a desirable outcome is not achieved. The figure also describes that the initiation of process mining techniques depends on the event logs which are extracted from real-world processes via information systems. Moreover, the quality of event logs may affect all types of process mining.

At the core, discovering the sequence of activities can be considered as a main part of process mining because it usually is inspiring for the people who either work in the field or study it. Therefore, the control-flow discovery can spark new possibilities as well as new ideas. However, the other forms of process mining should also be taken into consideration

with great concern. Additionally, the organizational, case and time perspectives are also in the scope of process mining. Hence, these aspects play a role as significant as, if not greater than, control-flow discovery [14].

As noted earlier, the quality of event logs plays also quite important role in process mining techniques, and it is judged based on several criteria. Considering the criteria, such as events should be trustworthy and complete, it is possible to define the maturity levels of event logs. Organizations must aim to acquire event logs of the highest quality to facilitate process mining techniques. Despite event logs, which are at lowest maturity levels, can also be utilized, the analysis and results are not typically reliable [3].

2.3 Healthcare data in Estonia

Estonia, a Northern-Eastern European country with a population of 1.4 million, is the first country to implement a nationwide electronic health records (EHR) system [17]. Estonian health data is collected by different governmental organizations; however, interoperability of the data is the main goal. All healthcare providers in Estonia are required to use three central health databases to achieve the goal [18]. These databases store information regarding health insurance claims, electronic health records, and drug prescriptions.

From a patient's point of view, all health information is available in the patient portal (Digilugu). Although it may be perceived as a centralized database, in reality, it retrieves the required data from different healthcare providers. Additionally, a patient's records are available for doctors, hence, doctors can access necessary data when it is needed.

3 Related work

This section summarizes previous related work and provides an analysis of available knowledge for the topic of the thesis.

Despite the fact that healthcare is quite an important industry, the application of process mining in the healthcare industry is not researched extensively. Nevertheless, it does not mean that there are not studies exploring process mining in healthcare. There are several papers that discuss the implication of process mining techniques for health services. However, to date, the limited academic literature [19] does not provide detailed insights about the implications on healthcare in Estonia.

In one of the first papers in the field, Mans et al. [20] analyzed the opportunities for applying process mining techniques in a Dutch Hospital. The authors investigated the healthcare processes from control-flow, organizational and performance perspectives. Raw data, which include information about 627 gynecological oncology patients, were extracted for analyzing healthcare processes, in which all diagnostic and treatment activities have been recorded. Overall, the paper demonstrates the possible implementations of process mining techniques to health data recorded in a large academic hospital in the Netherlands and it discusses the potential of process mining in healthcare. The focus of this thesis is also to analyze how feasible process mining techniques to healthcare domain. However, the scope is not limited to only one source of data recorded in a hospital as aforementioned paper. Therefore, it enables to explore the complexity of data acquired from multiple database systems at the national level in Estonia.

Perimal-Lewis et al. [21] used process mining methodology to assess the data quality of electronic health records. The researchers worked closely with domain experts to be able to validate the process models. The flow abnormalities were fetched during the analysis and represented to the experts. Moreover, according to feedbacks from the experts, the authors demonstrated that process mining can be considered as a feasible methodology to assess the data quality. The main findings showed the effects of incorrect timestamp data used in time-based performance metrics. Additionally, the authors were able to provide possible corrective actions that are necessary to address the data quality issues. The study provides generalizable actions to avoid common issues in most hospitals regarding data quality. This work extends it with the application of process mining from other perspectives.

Toth et al. [22] explored the applicability of process mining in healthcare with the main focus on challenges. The main challenges identified were the use of different codes for the same medical concept, the variability of treatments, and time-related problems. The authors suggested several recommendations to overcome these challenges, such as aggregating similar codes into one, sorting events in alphabetical order, and more. Additionally, the main contribution that the paper provides is a proposed workflow for generating more precise models for healthcare processes. That being stated, the outcomes of the paper can be considered as an insightful foundation of structuring methodology in this thesis. However, in essence, the paper focuses on only one type of process mining, process discovery. While the aim of this thesis is to identify the applicability of not only process discovery, but also

conformance checking, variant analysis and performance mining. Also, the proposed workflow overlaps with L* life-cycle model presented by Rojas et al [7].

There are a number of papers that discuss the data challenges within the context. Mans et al. [20] follow an interesting route to discover the challenges and provide solutions. They start with researching the most frequently posed questions by medical professionals. Mainly, the experts intend to identify whether the process complies with guidelines or not, the bottlenecks in the process, and common and exceptional paths in the process. Afterward, the authors classify the different types of event data found in health information systems and discuss the problems. As in the other papers, one of the main problems revealed is the granularity of the timestamps.

The challenges and promising directions in the field were demonstrated on the paper by Gatta et al.[4] recently, while providing a compelling overview as well. The main types of process mining techniques have several useful utilization possibilities. The authors explain that process discovery can be useful for both clinical and managerial issues, as well as, revealing the bottlenecks in the process. The main role of conformance checking can be assessing the quality of provided healthcare service, while testing the models in reality versus in guidelines. In the paper, process enhancement is described as the most unexplored type, and it is recommended to explore AI-empowered opportunities within the field to achieve better results than the outcomes of manually performed enhancements. On the other hand, the authors mention the challenges regarding scheduling, planning, patients' privacy as well as understandability [23], [24].

Kusuma et al. [25] reviewed a potential set of 156 articles that were related to disease trajectory models, their patterns over time and process mining. However, the number of papers that directly applied process mining to disease trajectory models were only four. The paper revealed that none of the studies has used national level health records data for analyzing the disease trajectories. National level health records are used in this thesis. It is worth mentioning that there are other studies which has used health records at national level. Fox et al. [26] applied a framework to a case study of dental care pathways covering 41 dental clinics in the Republic of Ireland; however, the focus of the study was to develop and apply a data quality framework for process mining.

In terms of methodology to use in process mining projects, van Eck et al. [27] introduced a methodology called PM^2 as a guide. The authors believe that previous well-known methodologies for process mining lacked suitability for every project. They argue that Process Diagnostic Method [28] and L* life cycle [7], which are the two most known methodologies, have limitations regarding the applicability for complex projects, iterative analysis and practical guidelines. Rebuge and Ferreira [29] presented a methodology for application of process mining on the data recorded in healthcare information systems. Essentially, all mentioned methodologies possess mostly same components and differ slightly in some areas from others. PM^2 is designed to support projects which aims at compliance and improving the performance of the process.

Vathy-Fogarassy et al. [30] proposed a methodology called Process Mining Methodology for Exploring Disease-specific Care Processes (MEDCP) in one of the most recent papers.

The main need for a such domain-specific process mining methodology has arisen because of the main problems of data quality and the high variability of treatment processes in healthcare. Like other methodologies, MEDCP follows a similar workflow and includes comparable steps. Data collection and creating event logs are the initial steps of the methodology. The paper proposes that the analysis should not necessarily be limited to events recorded from only one information system. Typically, the data is stored in different databases. Therefore, to ensure a comprehensive analysis of the disease under study, it is crucial to include all relevant activities in the data collection process. Furthermore, after creating structured event logs based on the acquired data, preprocessing of the log file is the next step. This step includes several sub-activities to be executed. Defining the starting and end events, deleting the treatment events which are not important, removing events which do not have direct relevance to the treatment can be considered as the main sub-activities of the step. The next step, which is quite vital to the analysis, is the integration of task-specific knowledge in the form of taxonomies. The key idea is to explore process models at different abstraction levels according to the taxonomies defined by domain experts. It is typically achieved by classifying events into higher-level event cohorts or groups, however it also involves other possible aggregations of events as well handling problems related to timestamps. Moreover, the last step is using process mining tool or domain-specific framework. Consequently, both methodologies, MEDCP and PM^2 consist of very useful principles, procedures and rules to guide this thesis. Therefore, this thesis utilizes the principles that are proposed in the papers of these methodologies to analyze the applicability of process mining for a specific disease related events that are recorded in Estonia.

Martin et al. [5] propose recommendation for process mining researchers who involve within the healthcare industry. The authors emphasize that healthcare guidelines and processes undergo changes due to various factors, including technological advancements and advances in medicine. Park et al. [31] presents process mining in a changing environment and explores the opportunities using the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) to analyze healthcare processes by utilizing different techniques of process mining. The paper outlines that currently there are not enough available data from administrative systems in healthcare. Therefore, it is important to obtain necessary information from other sources, which can be internal as well as external, to complement the data from health records. Valero-Ramon et al. [32] proposed an interactive tool, which is integrated to healthcare dashboard, to make clinical data more accessible to healthcare professionals. In fact, it supports clinical decision-making and performance review of the experts.

In conclusion, the literature review demonstrates that existing papers have studied process mining in healthcare with different techniques and methodologies. While some of the papers proposed new methodologies, the others empirically studied the opportunities. However, despite the growing interest in process mining application in healthcare, there is lack of studies that focus specifically on the context of Estonia. Given a healthcare system in Estonia produces enormous amount of electronic health records and it is not known that which process mining use cases can be applied to those records, there is a clear need for research to investigate the potential of process mining in this context.

4 Methodology

This section introduces the methodology used for this research. All the stages involved in addressing the research questions are described in Figure 3. It also includes the formulation process of research questions.

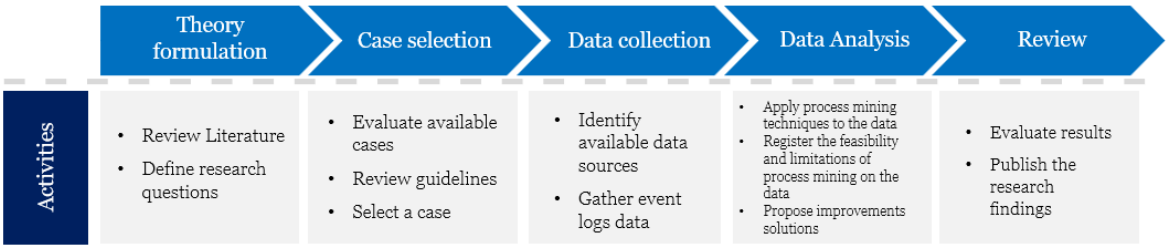


Figure 3. Methodology for the research

This work is divided into five main stages, which start with formulation of theory for the research. The main aim of the research is to investigate the application of process mining to healthcare data, which is obtained in Estonian health information systems. Thus, three research questions are formulated as follows:

RQ1: *What process mining use cases can be applied to Estonian healthcare data?* The question is associated with the main process mining categories, which are Process Discovery, Conformance Checking, Variant Analysis, Enhancement (Performance Mining), and their applicability to the acquired healthcare data. Therefore, the most common use cases which are related to these categories are the focus of this research.

RQ2: *What are the limitations that restrict the implication of process mining to Estonian healthcare data?* The motivation is to explore the extent to which Estonian healthcare data is limited or restricted by its characteristics in terms of the application of process mining use cases.

RQ3: *What changes would be required to improve the feasibility of process mining on Estonian healthcare data?* The last question relates to the data quality. The main idea is to identify how the dataset can be improved to increase process mining application in the given context. In addition to giving an overview of possible improvements based on the identified limitations, it also aims to provide possible implementations regarding improvements.

In order to address the research questions, a case of cervical cancer screening procedure was selected. Mainly because a guideline [33] exists for this particular screening process in Estonian healthcare settings. Aavik et al. [33] provide a national cervical cancer screening program, which includes the guidelines regarding diagnosis, monitoring and treatment of cancerous conditions of the cervix, and therefore, can be used as a reference process for this thesis. The main activities of cervical cancer screening, which will be actively used in the data analysis stage, will be explained in section 4.1. In the following subsections, the screening process, data collection and analysis are described in more detail.

4.1 Cervical cancer screening

Cervical cancer is among the most common cancers in women all over the world and Estonia is not an exception [33]. Similar to most of the other cancers, early detection is vital for better outcomes in treatment and therefore, several diagnostic tests should be carried out regularly. Consequently, the process mining techniques are potentially very useful for identifying the current pathways of the tests in comparison to general treatment guidelines.

The aim of cervical cancer screening is to identify precancerous cells and start treatment as early as possible [34]. There are a few crucial risk factors for cervical cancer, such as human papillomavirus (HPV). Thus, there is HPV test for checking high-risk HPV-related cervical cancer causes. There is also another test called Papanicolaou (PAP test) for collecting cervical cells and checking for changes. These can kind of tests can be done separately, but depending on the patient, co-testing can also be recommended. There are also several other factors such as patient's social and sexual history [34], which effect the screening procedure. Furthermore, depending on the test results and patient's age and other conditions, a flow of activities in a procedure can be changed. For example, there are abnormal test results, such as the observation of a high-grade squamous intraepithelial lesion (HSIL) or a low-grade squamous intraepithelial lesion (LSIL). Moreover, guidelines for cervical cancer screening [33] provide a recommended flow of activities when there are abnormal results. The flow is also dependent on the significance of issues. For example, atypical squamous cells of undetermined significance (ASC-US) almost always indicate an HPV infection [34]. As well as abnormal results, a negative result, which is called negative for intra epithelial lesion (NILM), is also possible. Within the scope of this thesis, it is necessary to understand that observation of such abnormal results is the result of particular tests (see section 5.2). Tests will be explicitly stated from observation of test results in the following sections when it is necessary to mention. However, this thesis does not require in-depth knowledge of activities in a cervical cancer screening procedure.

4.2 Data collection and description

The data acquisition process for this research is carried out within the University of Tartu. As it is already mentioned, the data is regarding cervical cancer screening in Estonia. Essentially, all electronic health records are stored in three Estonian national databases, which store information obtained from many healthcare settings [18]. The data, which is used in this research, is the result of the research project [18] conducted by Oja et al (see Figure 4).

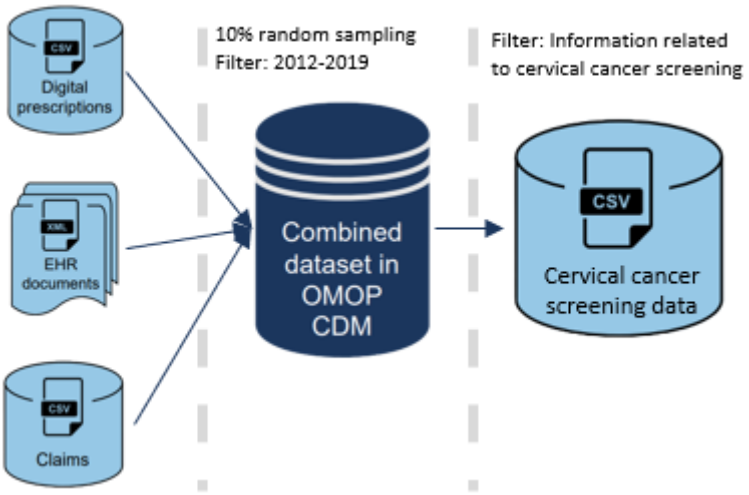


Figure 4. Data acquisition process [18]

The acquired data (see Appendix I) is a comma-separated values (CSV) file which includes information such as cohort ID, cohort name, cohort start date, subject ID, age. Cohorts are basically the set of similar activities that form an activity which is useful for process mining. For example, separate events related to contraception drug prescriptions are all combined into single cohort "Contraception drugs (G03A, G02B; some L02A)" - an approach previously described by Vathy-Fogarassy et al [30]. The reason why all activities should not be imported to the process mining tool without being combined into cohorts is the complexity of clinical processes. In addition to the complexity, activity of providing drug prescriptions are differentiated based on the drugs. Therefore, it is also necessary to combine them into one activity, as in essence, they all are the same activity if the prescription is written based on the same diagnosis. The need for event generalization when investigating health event sequences was also highlighted by Kunnapuu et al [35]. Furthermore, activities related to prescriptions are not the only activities which need to be combined. Otherwise, the analysis in process mining tools is not able to serve its purposes. Although it is still possible to use the activities without cohorts in practice. The analysis may not be as effective; however, it would allow to analyze specific cases more in-depth. For the clarity, effectiveness and wider implementations, cohorts are used in this thesis.

From the process mining perspective, the events data needs to be comprehended as follows:

Case ID: Subject ID is the case identifier in this dataset, which essentially identifies the person who is involved in the cervical cancer screening process. As each case is identified according to a person, it means that a case refers to a person's entire birth-to-death period and includes all activities related to the individual's healthcare history. However, it is worth to mention that the given dataset only covers 2012-2019 period. It thus rather refers to all activities that happened since a person registered in the system within the given period.

Activity: Cohort ID and cohort name can be interchangeably used as an activity identifier. However, in practice, activities should be easily read and understood by a human. Therefore, cohort ID is ignored, and cohort name is used for identifying the activity. In total, the dataset contained 33 different activities – all related to cancer screening. A full list of activities is given in Appendix II. There is an extra activity, "Female patients", which is artificially

created for the purpose of demonstrating that a person has registered at the clinic for the first time.

Timestamp: Cohort start date describes when the activity was recorded in the data. Note that this is the approximation of when the actual event took place and contain date information without timestamps – both limitations typical to any healthcare data. The order of the events occurring on the same day is given as it is in the original data.

Attribute: Age which indicates the age of a person in the process, is an event attribute. The main reason why it is an event attribute rather than a case attribute is that age of a person can be changed during the process. Therefore, it is obvious that there may be cases that have different ages in the time when different activities are registered.

In total, the data contained activities of 66,585 patients.

4.3 Data analysis

This section outlines the analysis process which is conducted during the research. The analysis is divided into two parts based on the research questions (see Figure 5).

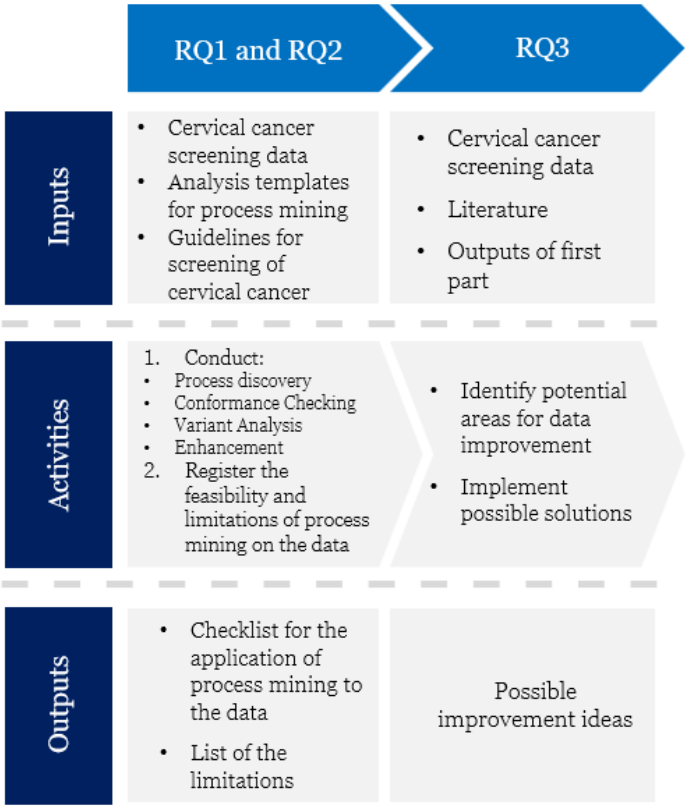


Figure 5. Parts of data analysis

In essence, all parts of data analysis are mainly conducted on a process mining tool called Apromore[1]. Data sensitivity is one of the main concerns when working with healthcare data. One of the requirements to conduct an analysis on the given dataset was to analyze the data within the premises of the University of Tartu (UT). Thus, Apromore was chosen as it has

[1] https://apromore.com/about-us/

provided an opportunity to access the tool within the premises and a full version was provided for academic purposes. The Apromore deployment which was used for this thesis was hosted by UT. Additionally, Apromore is considered as "Leader" in the latest report by Gartner[2].

The first part of the data analysis addresses the first and second research questions, which are related to the applicability of process mining use cases on Estonian healthcare data and corresponding limitations. There are three main requirements in order to be able to finish this part of the analysis: data, which is related to cervical cancer screening in the thesis, analysis templates for each of the categories of process mining [36]–[38], guidelines for cervical cancer screening [33]. The headings for the templates are described according to the categories (see Appendix III). The templates demonstrate what needs to be done for each category and also it explains how it can be done using Apromore. It is quite convenient that the tool used in the templates is Apromore, therefore they are very useful for the analysis stage in this research.

Essentially, as the current literature also supports, the templates and the given event logs are necessary to conduct analysis for three out of four main categories of process mining. Conformance checking is analyzed by comparing the process based on the real data and the guidelines provided. All other categories of process mining are also conducted to identify the possible application and feasibility of process mining. Additionally, elicitation of the limitations is performed. Moreover, the list of applicable use cases and the limitations are registered in the first part of the analysis.

The second part is the analysis of the third research question, which is, in essence, performed on top of the output that are acquired during the first part. As the output helps to identify the issues in the data, it also helps to ideation of potential areas for improvements. Furthermore, the author implemented the possible solutions given the circumstances. Thus, the output of the second part is ideas for improving the data quality for better process mining performance on the given data.

In conclusion, the outputs of the analysis are a checklist for the feasibility of process mining to the given data, a list of limitations and possible improvement ideas as it is described on Figure 5. The clarification of all steps implemented in the analysis will be explained in section 5, which includes the main findings of the thesis.

4.4 Ethics approval

This work was approved by the Estonian Bioethics and Human Research Council (EBIN, no. 1.1-12/653).

[2] https://www.gartner.com/doc/reprints?id=1-2CZI8XWU&ct=230320&st

5 Results

This section summarizes the results based on the main findings of the analysis of research questions. It starts with main findings on feasibility and limitations of process mining on the given event logs, followed by improvement opportunities.

After the data is imported to Apromore, it is possible to see log and temporal statistics for the given events data (see Figure 6).

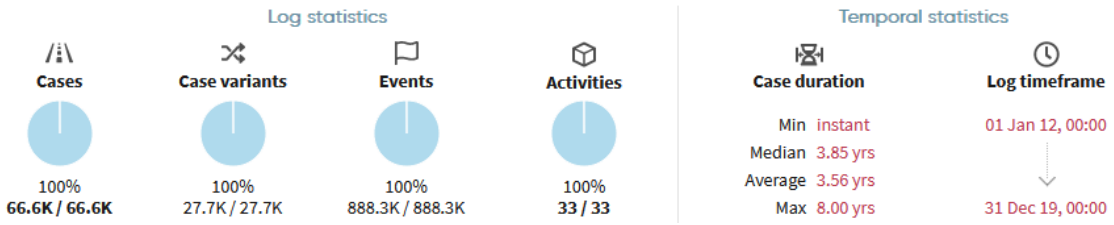


Figure 6. Log and temporal statistics

The number of case variants is 27.7 thousand while total number of cases is about 66.6 thousand. Temporal statistics shows minimum, median, average and maximum periods for the cases. The maximum period of the cases is 8 years which is in accordance with the data timeframe from 1st January of 2012 to 31st December of 2019. It can be considered normal that there would be some cases with instant duration, meaning that for all activities in a case the same timestamps were registered. However, after applying the filter to see the cases with only instant duration, it is revealed that the duration is instant for 34.9% of the cases. It may essentially indicate the problem with the registration dates of activities, which will be explained in the analysis of the first and second research questions.

5.1 Feasibility and limitations

RQ1 aimed at determining applicable process mining use cases within the context, while the main focus of RQ2 was to find out the main limitations. Following subsections address RQ1 and RQ2 while explaining the main steps of the analysis and results.

5.1.1 Automated Process Discovery

According to the analysis template for Automated Process Discovery [36], there are four parts that need to be executed: the analysis of flow, filtered flow, frequency and handoff analysis.

Flow analysis is consisted of analyzing the process structure, main case variants, case entry and exit points, and identifying parallelism, branching points as well as rework loops [36]. From the clinical pathway perspective, diagnosing the most frequent case variants is valuable. However, because of the complexity of clinical procedures, a case differs from other cases in one way or another (for example, see Figure 7).

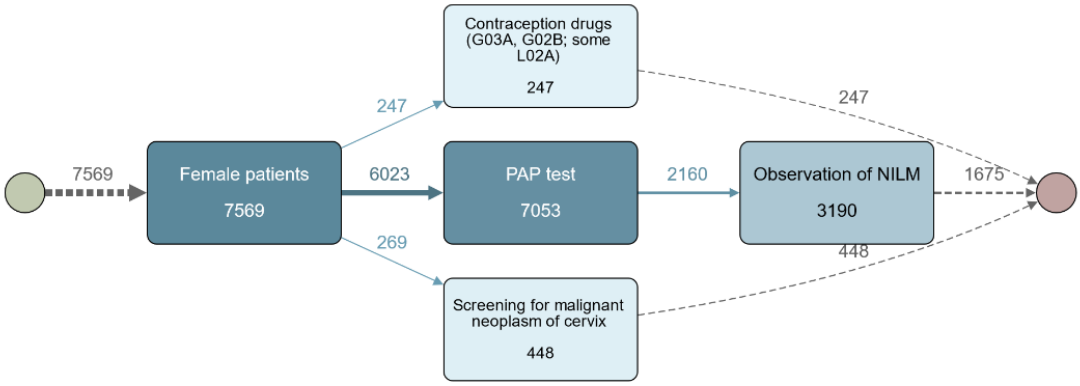


Figure 7. Top 10 most frequent case variants

In the given events log of cervical cancer screening, the top 10 most frequent case variants cover only about 11.4% of total (see Figure 7). Ignoring artificially created "Female patients" activity, PAP test seems to appear in the most frequent case variants, and also "Observation of NILM" also appears frequently. Furthermore, it implies that the results are negative, which means no cancerous changes, for a considerably high number of PAP tests. By using abstraction slider by "Case frequency" and "Average Duration", it is also possible to remove the nodes with low case frequency or average duration, which essentially enables simplifying the process map and see valuable details at first.

In essence, **filtered flow analysis** is the investigation of different components of the process separately. In the process discoverer, it is possible to retain or remove subsets of activities using event filtering. As there is information regarding the age of the patients in the given cervical cancer screening dataset, filtering the events based on the age is possible and useful. Also because there is a differentiation of guidelines according to the age of patients. In addition to filtering cases according to attributes (see Figure 8), the process discoverer also allows to see the frequency of the events with selected attributes.

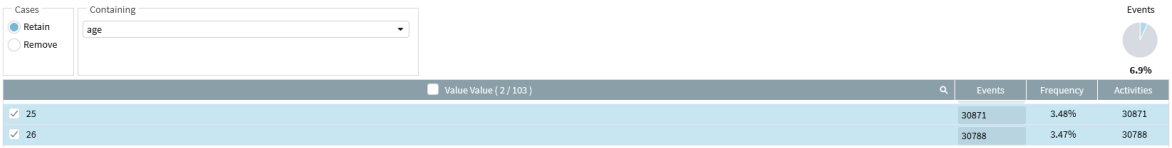


Figure 8. Filtering cases based on event attribute: age

For example, a doctor, physician, or expert in the field who research the cases can use this filtering to analyze particular age groups.

Frequency analysis can provide a visual color-coded representation for finding the most frequent arcs. There is also a possibility to switch between metrics of frequency such as minimum, average, maximum and others. In the given data, the most frequent activities are related to prescriptions regarding contraception (see Figure 9). Also, it reveals that "Observation of NILM" is one of the most frequent observations in the process.

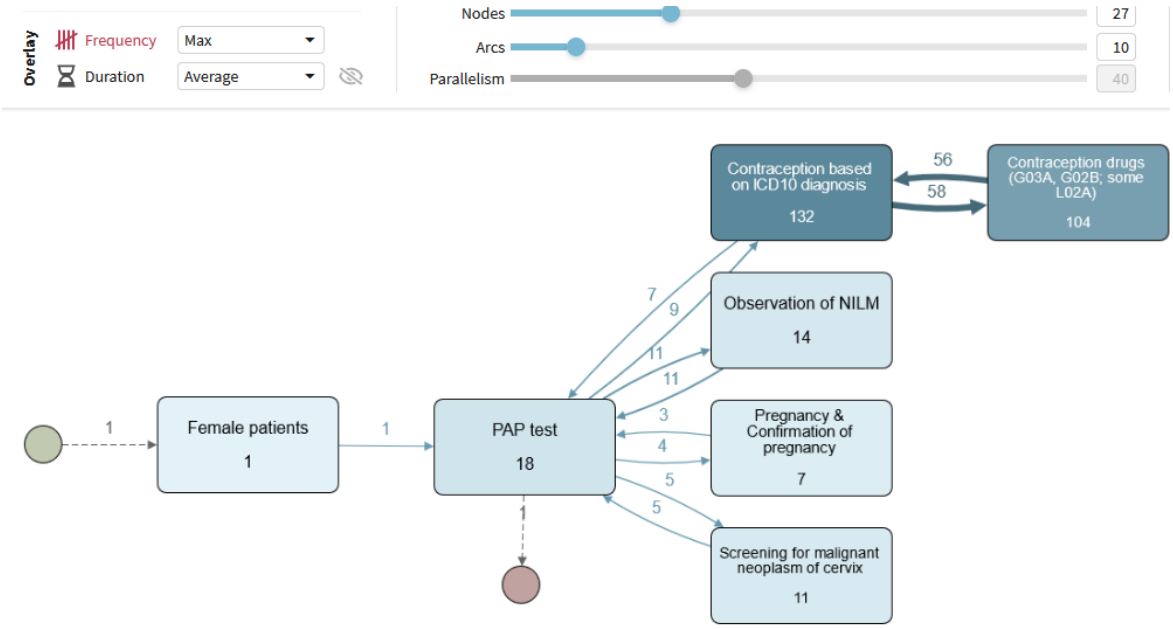


Figure 9. Representation of most frequent activities

Similarly, in the process discoverer, it is possible to list the activities according to their frequency and select the highest scores. For example, from Figure 9 one can observe the most frequent activities, evaluate the results and judge the cervical cancer screening process in general. That can help to identify the root-causes for delays in the screening process.

Handoff analysis is a useful way to analyze the efficiency of the process. It enables to investigate the resources such as workers or teams. Delays between activities can be identified with the handoff analysis. From managerial point of view in clinical procedures, it would be quite important aspect to cover and analyze. When handoff analysis is combined with filters, it becomes more insightful. However, as the given data does not include the resources such as doctors and physicians, it is not possible to conduct the analysis.

Consequently, initial three parts of process discovery, which are mentioned above, are practically possible to be applied to Estonian healthcare data. However, the outputs should not be considered as certain results. The main reasons are the lack of timestamps as well as start and end dates in the given event logs. These limitations can cause the order of activities in the process to be demonstrated incorrectly. Additionally, obtaining only either start or end date causes an issue to analyze the duration of activities. Therefore, the durations for all activities are identified as instant (see Figure 10).

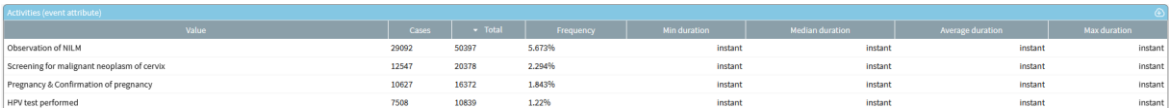


Figure 10. Duration of activities

Furthermore, in clinical practice, registration of events is usually done when a set of activities are done. For example, a patient visits a doctor and then gets a test. After that, according the test results, pathway for a patient can differ. However, these activities are not usually registered on time. In practice, doctors or physicians usually register the set of

activities when the related activities are done. This it can cause a misleading information regarding the flow of activities (see Figure 11).

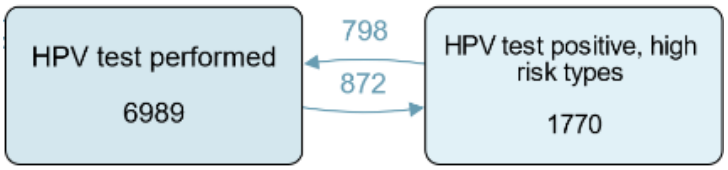


Figure 11. Anomaly in the flow of activities (more detailed in Appendix IV)

In Figure 11, with the above 25 years old age filter, "HPV test performed" activity is followed by "HPV test positive, high risk types", which is a normal flow in cervical cancer screening. However, the otherwise flow should not be that high, because after receiving positive test results, the subsequent activity should usually be "Colposcopy". It could be arguable that the clinical pathway is not very consistent with the guidelines, but the lack of timestamps is the main reason of the issue in this context. Given the frequency of this pattern in the process model of the given data, it is evident that the lack of timestamps is a significant issue.

5.1.2 Conformance checking

Essentially, conformance checking involves comparing a model based on real event logs with an expected model or guidelines. Moreover, it is not necessary to use a model to compare against the event log. Rules can be defined according to guidelines, which helps to make the comparison and describe the differences. Each and every difference is capable of making some impact on the process. Thus, identifying the differences enables to measure the success of clinical guidelines as well as exceptions. Additionally, conformance checking is also possible by analyzing the behavior in the log. Based on the analysis, unfitting behaviors in clinical pathways can be found, and the reasons can be investigated. As the only resource about clinical pathways is the guidelines and there is not manually designed process models for the pathways, only rules-based compliance and exceptional analysis are applicable in the healthcare settings.

According to the analysis template for Conformance checking [37], there are five parts which can be performed: flow, temporal, resource, model-to-log conformance checking and exception analysis.

Flow compliance checking, basically checking skipped mandatory activities, forbidden repetitions, and co-occurrence relations are feasible based on the template. Mandatory activities or forbidden repetitions are not directly applicable in healthcare settings, more explicitly in this cervical cancer screening event logs. Mainly because there are no mandatory or forbidden activities, considering that the data covers all events that may or may not be directly related to cervical cancer screening. But with the context of filtering based on other attributes such as age, they also become relevant.

The main analysis is done using a path filter in the process discoverer in Apromore. In the established guidelines [33], there are several procedures for different scenarios which need to be followed. For example, "Observation of ASC-H" activity should be directly followed

by “Colposcopy”, regardless of HPV status for women whose age is above 25 (see Appendix V). Path filter enables screening of cases where an activity either directly or eventually follows another specified activity (see Appendix VI). For getting better insight into the cases where the guidelines are being followed, firstly, it is helpful to filter only the cases where “Observation of ASC-H” has occurred. After the first filter, it is revealed that there are 524 cases where “Observation of ASC-H” has occurred (see Appendix VII).

Figure 12 represents all the cases where the established guidelines regarding mentioned activities have been followed.

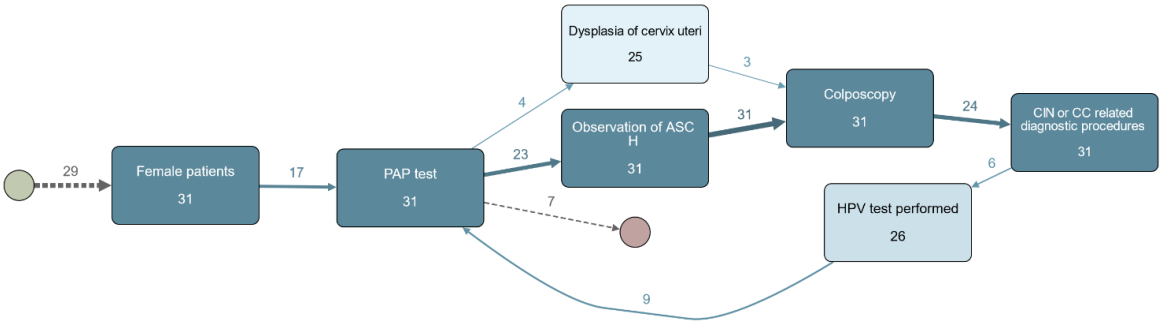


Figure 12. Cases where "Observation of ASC-H" has been directly followed by "Colposcopy"

Only 31 cases have followed the guidelines, which makes the compliance at about 5.9% (31 out of 524). However, the lack of timestamps can cause issues here too. Additionally, there may be activities which are not directly related to the procedure, however they appear in between two activities mentioned above. Activities regarding drug prescriptions can appear between these activities, and in practice, it is not an issue. Therefore, despite the fact that it is noted that “Colposcopy” has to be executed right after “Observation of ASC-H”, considering the circumstances of the given data and healthcare settings, it is also necessary to check the cases where it is executed not only directly after the observation, but also eventually. After using eventually-follows filtering (see Appendix VIII), the results are drastically changed into 75% (393 out of 524). Furthermore, by analyzing randomly selected specific cases, it is also revealed that registering a set of activities in one day also causes the poor results.

Temporal compliance checking is analyzing the conformance in terms of time-related aspects. Apromore enables to switch between overlays according to the frequency and duration of cases (see Figure 13). According to the guidelines established in between 2012-2019 [33], it was recommended to have a “PAP test” once a year to women with age from 21 to 29 years old. Figure 13 shows the median of durations for the cases of patients in the specified age group. The difference between the duration for repeating “PAP test” in the event log and the guidelines is not significant, which means that there is a compliance. However, it does not necessarily mean that the process is fully compliant with the guidelines. The main reason is that there are many different combinations of diagnoses, which result different recommendations.

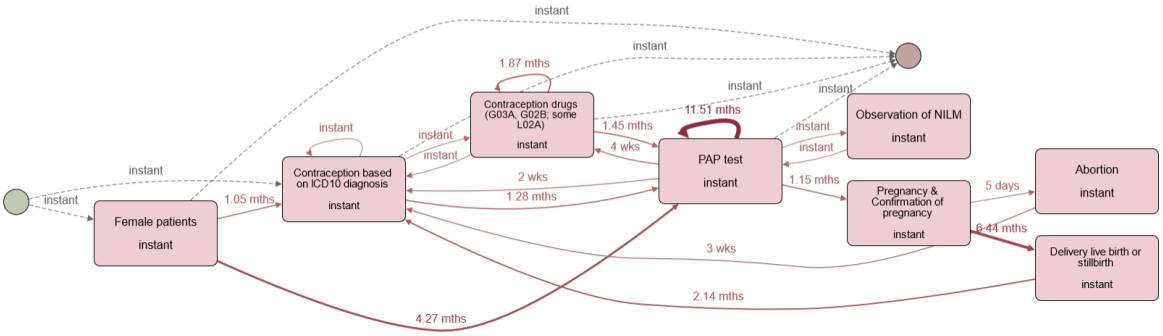


Figure 13. Duration overlay of the process for patients whose age is between 21 and 29

Temporal compliance checking is basically conducted based on the constraints which are related to time. Performance filter can be useful for identifying service level agreement violations, but it is not exactly needed in this particular healthcare settings. The total duration of cases is depended on many different aspects. The process can be stopped when a patient moves to another country, and it can also be related to some activities. Some surgical procedures can potentially end the process for a specific patient.

The other temporal compliance possibility is to use path filter with duration constraints, which is quite common to have in cervical cancer screening. For example, for women aged 21-24 with “Observation of ASC-H” and “Observation of HSIL”, it is necessary to repeat “PAP test” at 6-month intervals within 24 months. Temporal compliance checking enables to check the particular cases with constraints regarding time.

Resource compliance checking is crucial from a managerial or organizational point of view, considering how valuable is to find out the proper usage of resources such as doctors and physicians. In essence, it also has an undeniable impact on a clinical pathway. Unfortunately, in the given data of cervical cancer screening, resources for activities were not represented for privacy reason. Therefore, it is not possible to apply resource compliance checking to the event logs used in this research. It is worth to mention that even without specific resources of activities, the top level of resources can also be useful. Considering that the data is at national level, the top level of resources are hospitals, clinics and pharmacies.

Model-to-log conformance checking is not applicable as much as rules, guidelines and procedures in the healthcare settings, as it is already mentioned. There is usually not any established process model for diseases or any clinical pathways, which does not enable to analyze this use case. Therefore, it is better to use possible constraints to filter what is relevant and get insights about the process from event logs. Consequently, the result for this use case is that it is not feasible.

Exception analysis is analyzing the exceptions, which are highly infrequent behaviors that occur in the process. Abstraction slider in the process discoverer allows to invert the order of frequent cases, meaning that it enables to view the cases with higher infrequencies. Exception analysis can be useful to some extent in the cervical cancer screening (see Appendix IX). The analysis allows to identify the infrequent behaviors, their reasons and

impacts on the process. Figure 14 represents the number of variants identified with common and exceptional behavior.

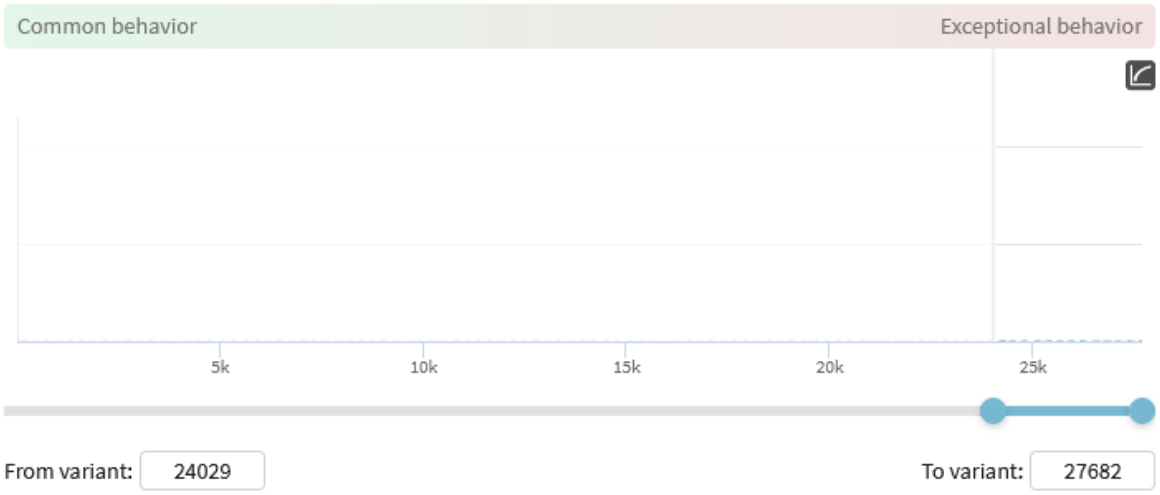


Figure 14. Case variant filtering based on behavior

The most interesting part is that all exceptional behavior case variants include only one case. It can already be understood from the number of total case variants. However, the results justify the complexity of clinical pathways and their variations from the standard procedures.

5.1.3 Variant analysis

For analyzing the variants, it is recommended to visualize two or more event logs of the same process. The variants can be filtered based on performance, attributes or timeframes. According to analysis templates [38], flow, frequency, rework, and bottleneck comparison are possible to be performed to analyze the variants. The main reason why variant analysis is considered as another category rather than being a part of exception analysis is the idea of comparing the variants side-by-side visually. Moreover, it relies on visual comparison of different variants at the same time. However, as it is already mentioned that the complexity of clinical pathways creates an enormous number of variants, which are different cases at core.

Flow comparison starts with discovering a process model from an event log for each variant. First, it is necessary to filter the event log based on temporal, logical or performance constraints, and then it is possible to save the filtered log. Discovering a process model can be performed after saving two or more event logs. For example, one way to conduct variant analysis in the given event log, comparing the variant of cases where “Abortion” activity appears with the cases where it does not appear (see Appendix X). The analysis may reveal that different procedures depend on whether “Abortion” has occurred or not. However, it requires more domain-specific knowledge to compare the variants and derive valuable insights from the comparison. Furthermore, it is good practice of visualizing the processes which are important to analyze from clinical point of view.

Frequency comparison is quite similar to flow comparison in terms of steps which need to be done. The main difference is the choice of overlay. Frequency overlay is recommended

for this analysis, while duration overlay is recommended for flow comparison. It is basically the only difference between these comparison variant analyses. Frequency overlay enables to detect the most frequent transitions [38] (see Appendix XI). Nevertheless, the **rework comparison** is not theoretically applicable to the given clinical screening process. Identifying rework in process mining is crucial to improve the efficiency of the process, which would also help to improve the process in healthcare settings from organizational point of view. In terms of clinical pathways, it is not possible to recognize the rework, because the activity can possibly be performed several times based on different factors such as doctor's intuition and a patient's correlated diseases. Moreover, due to the nature of the clinical processes, many of the activities are inherently repeatable.

Bottleneck comparison can be conducted from two perspectives to identify the bottlenecks in the process: resource and activity. For resource perspective, there should be resources which are assigned to each activity in the dataset. However, it is not available in the given event logs. On the other hand, for activity perspective, there should be start and end dates or preferable timestamps of each activity. Unfortunately, it is not available either in the obtained event logs. Therefore, bottleneck comparison is not applicable at all to the healthcare data retrieved.

5.1.4 Enhancement (Performance mining)

According to the analysis templates [38], the analyses which can be performed for enhancement are: bottleneck, workload, demand, rework and over-processing analyses.

The problems related to bottleneck and rework analysis are already mentioned in the previous paragraphs. For variant analysis, comparison of bottlenecks and reworks are necessary. Moreover, there is no possibility of comparison without analyzing items. Therefore, **bottleneck** and **rework** analysis are not also applicable in this context.

Once more, **workload analysis,** an analysis of resources is vital in process mining in order to get better overview of performance of processes. Thus, process mining tools possess different capabilities to investigate resources by identifying workload. Keeping in mind that there is no available data regarding resources, workload analysis is not applicable either.

On the other hand, **demand analysis** is conducted based on the information regarding active cases over time. In short, it is necessary to detect periods of time which are high in demand. Considering that obtained event logs cover the cases over the period between 2012 and 2019, it is not meaningful to analyze active cases between that timelines (see Appendix XII).

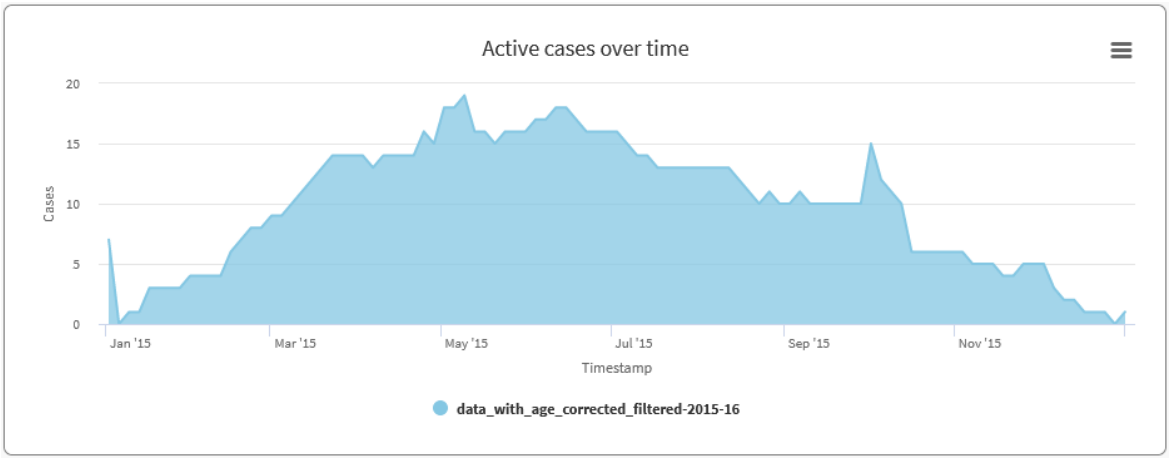


Figure 15. Active cases over time (period between January 2015 and 2016)

Therefore, additional filtering on timeframe is applied to get better results. Figure 15 represents the demand during 2015. It is possible to filter different timeframes to detect the months which usually have most active cases. It can help to allocate the resources properly.

Over-processing can be considered as a waste of activities. Decision activities have an important impact on how the processes flow or proceed. They may have a knock-out effect which results in flow of one or more unnecessary activities [38]. It can potentially be applied to healthcare data. However, it requires more in-depth domain knowledge for the analysis. Determining whether a decision activity can be performed earlier or not is the key point of over-processing analysis. Medical experts may determine such activities. Although it may be possible to perform in practice, but a theoretical value of it is arguable.

5.2 Improvement opportunities

This subsection outlines the improvement ideas suggested for the limitations, which identified in the analysis of the second research question. Additionally, there is an implementation for solutions when it is possible.

The main issues in the healthcare data are time related problems, such as the lack of timestamps as well as start and end dates, timestamps for activities. In terms of start and end times for activities, one possible solution can be assigning start dates based on end dates. It requires a piece of in-depth knowledge of each activity. For example, based on the usual amount of time each activity takes, it would be possible to calculate start dates from end dates, by subtracting the specified duration or otherwise. Moreover, if there are event logs with both start and end dates for cervical cancer screening in Estonia, machine learning algorithms would train on those data, and provide the solution. However, it is out of scope of the thesis and there is no access to such data.

On the other hand, the issue regarding the lack of timestamps can be solved at least partially. The key point is to know usual order of the activities in the event logs. Despite the fact that it is not a certain solution for the problem, it improves the overall quality of almost all categories of process mining. During the analysis of first and second research questions, it

is already mentioned that the lack of timestamps can potentially affect the flow of activities. It is practically revealed that there is quite high possibility of that issue, knowing that an activity related to a test result cannot be happened before the relevant test is happened (see Appendix XIII). To solve the issue, the author identified activities related to test as well as test results (see Appendix XIV). The main idea is to add artificial timestamps to all activities.

Moreover, timestamps for activities related to tests should be assigned earlier than activities regarding test results. Therefore, timestamp for tests is set "00:05", and "00:10" is set up for test results. The timestamp for all other activities is "00:00" (see Appendix XV). There are also other activities, which are not either test or test results, and it would be useful to set timestamps for those activities as well. However, it requires domain-expertise. Thus, the solution is implemented for certain activities. For example, in Figure 16, it is possible to see that there is a difference between event logs when artificial timestamps are added to activities and without timestamps. The anomaly which is described in Figure 11 is persistent in the given data without timestamps (see Figure 16 a). After applying timestamps for activities related to test and their results, it is possible to see that the nodes coming from test results to the tests are not reflecting the truth. The main reason why there are nodes to both way between "Observation of ASCUS" and "PAP test" is that they are registered under the same day. Even though the exact day for each activity cannot be retrieved, adding artificial timestamps solves the issue. A solution as simple as that improves the quality of discovery process and gives better overview of the flow of activities in the process model.

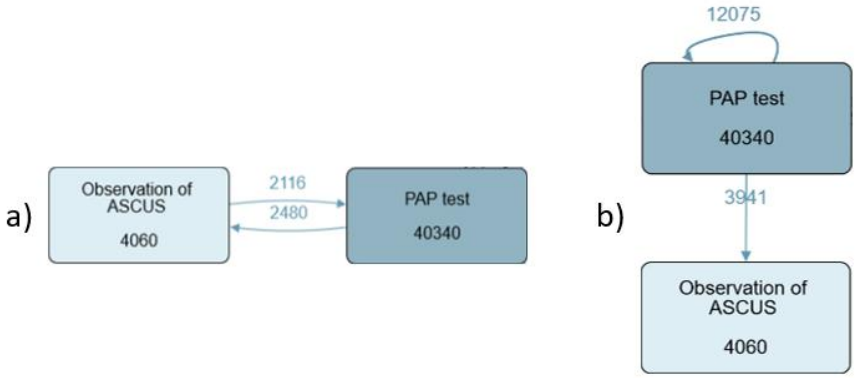


Figure 16. Difference between event logs a) without timestamps and b) with timestamps
*same abstraction level is used for retrieving both snapshots from the process model

There are also other issues and limitations in the given dataset and of process mining in general. The lack of resources data, the unclarity and complexity of clinical pathways are the other main examples of limitations. The improvement ideas for lack of resources can be that it is crucial to capture the data related to resources at activity level and it is also vital not to miss it out during the manipulation of several databases. Unfortunately, it is neither within the scope of the thesis nor possible within the given circumstances. For dealing with complexity of clinical pathways, Munoz-Gama et al. suggests to view the processes from a patient's point of view [39]. As a patient is the core of any healthcare process, it is necessary to approach each case individually when it is possible. Exceptions in the clinical pathways can give insightful overview about activities, which are not significantly visible at first glance.

6 Discussion

This section summarizes the results based on the main findings of the analysis of research questions and also contrasts them with the findings in the literature. Here, the discussion starts with the discovery of the relationship (see Table 1) between the findings of RQ1, which focuses on the feasibility of process mining, and RQ2, which investigates main limitations accordingly. It is followed by a reflection on possible improvement ideas. Finally, this section also discusses the limitations of the study.

6.1 Applicability of process mining use cases and limitations

During the analysis of the first research question, the feasibility of process mining use cases has been identified. The analysis is involved in four categories of process mining. Therefore, the main findings are also categorized under the specified types. All main findings are presented in tabular format in "Feasibility" column of Table 1, which shows the results.

Table 1. Summary of results for RQ1 and RQ2

PM Analysis	Feasibility	Main limitations
Automated Process Discovery		
Flow analysis	Yes	T, D
Filtered flow analysis	Yes	T, D
Frequency analysis	Partially	T, D
Handoff analysis	No	R
Conformance checking		
Flow compliance	Partially	C
Temporal compliance	Yes	T
Resource compliance	No	R
Model-to-log conformance	No	C
Exception analysis	Partially	C
Variant Analysis		
Flow comparison	Partially	C
Frequency & rework	Partially	T, C
Bottleneck comparison	No	C, D
Enhancement		
Bottleneck analysis	No	R, D
Workload & demand	Partially	R
Rework analysis	No	C
Over-processing	Partially	C

- T: Lack of timestamps
- D: Undefined start and end date
- R: Lack of resource information
- C: Complexity of clinical pathways

The feasibility of each use case is also evaluated into three categories which are quite self-explanatory. Basically, "Yes" means that the corresponding use case is feasible, while "No" means it is not feasible at all. The main difference between "Yes" and "Partially" is the level of issues caused by limitations. For example, if a use case has some limitations but its

feasibility recorded as “Yes”, it means that the results are reliable; however, there is a room for improvement. However, when the feasibility is recorded as “Partially” for a use case, it means that limitations make the results unreliable to some extent.

Overall, out of the 16 analysis types that can be performed using the templates, 10 use cases were found to be feasible at varying levels. Table 1 addresses **RQ1.** *What process mining use cases can be applied to Estonian healthcare data?*

Consequently, flow and filtered flow analysis as well as temporal compliance, can be easily applied to Estonian healthcare data. Additionally, per each category, there are also other use cases that are applicable. As it can be seen from the table, frequency analysis can be applied partially during process discovery. Flow compliance and exception analysis can provide valuable information in alignment with guidelines for conformance checking. Flow and frequency comparisons are also applicable to the given event logs. In terms of enhancement, only demand and over-processing analyses can be put into practice.

There are several studies investigated the applicability of process mining use cases [40], but no detailed findings have been found that have explicitly used the templates employed in this work within the healthcare domain. However, at top level, the analysis process is somehow aligned with the workflow [22] introduced by Toth et al. Furthermore, in most of the papers, process mining is divided into three categories. All of the categories are similar to the ones that are used in this research, however there is one additional type, variant analysis. Besides that, the categorization for implications of process mining are consistent with the most recent papers in the field [4]. According to the results of this thesis, the most applicable category is Automated Process Discovery, which provides valuable insights about the cervical cancer screening in Estonia. The only barrier for that category is resource information that is not available.

Main limitations have been identified during the analysis of the second research question. In most of the cases, the limitations either restricts the analysis fully or partially. There are a few cases where the impact of issues is not significant. All main limitations are compiled into “Main limitations column” of Table 1, which also addresses **RQ2.** *What are the limitations that restricts the implication of process mining to Estonian healthcare data?*

The first limitation is lack of timestamps, which is not as problematic as others. It occurs in flow, filtered flow and frequency analysis as well as temporal compliance. Frequency comparison, which is quite similar to frequency analysis, is also affected by this barrier. The main issue it causes is uncertainty in the flow of activities. The second obstacle, the undefined start and end date is also related to time. This limitation affects flow, filtered flow and frequency analysis. However, it is the key drawback of bottleneck-related analysis and comparison. Additionally, it is worth to mention that time information can also be sometimes incorrect. In the given event log, there have been cases where the test results appear to be before the test was taken. In addition to lack of timestamps, it is possible that a test is taken in one institution while it is analyzed in another one. Due to the differences between regular information processing procedures of different institutions, the time information is recorded incorrectly. For example, test results are recorded automatically in

the system, however information regarding when the test was taken needs be recorded by a doctor. If a doctor records the data later it may cause problems. And it is also evident that the doctor may not even record the exact date of taking test. Additionally, the event log reveals that it is quite common practice that doctors or physicians records the information related to particular health process of a patient at the end of the process. Therefore, the exact order of events are not always catched precisely. Augusto et al. [41] also encounters the same problem when applying process mining to vaccination patterns.

Lack of resource information has a moderately significant impact on the applicability of process mining within the context. Essentially, resource information is available in the Estonian healthcare data. However, it was not used in this work because of two main reasons: privacy of the institutions and cohort-based approach. The reason why the cohort-based approach creates an issue is that when several events from different institutions are combined into a single cohort, it is not possible to hold details of the institutions. Also, privacy of the institutions was a major concern. Considering the privacy reasons and the approach, which is used in this work, resource information was not accessible. Therefore, resource compliance, handoff, workload and bottleneck analysis were investigated in full potential. In contrast to this work, Park et al. [31] managed to analyze the resource-related use cases, as they had the department names for each activity. Rather than just institution about the information as a whole, the data they used had recordings of the department where the activity was performed.

Last but not least limitation is the complexity of clinical pathways, which has quite substantial effect on the implication of process mining to healthcare data obtained in Estonia. In essence, it almost completely restricts the applicability of model-to-log conformance, rework analysis and comparison. While these use cases are not entirely implausible, the extreme difficulties of conducting these analyses should not be underestimated. Furthermore, this barrier also limits implication of exception and over-processing analysis as well as flow compliance to some extent. The complexity of all healthcare procedures has been mentioned in many studies and some used medical expertise to ease this limitation to some degree [6], [22], [31], [39]. In addition to main limitations, there is also one aspect which needs to be considered. For applying process mining techniques, it is recommended to have clear start and end points, but screening process is ongoing throughout a patient's lifetime. Therefore, it is very difficult to identify what is the starting point of a particular process in medicine. Birth date of patient or first visit to doctor are the main possible identifiers in medical records. Thus, in this work, the starting point for a case is when a person is registered first time in institutions within the given time period.

Additionally, a recent work by Mooses [42] tried to investigate cervical cancer screening process in Estonian health dataset by using other analytical methods. This allows to compare the results from both works. The most of the main findings of that work, such as the percentage of women aged 21-65 had at least PAP or HPV tests ("PAP test", "HPV test performed" activities in the event log), are also revealed in this work (see Appendix XVI). The results of other analytical methods [42] were repeatable with the application of process mining to the same electronic healthcare data. Moreover, it was possible to reveal whether

official cancer screening guidelines were followed within the studied time period or not. However, it is necessary to keep the limitations in mind and therefore the slight differences between the results with process mining and other analytical methods. These differences (see Appendix XVII) can be minimized by improving the event log quality.

6.2 Improvement options

This subsection addresses **RQ3.** *What changes would be required to improve the feasibility of process mining on Estonian healthcare data?*

Table 2 summarizes the main findings related to improvement suggestions to overcome the limitations. Acquisition of relevant information precisely is necessary and recommended to minimize the effects of issues for all limitations.

Table 2. Main suggestions to overcome the limitations

Limitation	Suggestions
Lack of timestamps	• Adding artificial timestamps to activities based on the characteristics of the activity
Undefined start and end date	• Machine learning and AI solutions to predict start date based on the end date and learned activity duration
Lack of resource information	• Approaching each case individually • Recording specific information regarding resources (doctors, nurses, physicians and others)
Complexity of clinical pathways	• Multidisciplinary empirical studies

The key point to address this research question is to conduct comprehensive analysis on the limitations. In the given context, there are four limitations identified. First, the most convenient one to overcome is lack of timestamps, which is solved by adding artificial timestamps to each activity. The solution is implemented by assigning the activities to groups, which have logical order in terms of the flow in the process. For example, test result cannot be retrieved before the actual test itself. In comparison to solution used by Augusto et al. [41], the solution used in this work can be considered better in terms of providing precise order of activities. They have created a standard order based on the alphabetical order of the labels of activities. Additionally, they have also tried to create clinically meaningful activity order, which is improved by implementing logical clinical order into activities when it is possible in this work. Applying a logical order rather than an alphabetical order improved the quality of flow, filtered flow and frequency analysis mainly. Moreover, it has also impact on flow and frequency compliance for conformance checking. Considering the attempts to improve the data quality for process mining, there are also similar approaches in the literature. Park et al. has tried adding arbitrary timestamps to dates without timestamps in the relatively small event log for outpatient process [31]. There are also other papers which suggest the artificially created timestamps. But in general, it is not always possible to specify clinically logical order for all activities, because for some activities there are no required or logical order, meaning that they can be occurred anytime.

Solution for the second limitation can be varied depending on opportunities. It is important to note that any data related to the listed limitations should be acquired if possible. Thus, it is a common solution for at least three out of four limitations. However, with the emergence of machine learning algorithms and artificial intelligence solutions, it is possible to define start date when end date for an activity is defined clearly. This solution requires additional knowledge on each activity and its main properties. Nevertheless, it should be implementable in one way or another with correct amount of resources and capabilities. Moreover, the aforementioned solutions could potentially address the lack of resources information as well. However, as it is mentioned that a cohort-based approach and privacy reasons caused that limitation, it is possible to analyze the data without cohorts. But this approach is more reasonable for investigating individual cases separately, rather than analysis for overview. Also, the ultimate goal should be assigning specific resources such as a doctor, nurse, physician and others to events, instead of recording only the medical institution.

In terms of complexity in the nature of clinical pathways, it is not easily feasible to address the issue. However, reducing the impact of the barrier is achievable. A medical expertise within specific domains in the field can be considered as a main solution to this drawback. Multidisciplinary empirical solutions are not always easy to manage, but they are not impossible either. Considering the benefits can these solutions provide, it is worth to handle. Munoz-Gama et al. emphasizes that the focus should be on a patient [39], which is perfectly aligned with suggested medical expertise in this thesis.

6.3 Limitations

The main part of the thesis is quantitative data analysis on healthcare data using a process mining tool. As the cervical cancer-related event log is the key point of the analysis used for addressing the research questions, data quality is crucial for the validity of the results. One of the main limitations is that the data was prefiltered for cervical cancer screening data only. It means that all activities in the event logs were supposed to be related to the screening procedure. However, in real world, different healthcare processes happen simultaneously. It thus is challenging to distinguish the activities whether should or should not be considered as a related activity to the procedure. Therefore, there is a risk that other healthcare processes occurring simultaneously, which could potentially be related to cervical cancer screening, may have been excluded. The other limitation of this work was related to resource information as it is already discussed. The information was not accessed because of privacy reasons as well as the approach used in this work. It has limited the investigation of feasibility of process mining on Estonian healthcare data.

In terms of the generalizability of the results, it is arguable how expandable the study is outside of Estonia. The approach and main findings in this thesis can potentially be used to identify clinical pathways for other diseases or screening procedures in Estonia. However, considering the possible differences between healthcare data in other countries and Estonia, it thus is not certain how the main findings will be interpreted within the context of other countries without future research.

7 Conclusions

The aim of this thesis was to research the applicability of process mining techniques to aggregated electronic health records, which are obtained in Estonia. Alongside investigation of the feasibility, main limitations and improvement suggestions were also a part of the aim to explore. Three research questions were aimed to be elaborately analyzed and addressed in this study.

To investigate the applicability of process mining use cases, a real-world event log related to cervical cancer screening, along with guidelines regarding the disease and analysis templates for the main types of process mining, were utilized simultaneously. This approach allowed to address the question of which use cases are applicable within the context in a comprehensive manner. As a result, 10 out of 16 use cases were identified to be applicable to some extent. The main limitations, which were also in the aim, were an inseparable part of the analysis too. Based on the use cases, four main limitations were identified as following: lack of timestamps, undefined start and end dates, lack of information regarding resources, and complexity of clinical pathways. To overcome each limitation, possible improvement ideas were suggested. Only one of the suggested ideas was implementable given the circumstances and the scope. The implemented solution improved the overall quality of process models by artificially imputing necessary information.

In terms of future work, there are several opportunities. First, the identified applicable use cases can be used for discovering and analyzing clinical pathways of various diseases, screening and medical procedures, for example, event logs, which are obtained at the national level, for clinical pathway investigation. Second, suggested improvement ideas can be researched and attempted to be implemented to improve the overall quality of aggregated electronic health records acquired in Estonia.

References

[1] A. Hellström, S. Lifvergren, and J. Quist, “Process management in healthcare: investigating why it’s easier said than done,” *J. Manuf. Technol. Manag.*, vol. 21, no. 4, pp. 499–511, May 2010, doi: 10.1108/17410381011046607.

[2] I. Beerepoot, N. Martin, and J. Koorn, “From Insights to INTEL: Evaluating Process Mining Insights with Healthcare Professionals,” p. 10.

[3] W. van der Aalst *et al.*, “Process Mining Manifesto,” in *Business Process Management Workshops*, F. Daniel, K. Barkaoui, and S. Dustdar, Eds., in Lecture Notes in Business Information Processing. Berlin, Heidelberg: Springer, 2012, pp. 169–194. doi: 10.1007/978-3-642-28108-2_19.

[4] R. Gatta, S. Orini, and M. Vallati, “Process Mining in Healthcare: Challenges and Promising Directions,” in *Artificial Intelligence in Healthcare*, T. Chen, J. Carter, M. Mahmud, and A. S. Khuman, Eds., in Brain Informatics and Health. Singapore: Springer Nature Singapore, 2022, pp. 47–61. doi: 10.1007/978-981-19-5272-2_2.

[5] L. Martin, M. P. White, A. Hunt, M. Richardson, S. Pahl, and J. Burt, “Nature contact, nature connectedness and associations with health, wellbeing and pro-environmental behaviours,” *J. Environ. Psychol.*, vol. 68, p. 101389, Apr. 2020, doi: 10.1016/j.jenvp.2020.101389.

[6] M. R. Dallagassa, C. dos S. Garcia, E. E. Scalabrin, S. O. Ioshii, and D. R. Carvalho, “Opportunities and challenges for applying process mining in healthcare: a systematic mapping study,” *J. Ambient Intell. Humaniz. Comput.*, vol. 13, no. 1, 2022, doi: 10.1007/s12652-021-02894-7.

[7] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, “Process mining in healthcare: A literature review,” *J. Biomed. Inform.*, vol. 61, pp. 224–236, Jun. 2016, doi: 10.1016/j.jbi.2016.04.007.

[8] M. Hammer and J. Champy, *Reengineering the corporation: a manifesto for business revolution*. New York: HarperBusiness, 2001.

[9] M. Zairi, “Business process management: a boundaryless approach to modern competitiveness,” *Bus. Process Manag. J.*, vol. 3, no. 1, pp. 64–80, Apr. 1997, doi: 10.1108/14637159710161585.

[10] M. Dumas, M. La Rosa, J. Mendling, and H. A. Reijers, *Fundamentals of Business Process Management*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2018. doi: 10.1007/978-3-662-56509-4.

[11] W. van der Aalst, “Using Process Mining to Bridge the Gap between BI and BPM,” *Computer*, vol. 44, no. 12, pp. 77–80, Dec. 2011, doi: 10.1109/MC.2011.384.

[12] R. Pérez-Castillo, B. Weber, J. Pinggera, S. Zugal, I. G.-R. de Guzmán, and M. Piattini, “Generating event logs from non-process-aware systems enabling business process mining,” *Enterp. Inf. Syst.*, vol. 5, no. 3, pp. 301–335, Aug. 2011, doi: 10.1080/17517575.2011.587545.

[13] W. M. P. van der Aalst *et al.*, “Business process mining: An industrial application,” *Inf. Syst.*, vol. 32, no. 5, pp. 713–732, Jul. 2007, doi: 10.1016/j.is.2006.05.003.

[14] W. van der Aalst, “Process Mining: Overview and Opportunities,” *ACM Trans. Manag. Inf. Syst.*, vol. 3, no. 2, pp. 1–17, Jul. 2012, doi: 10.1145/2229156.2229157.
[15] M. Weske, *Business process management: concepts, languages, architectures*. Berlin ; New York: Springer, 2007.
[16] W. M. P. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-19345-3.
[17] L. Bos and International Council on Medical and Care Compunetics, Eds., *Medical and care compunetics 6*. in Studies in health technology and informatics, no. v. 156. Amsterdam ; Washington, DC: IOS Press, 2010.
[18] M. Oja *et al.*, “Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned,” Health Informatics, preprint, Feb. 2023. doi: 10.1101/2023.02.16.23285697.
[19] T. G. Erdogan and A. Tarhan, “Systematic Mapping of Process Mining Studies in Healthcare,” *IEEE Access*, vol. 6, pp. 24543–24567, 2018, doi: 10.1109/ACCESS.2018.2831244.
[20] R. S. Mans, M. H. Schonenberg, M. Song, W. M. P. van der Aalst, and P. J. M. Bakker, “Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital,” in *Biomedical Engineering Systems and Technologies*, A. Fred, J. Filipe, and H. Gamboa, Eds., in Communications in Computer and Information Science, vol. 25. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 425–438. doi: 10.1007/978-3-540-92219-3_32.
[21] L. Perimal-Lewis, D. Teubner, P. Hakendorf, and C. Horwood, “Application of process mining to assess the data quality of routinely collected time-based performance data sourced from electronic health records by validating process conformance,” *Health Informatics J.*, vol. 22, no. 4, pp. 1017–1029, Dec. 2016, doi: 10.1177/1460458215604348.
[22] K. Toth, K. Machalik, G. Fogarassy, and A. Vathy-Fogarassy, “Applicability of process mining in the exploration of healthcare sequences,” in *2017 IEEE 30th Neumann Colloquium (NC)*, Budapest, Hungary: IEEE, Nov. 2017, pp. 000151–000156. doi: 10.1109/NC.2017.8263273.
[23] S. Kundu, “AI in medicine must be explainable,” *Nat. Med.*, vol. 27, no. 8, pp. 1328–1328, Aug. 2021, doi: 10.1038/s41591-021-01461-z.
[24] the Precise4Q consortium, J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, “Explainability for artificial intelligence in healthcare: a multidisciplinary perspective,” *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, p. 310, Dec. 2020, doi: 10.1186/s12911-020-01332-6.
[25] G. P. Kusuma, A. P. Kurniati, E. Rojas, C. D. McInerney, C. P. Gale, and O. A. Johnson, “Process Mining of Disease Trajectories: A Literature Review,” in *Studies in Health Technology and Informatics*, J. Mantas, L. Stoicu-Tivadar, C. Chronaki, A. Hasman, P. Weber, P. Gallos, M. Crişan-Vida, E. Zoulias, and O. S. Chirila, Eds., IOS Press, 2021. doi: 10.3233/SHTI210200.

[26] F. Fox, V. R. Aggarwal, H. Whelton, and O. Johnson, “A Data Quality Framework for Process Mining of Electronic Health Record Data,” in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, New York, NY: IEEE, Jun. 2018, pp. 12–21. doi: 10.1109/ICHI.2018.00009.

[27] M. L. van Eck, X. Lu, S. J. J. Leemans, and W. M. P. van der Aalst, “PM^{2}: A Process Mining Project Methodology,” in *Advanced Information Systems Engineering*, J. Zdravkovic, M. Kirikova, and P. Johannesson, Eds., in Lecture Notes in Computer Science, vol. 9097. Cham: Springer International Publishing, 2015, pp. 297–313. doi: 10.1007/978-3-319-19069-3_19.

[28] M. Bozkaya, J. Gabriels, and J. M. van der Werf, “Process Diagnostics: A Method Based on Process Mining,” in *2009 International Conference on Information, Process, and Knowledge Management*, Cancun: IEEE, Feb. 2009, pp. 22–27. doi: 10.1109/eKNOW.2009.29.

[29] Á. Rebuge and D. R. Ferreira, “Business process analysis in healthcare environments: A methodology based on process mining,” *Inf. Syst.*, vol. 37, no. 2, pp. 99–116, Apr. 2012, doi: 10.1016/j.is.2011.01.003.

[30] Á. Vathy-Fogarassy, I. Vassányi, and I. Kósa, “Multi-level process mining methodology for exploring disease-specific care processes,” *J. Biomed. Inform.*, vol. 125, p. 103979, Jan. 2022, doi: 10.1016/j.jbi.2021.103979.

[31] K. Park *et al.*, “Exploring the potential of OMOP common data model for process mining in healthcare,” *PLOS ONE*, vol. 18, no. 1, p. e0279641, Jan. 2023, doi: 10.1371/journal.pone.0279641.

[32] Z. Valero-Ramon *et al.*, “Analytical exploratory tool for healthcare professionals to monitor cancer patients’ progress,” *Front. Oncol.*, vol. 12, p. 1043411, Jan. 2023, doi: 10.3389/fonc.2022.1043411.

[33] I. Aavik, L. Padrik, T. Raud, K. Täär, and P. Veerus, “Emakakaela, tupe ja vulva vähieelsete muutuste diagnoosimine, jälgimine ja ravi. HPV vastase vaktsineerimise soovitused. Eesti Naistearstide Seltsi ravijuhend. ‘Estonian Gynaecologists Society.’” [Online]. Available: https://www.ens.ee/

[34] S. Shami and J. Coombs, “Cervical cancer screening guidelines: An update,” *J. Am. Acad. Physician Assist.*, vol. 34, no. 9, pp. 21–24, Sep. 2021, doi: 10.1097/01.JAA.0000769656.60157.95.

[35] K. Künnapuu *et al.*, “Trajectories: a framework for detecting temporal clinical event sequences from health data standardized to the Observational Medical Outcomes Partnership (OMOP) Common Data Model,” *JAMIA Open*, vol. 5, no. 1, p. ooac021, Jan. 2022, doi: 10.1093/jamiaopen/ooac021.

[36] M. Dumas, “Business Process Mining Lecture 4: Automated Process Discovery,” [Online]. Available: https://courses.cs.ut.ee/LTAT.05.025/2021_spring/uploads/Main/Lecture4.pdf

[37] M. Dumas, “Business Process Mining Lecture 5: Conformance Checking,” [Online]. Available: https://courses.cs.ut.ee/LTAT.05.025/2021_spring/uploads/Main/Lecture5.pdf

[38] M. Dumas, “Business Process Mining Lecture 6: Process Performance Mining,” [Online]. Available: https://courses.cs.ut.ee/LTAT.05.025/2021_spring/uploads/Main/Lecture6.pdf

[39] J. Munoz-Gama *et al.*, “Process mining for healthcare: Characteristics and challenges,” *J. Biomed. Inform.*, vol. 127, p. 103994, Mar. 2022, doi: 10.1016/j.jbi.2022.103994.

[40] F. Milani, K. Lashkevich, F. M. Maggi, and C. Di Francescomarino, “Process Mining: A Guide for Practitioners,” in *Research Challenges in Information Science*, R. Guizzardi, J. Ralyté, and X. Franch, Eds., in Lecture Notes in Business Information Processing, vol. 446. Cham: Springer International Publishing, 2022, pp. 265–282. doi: 10.1007/978-3-031-05760-1_16.

[41] A. Augusto, T. Deitz, N. Faux, J.-A. Manski-Nankervis, and D. Capurro, “Process mining-driven analysis of COVID-19’s impact on vaccination patterns,” *J. Biomed. Inform.*, vol. 130, p. 104081, Jun. 2022, doi: 10.1016/j.jbi.2022.104081.

[42] K. Mooses, “Early Detection of Cervical Cancer and Monitoring in Estonia 2012-2019,” University of Tartu, 2022. [Online]. Available: https://comserv.cs.ut.ee/ati_thesis/datasheet.php?id=74841

Appendix

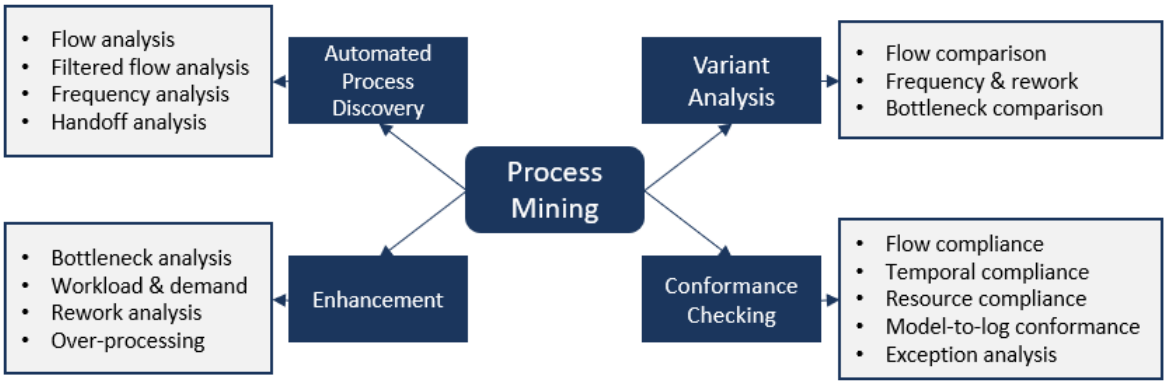
I. A snippet from acquired dataset

cohort_id	cohort_name	cohort_start_date	subject_id	age
623	Female patients	3/16/2012	11468	58
635	PAP test	1/8/2018	21691	59
603	Contraception based on ICD10 diagnosis	7/8/2012	44815	39
600	Pregnancy & Confirmation of pregnancy	3/1/2014	57416	29
631	HPV test negative	8/23/2017	40324	39
605	Contraception drugs (G03A, G02B; some L02A)	10/9/2017	50298	28

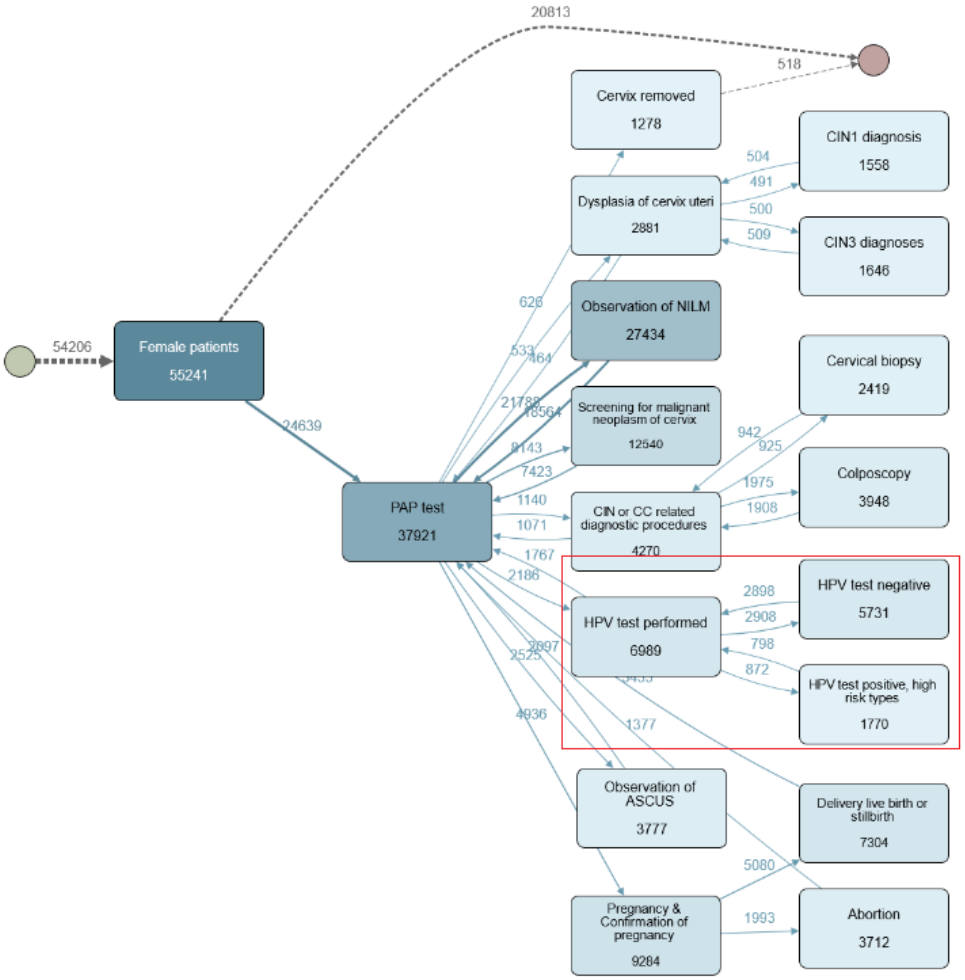
II. A list of all activities in the given dataset

Abortion
Cervical biopsy
Cervix removed
CIN or CC related diagnostic procedures
CIN1 diagnosis
CIN2 diagnosis
CIN3 diagnoses
Colposcopy
Cone biopsy of cervix
Contraception based on ICD10 diagnosis
Contraception drugs (G03A, G02B; some L02A)
Delivery - live birth or stillbirth
Destruction of lesion of cervix
Dysplasia of cervix uteri
Female patients
History of hysterectomy
HIV treatment diagnosis
HPV test negative
HPV test performed
HPV test positive, high risk types
HPV test positive, multiple high risk types
Malignant neoplasm of cervix
Observation of AGC-FN
Observation of AGC-NOS
Observation of ASC-H
Observation of ASCUS
Observation of HSIL
Observation of LSIL
Observation of NILM
PAP test
Pregnancy & Confirmation of pregnancy
Screening for malignant neoplasm of cervix
Surgical procedures on cervix

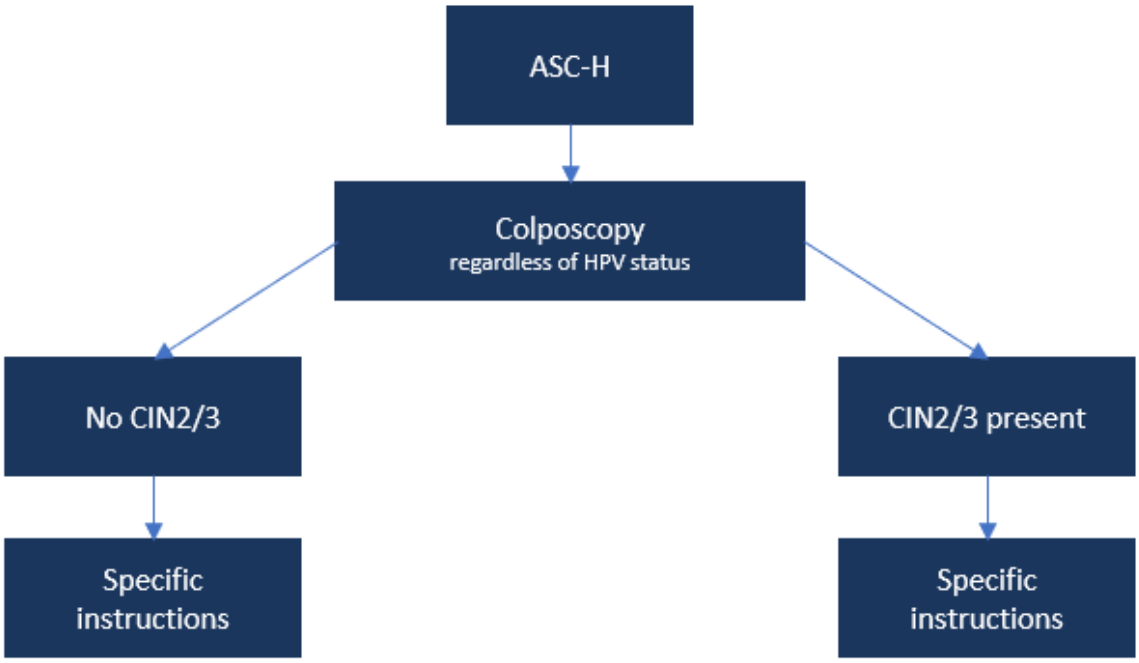
III. Process mining categories and analysis template headings



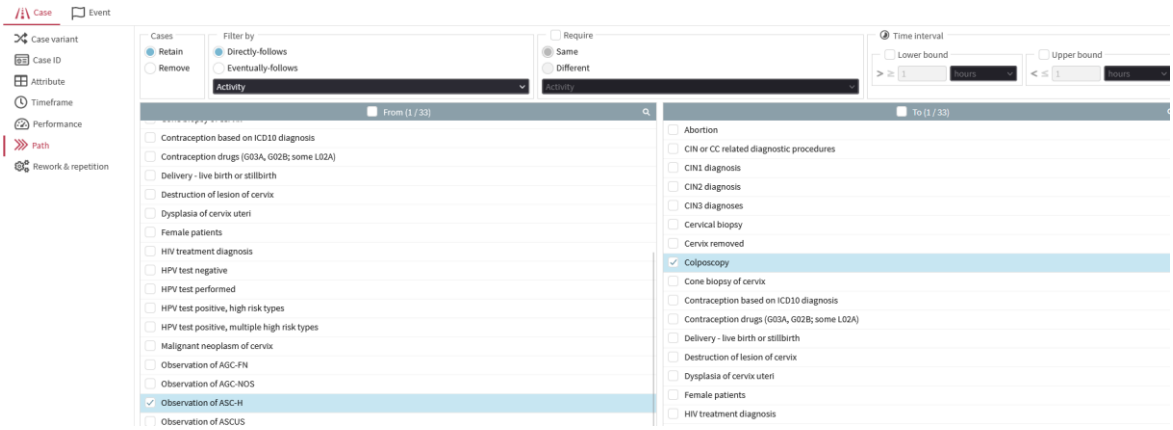
IV. Anomaly in the flow of activities



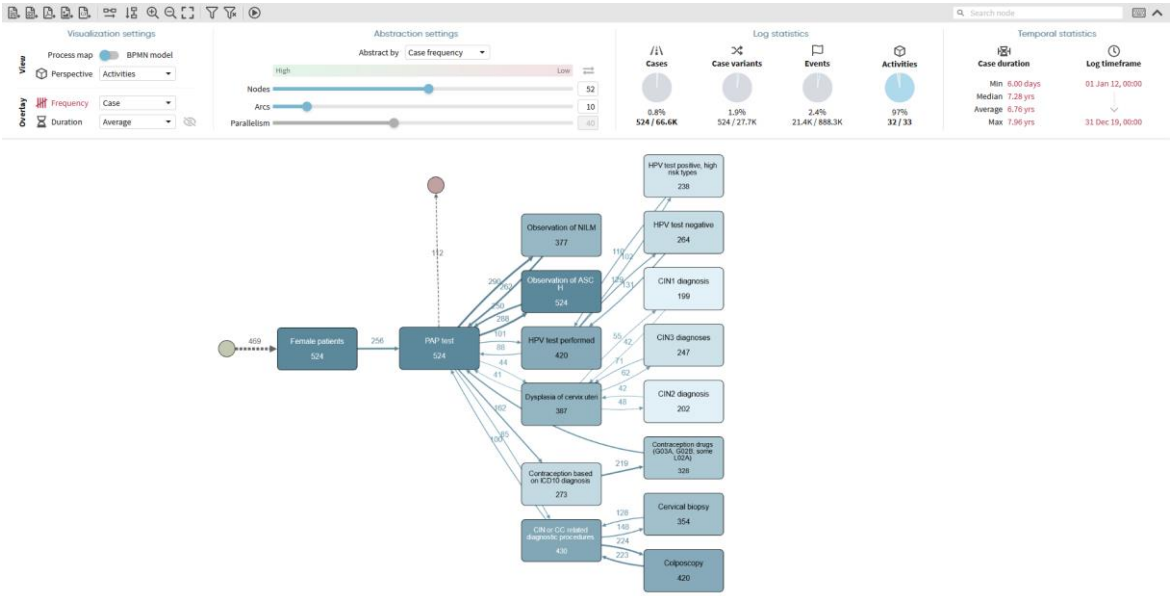
V. Follow-up for women whose age above 25, diagnosed ASC-H [33]



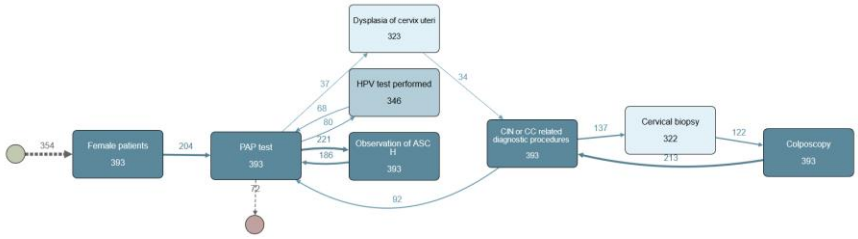
VI. Path filter for “Observation of ASC-H” directly followed by “Colposcopy”



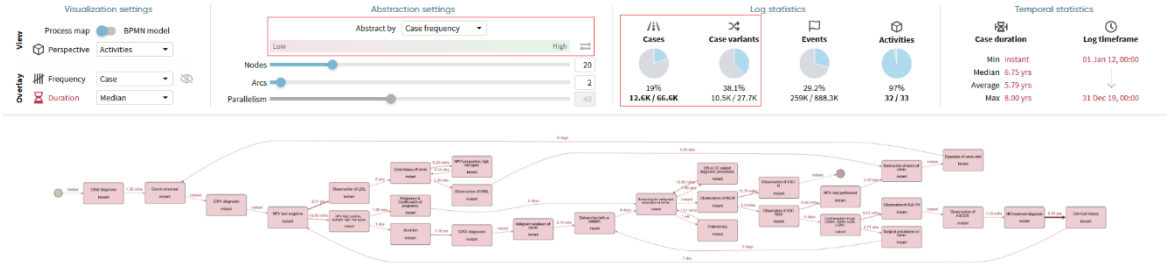
VII. 524 cases where “Observation of ASC-H” has occurred



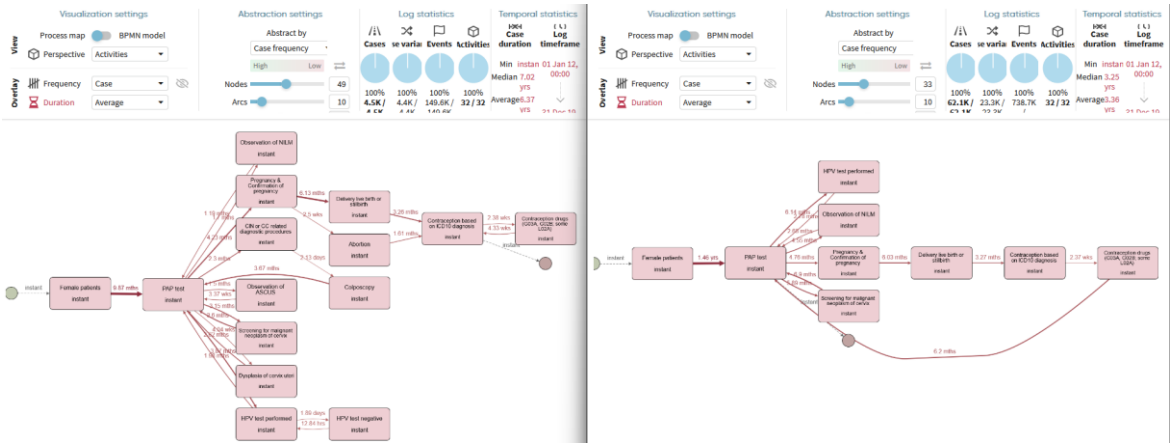
VIII. 393 cases where “Observation of ASC-H” eventually followed by “Colposcopy”



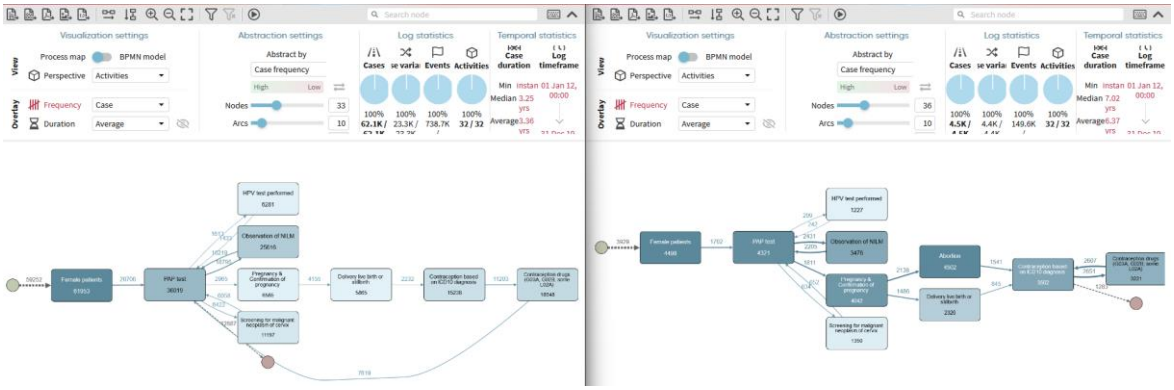
IX. The use of abstraction slider to identify highly infrequent cases



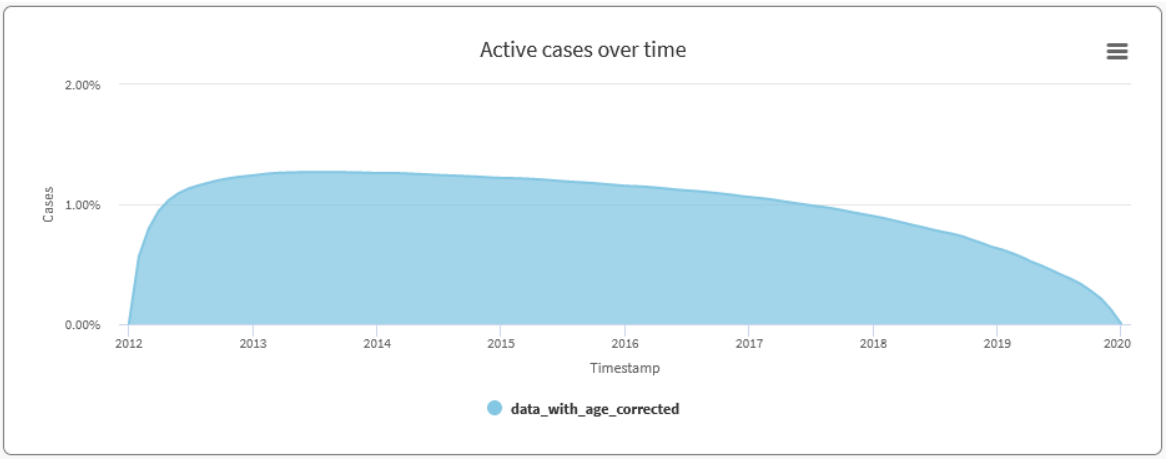
X. Visual comparison of two variants: “Abortion” retained and “Abortion” removed cases



XI. Visual comparison of two variants: “Abortion” retained and “Abortion” removed cases – Frequency overlay



XII. Demand analysis over all timeframe (2012-2019)



XIII. Impact of the lack of timestamps on the flow of activities

The most common events right after HSIL (out of 852)

Contraception based on ICD10 diagnosis	323
Contraception drugs (G03A, G02B; some L02A)	196
PAP test	102
Female patients	58
Observation of NILM	47
Screening for malignant neoplasm of cervix	23
Pregnancy & Confirmation of pregnancy	20
HPV test performed	14
HPV test negative	9
Colposcopy	8
Delivery - live birth or stillbirth	8
Cervical biopsy	5

The most common events right after ASCUS (out of 5030)

Contraception based on ICD10 diagnosis	1848
Contraception drugs (G03A, G02B; some L02A)	1225
PAP test	671
Female patients	355
Observation of NILM	284
Screening for malignant neoplasm of cervix	100
Pregnancy & Confirmation of pregnancy	82
HPV test performed	61
HPV test negative	53
Delivery - live birth or stillbirth	53
CIN or CC related diagnostic procedures	51

XIV. Python scripts for assigning activities to lists: “test” and “test results”

```
tests = [
"PAP test",
"HPV test",
"Colposcopy",
"Cervical biopsy",
"Cone biopsy of cervix",
"Destruction of lesion of cervix"
]
```

```
test_results = [
"HPV test positive, high risk types",
"HPV test positive, multiple high risk types",
"CIN1 diagnosis",
"CIN2 diagnosis",
"CIN3 diagnoses",
"Dysplasia of cervix uteri",
"Malignant neoplasm of cervix",
"Observation of ASCUS",
"Observation of LSIL",
"Observation of HSIL",
"Observation of AGC-NOS",
"Observation of AGC-FN",
"Observation of NILM"
]
```

XV. Python scripts for assigning timestamps for all activities

```
def add_timestamp(row):
    if row['cohort_name'] in tests:
        # adding 00:05 timestamps to activities which are related to tests
        return row['cohort_start_date'] + ' 00:05'
    elif row['cohort_name'] in test_results:
        # adding 00:10 timestamps to activities which are related to test results
        return row['cohort_start_date'] + ' 00:10'
    else:
        return row['cohort_start_date']
```

```
data['cohort_start_date'] = data.apply(add_timestamp, axis=1)
```

```
data['cohort_start_date'] = pd.to_datetime(data['cohort_start_date'], infer_datetime_format=True)

# adding "00:00" too all activities which are not either test or test results
data['cohort_start_date'] = data['cohort_start_date'].apply(lambda x: x.replace(hour=0, minute=0) if x.hour == 0 and x.minute ==
```

```
data['cohort_start_date'] = data['cohort_start_date'].dt.strftime("%Y-%m-%d %H:%M:%S")
```

XVI. Frequency of PAP and HPV tests -related activities

Female patients	44280	100.00%	44280
PAP test	33301	75.20%	103291
Observation of NILM	24999	56.46%	44655
Contraception drugs (G03A, G02B; some L02A)	18227	41.16%	186884
Contraception based on ICD10 diagnosis	16864	38.08%	299166
Screening for malignant neoplasm of cervix	12524	28.28%	20351
Pregnancy & Confirmation of pregnancy	10040	22.67%	15486
Delivery - live birth or stillbirth	7844	17.71%	10261
HPV test performed	7034	15.88%	10254

XVII. HPV tests of women aged 21-24 – 11% is reported on Mooses's work [42], while it is 15.39% in this work.

Value (1 / 32)	Cases	Frequency	Total
Female patients	3498	100.00%	3498
PAP test	2706	77.36%	7289
Contraception based on ICD10 diagnosis	2558	73.13%	54560
Contraception drugs (G03A, G02B; some L02A)	2442	69.81%	31494
Observation of NILM	1915	54.75%	3031
Pregnancy & Confirmation of pregnancy	1325	37.88%	2151
Delivery - live birth or stillbirth	982	28.07%	1311
Abortion	590	16.87%	885
HPV test performed	537	15.35%	766

8 License

Non-exclusive licence to reproduce the thesis and make the thesis public

I, **Musa Salamov**,

1. grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

Process mining on Estonian healthcare data

supervised by Fredrik Milani and Sulev Reisberg.

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Musa Salamov
09/05/2023