Doctoral Dissertation

A Data Analysis Methodology for Process Diagnosis and Redesign in Healthcare

Minsu Cho

Department of Management Engineering

Graduate School of UNIST

2018

A Data Analysis Methodology for Process Diagnosis and Redesign in Healthcare

Minsu Cho

Department of Management Engineering

Graduate School of UNIST

A Data Analysis Methodology for Process Diagnosis and Redesign in Healthcare

A dissertation
submitted to the Graduate School of UNIST
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Minsu Cho

06/21/2018 of submission
Approved by

Advisor
Marco Comuzzi

A Data Analysis Methodology for Process Diagnosis and Redesign in Healthcare

Minsu Cho

This certifies that the dissertation of Minsu Cho is approved.

06/21/2018 of submission

signature

Advisor: Marco Comuzzi

signature

Minseok Song: Thesis Committee Member #1

signature

Sooyoung Yoo: Thesis Committee Member #2

signature

Daeil Kwon: Thesis Committee Member #3

signature

Chiehyeon Lim: Thesis Committee Member #4;
three signatures total in case of masters

Abstract

Despite the disruptive and continuous development of healthcare environments, it still faces numerous challenges. Many of these are connected to clinical processes within the healthcare environment, which can be resolved through process analysis. At the same time, through the digitalization of healthcare, information from the various stakeholders in hospitals can be collected and stored in hospital information systems. On the basis of this stored data, evidence-based healthcare is possible, and this data-driven approach has become key to resolving medical issues. However, a more systematic data analysis methodology that covers the diagnosis and the redesign of clinical processes is required.

Process mining, which aims to derive knowledgeable process-related insights from event logs, is a promising data-driven approach that is commonly used to address the challenges in healthcare. In other words, process mining has become a way to improve business process management in healthcare. For this reason, there have been numerous studies on clinical process analysis using process mining. However, these have mainly focused on investigating challenges facing clinical processes and have not reached a virtuous cycle until process improvement. Thus, a comprehensive data analysis framework for process diagnosis and redesign in healthcare is still required.

We identify three challenges in this research: 1) a lack of guidelines for data analysis to help understand clinical processes, 2) the research gap between clinical data analysis and process redesign in healthcare, and 3) a lack of accuracy and reliability in redesign assessment in healthcare.

Based on these problem statements, this doctoral dissertation focuses on a comprehensive data analysis methodology for process diagnosis and redesign in healthcare. In particular, three frameworks are established to address important research issues in healthcare: 1) a framework for diagnosing clinical processes for outpatients, inpatients, and clinical pathways, 2) a framework for redesigning clinical processes with a simulation-based approach, and 3) a framework for evaluating the effects of process redesign.

The proposed methodology has four steps: data preparation, data preprocessing, data analysis, and post-hoc analysis. The data preparation phase aims to extract data in a suitable format (i.e., event logs) for process mining data analysis. In this step, a method for obtaining clinical event logs from electronic health record data mapped using the common data model needs to be developed. To this end, we build an event log specification that can be used to derive event logs that consider the purpose, content, and scope of the data analysis desired by the user. After compiling the event logs, they are preprocessed to improve the accuracy and validity of the data analysis. The data analysis phase, which is the core component of the proposed methodology, consists of three components for process mining analysis: clinical process types, process mining

types, and clinical perspectives. In the last phase, we interpret the results obtained from the data analysis with domain experts and perform a post-hoc analysis to improve clinical processes using simulations and to evaluate the previous data analysis results.

For the first research issue, we propose a data analysis framework for three clinical process types: outpatients, inpatients, and clinical pathways. For each category, we provide a specific goal and include suitable fine-grained techniques in the framework which are either newly developed or based on existing approaches. We also provide four real-life case studies to validate the usefulness of this approach.

For the second research issue, we develop a data-driven framework in order to build a discrete event simulation model. The proposed framework consists of four steps: data preparation and preprocessing, data analysis, post-hoc analysis, and further analysis. Here, we propose a mechanism for obtaining simulation parameters from process mining analysis from a control flow and performance perspective and automatically build a reliable and robust simulation model based on these parameters. This model includes realistic arrival rates and service times in a clinical setting. The proposed framework is constructed with a specific goal in mind (e.g., a decrease in waiting times), and the applicability of the framework is validated with a case study.

For the final research issue, we develop a framework for evaluating the effects of process redesign. Two types of indicators are used for this: best practice implementation indicators to assess whether a specific best practice has been applied well or not and process performance indicators to understand the impact of the application of best practices. These indicators are explicitly connected to process mining functionalities. In other words, we provide a comprehensive method for assessing these indicators using clinical event logs. The usefulness of the methodology is demonstrated with real-life logs before and after a redesign.

Compared to other existing frameworks in healthcare, this research is unique in constructing a healthcare-oriented data analysis methodology, rather than a generic model, that covers redesign in addition to diagnosis and in providing concrete analysis methods and data. As such, it is believed that this research will act as a motivation to extend the use of process mining in healthcare and will serve as a practical guideline for analyzing and improving clinical processes for non-experts.

Contents

List of Figures

List of Tables

I Introduction

This doctoral dissertation proposes a comprehensive data analysis methodology for process diagnosis and redesign in a medical setting. This chapter provides an introduction to this dissertation. Chapter 1.1 describes the background to our research, while Chapter 1.2 establishes the problems that are to be addressed. In Chapter 1.3, we introduce the objectives of this work and the scope of the research. Finally, Chapter 1.4 outlines the structure of the dissertation.

1.1 Research Background

The plethora of technical developments in medical environments, such as genomics, stem cells, new drugs, and innovative medical devices, have encouraged a number of stakeholders, including the government, care sites, clinical experts, and patients, to develop a keen interest in healthcare. As a result of this, residents of OECD countries have an average life expectancy at birth that exceeds 80 years and their living conditions have improved dramatically due to healthier lifestyles and universal public health coverage [5]. Furthermore, care quality has also significantly improved with the proactive detection of disease and the availability of more effective treatments [5].

Despite the continuous development of the healthcare system, a number of challenges remain [6]. Typical examples include the increase in visit occurrences due to the aging population, the increase in healthcare costs, the shortage of staff, and the increase in receivables [5, 7]. Most of these problems are related to clinical processes within the healthcare environment and can thus be improved through process analysis and redesign. As such, business process management (BPM) in healthcare is of paramount importance; it can be utilized to design, analyze, implement, and improve clinical business processes by applying useful methods and techniques [8].

There is a growing opportunity to deal with healthcare challenges by using the abundance of data that is collected within medical systems [9]. Due to the digitalization of healthcare information, data for hospital stakeholders is able to be collected and stored in hospital information systems [1]. On the basis of this stored data, evidence-based healthcare [10] and data-driven approaches can be employed to resolve current medical issues.

Analyzing clinical processes using data-driven methods is essential but currently this strategy is not effectively employed in healthcare organizations [1]. It is true that most hospitals take full advantage of data-based clinical expert systems [11–13]. A typical example is clinical decision support systems (CDSS), which can be used for alerts and reminders, diagnostic assistance, prescription decision support, information retrieval, image recognition and interpretation, and therapy critiquing and planning [11]. They focus on supporting the decisions of care providers with data analytics in order to increase patient outcomes. Healthcare stakeholders also use business intelligence systems for healthcare decisions [14,15]. However, these systems focus primarily on monitoring predefined key performance indicators using clinical data and establishing methods for improvement. As a result, there is a lack of applicable systems and methods that can

be used to analyze clinical processes from a process perspective; thus, a comprehensive data analysis methodology in healthcare is still required.

In this respect, process mining, which is used to obtain process-related information from event logs that can then be used in the decision-making process, is a promising data-driven approach [3]. It can be used to bolster BPM in healthcare. Its primary advantage is that the clinical process itself is explored and analyzed from a holistic perspective, while conventional data mining techniques only focus on a specific problem within a process [3]. For this reason, there has been a large volume of research conducted on clinical process analysis using process mining [16].

An overview of process mining in healthcare developed by Mans et al. [1] is presented in Figure 1. It starts with hospital information systems, which record clinical behavior with the support of healthcare reference models. Afterward, the collected data is converted into clinical event logs based on the extract, transform, and load (ETL) process; these event logs are utilized in two types of process mining, process discovery, and replay. As a result, meaningful insights can be developed and unseen challenges can be investigated.

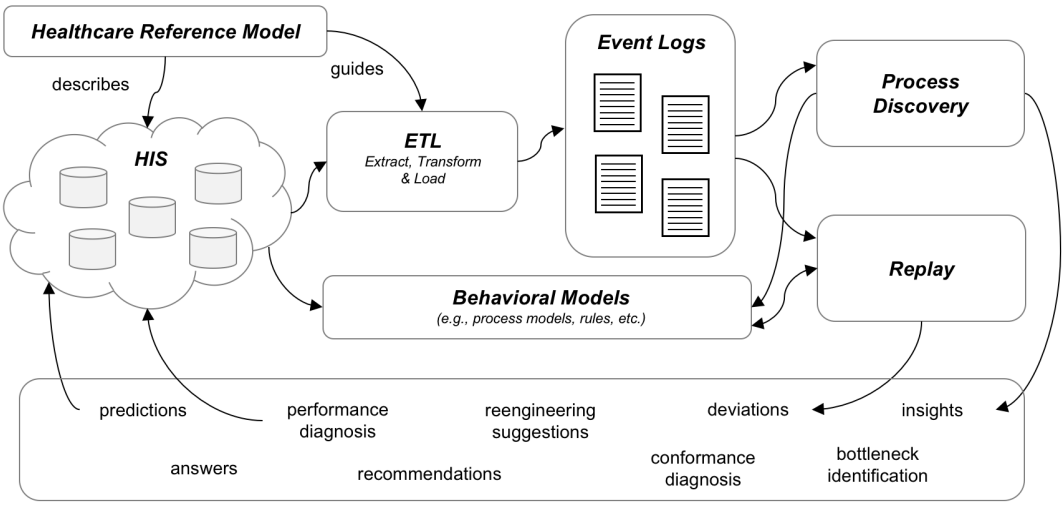


Figure 1. Process mining in healthcare [1]

This dissertation not only covers process mining in healthcare but also focuses on a comprehensive data analysis for process diagnosis and redesign. Here, process diagnosis involves understanding clinical processes from multiple perspectives, while process redesign represents a useful approach to improving clinical processes. For both concepts, data-driven process analysis using process mining is essential; thus, a data analysis methodology for process diagnosis and redesign in healthcare is proposed in this study.

1.2 Research Motivation and Problem Statement

This dissertation focuses on process mining for process diagnosis and redesign in healthcare, a topic of research that has been underserved to date. As previously mentioned, process mining in healthcare has been the subject of numerous studies, and there has been constant demand for clinical process analysis tools and methodologies that are applicable for practical use. However, there is a lack of comprehensive data analysis methodologies that utilize process mining, both academically and in practice. Existing works tend to either not be healthcare-oriented (e.g., healthcare reference models) or not include detailed algorithms and techniques; as a result, it is difficult to directly apply these methodologies in practice. Furthermore, most existing methodologies focus only on the diagnosis of clinical processes. However, investigating problems in clinical processes should be treated as a cyclical process that continues until the target process has improved and been evaluated. This means that an effective data analysis methodology should cover the following applications: understanding the existing clinical process, deriving re-design scenarios based on simulations, and evaluating the improved process.

In summary, the following research gaps have been identified, with the goal of addressing them in the present study.

1. **A lack of guidelines for data analysis directed at understanding clinical processes**

   As briefly introduced above, past research on data analysis from a process perspective in healthcare (i.e., applications of process mining in healthcare) is relatively common. This research includes newly-developed algorithms, such as discovery methods that produce clinical process models or analyze the ordering of clinical activities, and performance measurement methods that assess the length of hospital stays, the waiting or working time for specific activities, and the idle time for care providers. These studies have contributed greatly to the understanding of various medical processes. However, despite these technological advancements, there is a general lack of guidelines or frameworks for comprehensive clinical process analysis that can facilitate the practical use of these technologies for non-experts. In fact, past approaches have mainly been scenario-specific. In other words, the application of process mining techniques to clinical processes still operates on an ad-hoc basis. Users and researchers faced with process mining applications in practice struggle to find useful guidelines to follow in order to conduct their analysis, helping them to understand which data should be used and for what purpose or which process mining techniques are more useful for addressing various concerns typical of clinical process analysis.

2. **The research gap between clinical data analysis and process redesign in healthcare**

   Most process redesign approaches in the past have been manual or heuristic, leading to

difficulties predicting the effects of proposed improvements. Researchers and practitioners have employed discrete event simulation, which seeks to identify the most effective methods by predicting the expected effects of redesign options. However, this has a limitation in that it generally requires a great deal of time and effort to build an accurate simulation model because it is typically created by hand. While some previous methods have constructed a simulation model based on collected data, no concrete methods have been devised for the healthcare environment. For instance, clinical service time (i.e., an example simulation parameter) needs to be predicted using hospital information systems that only record the completion time for clinical activities. In summary, a sound method for building a simulation model for redesign options based on clinical data analysis in healthcare is still required.

3. **A lack of accuracy and reliability in redesign assessment in healthcare**

   The contextual method used to evaluate the effects of process redesign tends to simply compare the financial effects before and after a redesign. Qualitative evaluations of redesign outcomes, such as interviews or surveys with stakeholders, are also common, but these methods have a number of limitations. Most importantly, they require a certain amount of time to determine the effectiveness of the redesign, meaning the impacts of the improvement cannot be immediately identified. The qualitative approach may also lead to ambiguity about the reliability of the effects. To overcome these problems, some past studies have proposed a quantitative approach to measuring performance using process performance indicators. However, this approach only focuses on measuring the performance of specific processes and is not able to determine whether the results originate from the redesign or not.

1.3 Goals and Scope of Research

Based on the challenges presented in the previous section, this section explains the objectives and scope of the present research. The primary goal of the dissertation is to develop a data analysis methodology for process diagnosis and redesign using process mining in healthcare. In other words, the proposed methodology is comprehensively holistic, including data preparation, data preprocessing, data analysis, interpretation, redesign, and evaluation. On the basis of this proposed methodology, we establish three research frameworks in healthcare that can act as practical guidelines: 1) a framework for diagnosing clinical processes for outpatients, inpatients, and clinical pathways, 2) a framework for redesigning clinical processes with a simulation-based approach, and 3) a framework for evaluating the effects of process redesign. Figure 2 presents the goals and frameworks covered in this research.

The first research method is the development of a clinical process analysis framework using process mining, which covers data preparation, preprocessing, and analysis. In this framework,

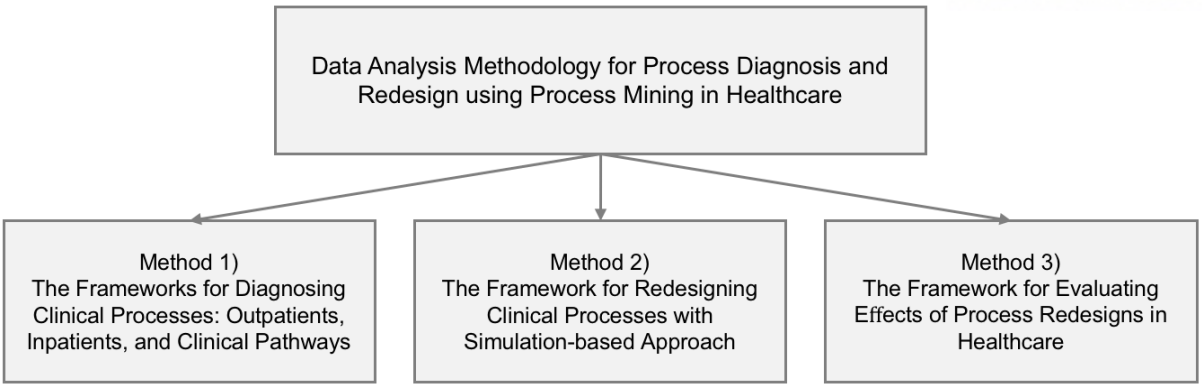


Figure 2. The goal and research methods of this dissertation

a series of specific procedures are delineated to produce valuable insights. To this end, the framework includes a homogenized event log specification template that helps to create a suitable format for analyzing data with the guidance of a clinical common data model [4]. It also has three orientations for clinical process analysis – outpatients, inpatients, and clinical pathways – which are engaged in clinical process types. We also propose frameworks for each analysis type, with the fine-grained techniques included in the framework either newly developed or based on existing approaches. The proposed framework is established based on a literature review and insights from case studies. It has a clear advantage in that it allows non-specialists to analyze data with ease and precision.

The second research method is the development of a redesign methodology for clinical processes based on process mining and discrete event simulation (DES). In the proposed method, a DES model is semi-automatically constructed with simulation parameters derived from process mining analysis (e.g., process discovery, patient arrival rate, and service time). Here, we develop a new method to calculate both arrival rates and service times considering the specific characteristics of hospitals. As such, it greatly simplifies the application of DES to clinical settings. This framework also proposes a series of steps for identifying an optimal process model from the simulation analysis. This allows the contextual rule-of-thumb approach to be avoided and reduces computing time and resources due to the efficiency of simulation analysis.

The final research method is the development of a structured assessment methodology for clinical process redesign. This approach proposes two types of indicator for evidence-based quantitative assessment based on redesign best practices defined in the existing research [17]: *best practice implementation indicators* (BPIs) and *process performance indicators* (PPIs). BPIs for each best practice are defined to determine whether it has been correctly applied or not. PPIs are required to determine whether a particular method generates actual benefits or not. They can be categorized into four categories: time, cost, quality, and flexibility. This framework provides practical implications in that it represents a ready-to-use tool for practitioners in conducting advanced redesign process analysis. Another benefit is that the evaluation of redesigns shifts

from a qualitative approach to a data-driven, quantitative one.

Finally, Figure 3 presents an overview of this research, including its research objectives and scope. We argue that it is an extension of process mining for healthcare. The *Observational Medical Outcomes Partnership Common Data Model (CDM)* is applied to describe the data stored in hospital information systems for the purpose of enhancing flexibility. Clinical event logs are then constructed for process mining based on the collected healthcare data using the ETL process. In this step, we propose an event log specification template and the use of structured query language (SQL) based on the clinical CDM. These clinical event logs become the foundation for clinical process analysis, including process discovery, performance analysis, and replay. This clinical process analysis is then used to conduct simulation analysis for redesign purposes. After the implementation of the improved process, an evaluation of a redesign is also conducted by comparing the as-is and to-be processes. These results finally lead to changes in clinical processes and healthcare information systems. As a result, this process has a cyclical structure.

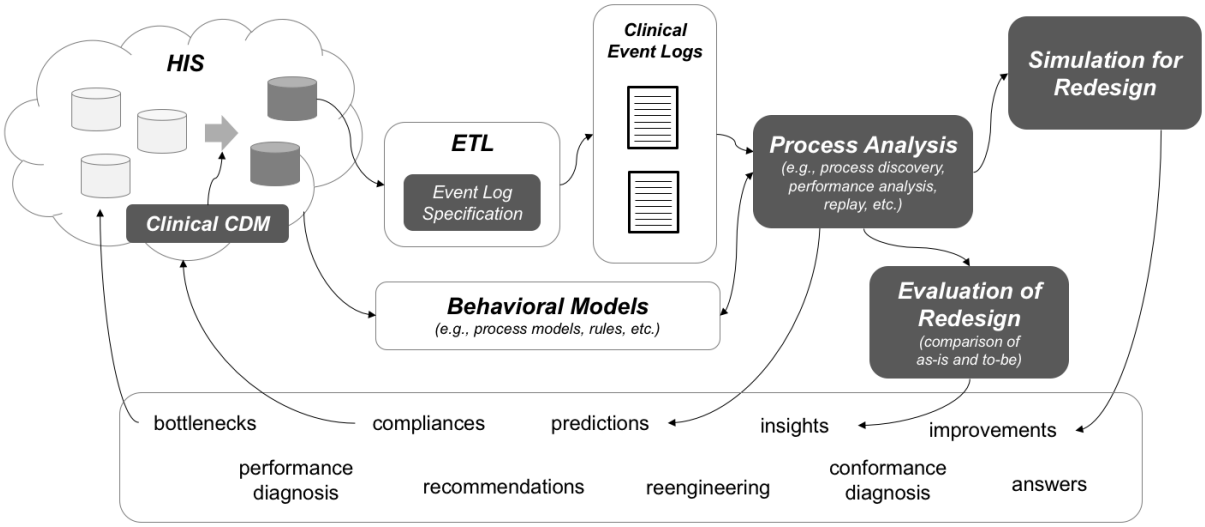


Figure 3. The overview of this research

1.4 Structure of This Dissertation

This dissertation is organized as follows:

- **Chapter II: Literature Reviews**

  This chapter reviews existing literature on business process management and data-driven approaches in healthcare. The literatures are categorized into the three corresponding concepts: (1) Business Process Management (BPM), (2) Process Mining, and (3) Data Science in healthcare. Finally, the comparison of the proposed methodology with the existing works is provided.

- **Chapter III: A Data Analysis Methodology for Process Diagnosis and Redesign in Healthcare**

  This chapter proposes a data analysis methodology for clinical processes in healthcare. Regarding the each phase in the methodology, the detailed explanation is provided such as building event logs with the guidance of CDM and a couple of aspects for data analysis in healthcare: clinical process types and process mining types.

- **Chapter IV: Diagnosing Clinical Processes of Outpatients, Inpatients, and Clinical Pathways**

  This chapter introduces frameworks for diagnosing clinical processes based on the proposed methodology. Here, three different clinical processes including outpatients, inpatients, and clinical pathways are considered; thus, the corresponding frameworks are developed. Furthermore, we provide the specific objectives and the detailed relevant analysis methods.

- **Chapter V: Redesigning Clinical Processes with the Simulation-based Approach**

  This chapter provides a framework for redesigning clinical processes with discrete event simulation and process mining. The presented framework covers building a data-driven clinical simulation model that overcomes the challenges from the existing approach, deriving applicable redesign recipes from data analysis, and validating them with simulation model analysis.

- **Chapter VI: Evaluating Effects of Process Redesigns in Healthcare**

  This chapter develops a framework for evaluating effects of process redesigns in healthcare with a data-driven quantitative approach. More in detail, it provides process performance indicators and indicators to assess whether process redesign best practices have been applied and to what extent.

- **Chapter VII: Conclusion**

  This chapter concludes this dissertation by summarizing the results discussed and describes the future research.

II Literature Reviews

This chapter discusses the literature reviews required for process mining in healthcare. More in detail, Chapter 2.1 discusses business process management including its basic concept and lifecycle. Chapter 2.2 describes process mining placed a key role in this dissertation. Chapter 2.3 provides an overview of multiple existing works for data-driven approaches in healthcare.

2.1 Business Process Management

According to Dumas et al., *business process management* (*BPM*) can be defined as follows: "*art and science of overseeing how work is performed in an organization to ensure consistent outcomes and to take advantage of improvement opportunities*" [8]. That is, BPM aims to build an efficient system of business-related fundamentals including events, activities, and decisions to improve the performances depending on the objective of an organization [18]. This continuous interests in business process management in academia and industry are not what happens today. As far as the research standpoint is concerned, researchers have continuously developed novel approaches for radical and gradual enhancement of processes. The typical examples were new methods for *business process redesign* [8, 19–22], developments of *business process modeling language* for a better representation [8, 23–27], and implementation of *business process management systems* for an efficient enactment of processes [8, 28, 29]. Besides, recent research has been conducted to incorporate uprising innovative smart technologies and BPM to give process participants automated and intelligent insights into the process perspective [2]. With regard to the industry viewpoint, there have been multiple applications to gain competitive advantages including cost reduction, execution time reduction, error rate reduction, and revenue growth through business process management and innovation [8].

The key concept of business process management, a tool that flexibly reacts to changes in business environments, is *business processes*. There have been various definitions of business processes by different researchers so far. Foremost, Hammer and Champy [19] described a business process as "*a collection of activities that takes one or more kinds of input and creates an output that is of value to the customer*". Also, Davenport [22] referred as "*a specific ordering of work activities across time and place, with a beginning, an end, and clearly identified inputs and outputs*". These definitions focused on a partial order of activities and the values of a process. Extending these definitions, Ko [30] defined a business process as "*a series or network of value-added activities, performed by their relevant roles or collaborators, to purposefully achieve the common business goal*". Furthermore, Weske [31] indicated as "*a set of activities that are performed in coordination in an organizational and technical environment. These activities jointly realize a business goal. Each business process is enacted by a single organization, but it may interact with business processes performed by other organizations*". In a nutshell, the primary concepts of a business process are a partial ordering of activities with an input and output, an

organization that performs an activity, and a specific goal of a process.

A business process can be represented as a document, i.e., a series of sentences [32,33]. However, it triggers a time or cost spent to comprehend a whole business process [8]. For overcoming this limitation and an explicit representation of a business process, multiple graphical modeling notations have been developed. Typical examples of representation methods are *Petri-net* [23,24], *Yet Another Workflow Language (YAWL)* [25], *Event-Driven Process Chains (EPC)* [27], and *Business Process Modeling Notation (BPMN)* [26]. These notations have the distinction of being useful in representing process control flows (e.g., AND-split, XOR-split, OR-split, AND-join, XOR-join, and OR-join) and containing numerous BP-related information including definition of activities or rules and decision points [34].

We provide a simple business process with a clinical setting as depicted in Figure 4. It aims at decreasing waiting time through building an efficient procedure for outpatients. The relevant process is represented as a process model with BPMN. In the figure, we can identify that the process contains two events (e.g., *Patient Registered and Patient Left*) and 6 different activities with a specific order (e.g., *Register, Check a Vital Sign, Consult & Treat, Conduct Lab Test, Pay for Care, and Appoint for a Further Visit*). Also, each organization group (e.g., *Administrative Office, Lab & Vital Test Office, and Physician Office*) serves a couple of activities, whereas the process includes the XOR control-flow (e.g., *Needed Lab Test? and Needed Further Visit?*). As such, process modeling enables users to understand the process at ease.

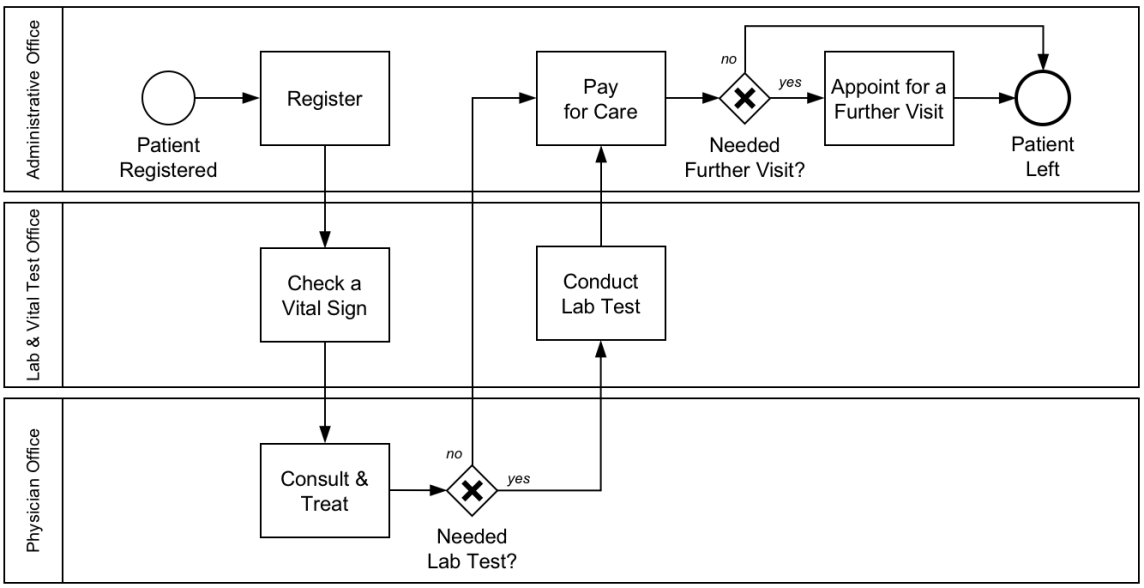


Figure 4. A simple example of a clinical business process

Heretofore, we have described BPM, BP, and modeling notation in detail, and it has demonstrated that these concepts are essential. Nonetheless, a specification of BPM is required for practical use. *Business process management lifecycle* becomes a solution, and there is numerous literature to conceptualize it. This dissertation introduces the BPM lifecycle presented by

Mendling et al. [2]. It represents an extended version of BPM lifecycle that differs from the others which only focused on the process model level. Figure 5 depicts the BPM lifecycle.

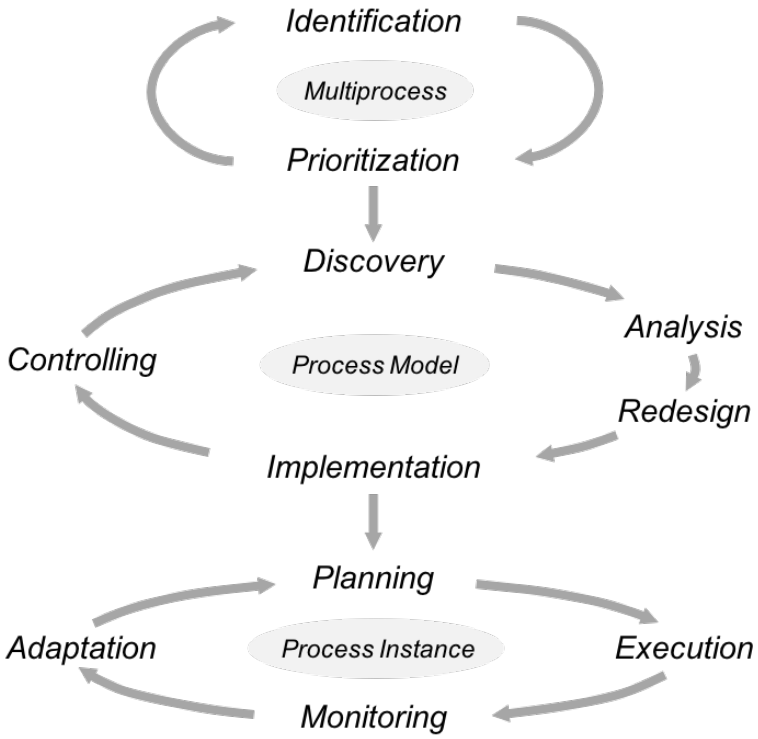


Figure 5. Business process management lifecycle [2]

BPM lifecycle is composed of three different levels: *multiprocess*, *process model*, and *process instance*. The top level, i.e., multiprocess level, has two steps. First, a couple of fundamental processes for an organization is identified in the *identification* phase. Afterward, priorities of these processes are evaluated in the *prioritization* phase. These two are necessary for the overall performance of an organization and connected with a process repository.

The middle level, i.e., process model level, is required for managing a single process among them identified in the multiprocess level. The *discovery* step (i.e., as-is process modeling) aims at documenting a current business process. In the *analysis* phase, multiple problematic issues are investigated based on the thorough qualitative and quantitative analyses on the as-is process model. These results become the foundation for deriving a to-be process model. On the basis of products from the analysis phase, the *redesign* phase builds the improved process model which resolves challenges identified before. In such a step, multiple redesign candidates can be considered; as such, the analysis and redesign steps are repeatedly performed to choose an optimal solution. The *implementation* phase applies the derived to-be model to an organization in practice. As a result, it may result in changes in organizations and process flows as well as infrastructures from automation with technical development. Finally, the *controlling* phase has a goal to assess results from the redesign, which utilizes multiple indicators, a predefined

template for evaluation, and data applied improvements before and after. Here, if the effects from redesigns are not satisfied, new iteration goes to the first phase in the cycle.

The process instance management, i.e., the bottom level, consists of four phases and aims to manage each process instance generated by executing a process. First, each process instance receives a plan considering process flows and resources in the *planning* phase. After that, instances are enacted based on the prepared schedules in the *execution* step. Then, the *monitoring* step is performed with the event stream data [35]. In such a step, a series of indicators are employed, and alerts are triggered if needed. The monitoring results are connected to the *adaptation* phase to get a better enactment for each process instance.

Among these three levels, process mining is actively applied to the BPM lifecycle for the process model management, in particular, the analysis phase. However, it has plenty of possibilities to extend the scope of applicability to other phases in the cycle. Therefore, this dissertation contributes to utilize process mining for re-engineering in healthcare.

2.2 Process Mining

As discussed before, the research for BPM has been manifold, and one of the key disciplines is *process mining* [3]. This concept is a relatively young concept focused on extracting process-oriented knowledge from event logs stored in information systems [3, 36–38]. In other words, process mining aims at analyzing data in a process perspective to fully comprehend the whole process and identify its improvements.

Event logs are the primary artifact of process mining [3, 39]. Event logs, i.e., the inputs of process mining, are a collection of *cases (i.e., process instances)*, where a case is a sequence of *events* (describing a trace). In other words, each event belongs to a single case. Events can have multiple attributes including an activity, an originator, an event type, and a timestamp. Thus, events can be expressed as assigned values for these attributes. Here, the π function is used to represent events as the attribute values. For instance, $\pi_{act}(e_1) =$ *'Consultation'* represents that the name of the activity of the event e_1 is consultation. Definition 1 provides a formal explanation of event logs.

Definition 1 (Events, Cases, and Event Log) *An event log L is a set of traces T, where a process instance (i.e., a patient in clinical event logs) has one trace. A trace is a finite sequence of events E. An event $e \in E$ includes multiple required attributes AT including the name of the activity (i.e., act), the completion time (i.e., ctime), the reservation time (i.e., rtime), and the name of the resource (i.e., res). For the specific attribute, we can get the corresponding value using π function. Here, $\pi : E \rightarrow (AT \nrightarrow V)$ is a function which obtains attribute values recorded for an event. Hence, $\pi_{at}(e) \in AT \nrightarrow V$ signifies to obtain the corresponding value $v \in V$ recorded for attribute $at \in AT$.*

A simple example log is provided in Table 1. In the table, 8 events for two cases are included,

and each line corresponds to an event. In the example, we can identify that the trace of the case 1 including the first six events refers to a process instance where *registration* was conducted by Paul at 09:00, *test* was performed by Allen at 09:45, *consultation registration* was given by Mike at 10:20, *consultation* was conducted by Chris at 10:40, and finally *payment* was performed by Paul at 11:00 in 2018-01-01.

Table 1. A partial example of event logs

Case	Event	Activity	Originator	Timestamp
Case 1	E1	Registration	Paul	2018-01-01 09:00
Case 1	E2	Test	Allen	2018-01-01 09:45
Case 1	E3	Consultation Registration	Mike	2018-01-01 10:20
Case 1	E4	Consultation	Chris	2018-01-01 10:40
Case 1	E5	Payment	Paul	2018-01-01 11:00
Case 2	E6	Registration	Paul	2018-01-01 09:30
Case 2	E7	Consultation Registration	Mike	2018-01-01 10:00

Process mining consists of three main types, namely process *discovery*, *conformance*, and *enhancement* [3, 37, 38]. Figure 6 depicts the overview of scopes of process mining.

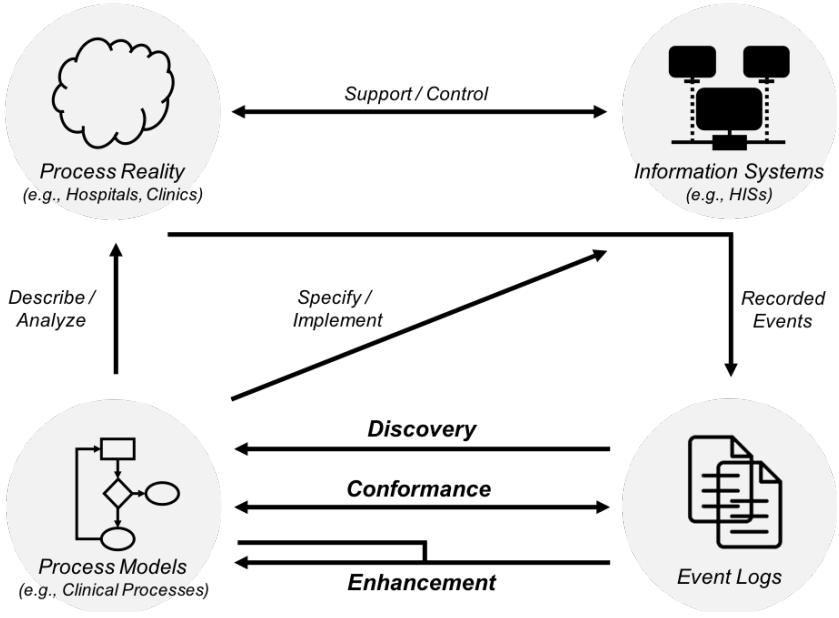


Figure 6. An overview of the three main types in process mining [3]

First and foremost, discovery of process models, one of the most challenging process mining tasks, deals with automatically constructing a process model using event logs without a-priori information [3, 40–45]. Because it is built from actual data in an event log, these models cap-

ture the real behaviors seen in the logs, as opposed, for instance, to process models in work instructions, which often capture an ideal process that is executed differently in reality. There are many techniques available for process discovery, which differ in terms of the algorithm used for discovery and the type of model discovered. The typical examples of discovery algorithms are *alpha-mining* [3], *heuristic mining* [45], *genetic mining* [41], *fuzzy mining* [42] and *inductive mining* [43].

Conformance checking aims at investigating a discrepancy between actual behaviors observed in a log and a model discovered from the log or designed by manually [3, 46]. By doing so, the quality of the model is assessed with leveraged four indicators: *fitness* (i.e., identifying the observed behavior is captured), *precision* (i.e., investigating models are general), *generalization* (i.e., identifying models are overfitted), and *simplicity* (i.e., examining models are enough simple) [3, 46].

Finally, the last type, i.e., enhancement, refers to changing or improving the discovered model based on other insights derived from event logs [3]. Here, several perspectives including organizational, performance, and case are combined into the model from the control-flow view, and we can derive a more useful process model [3].

In orthogonal to the three types of analysis, process mining defines four perspectives, i.e., control-flow, case, organizational, and time perspectives [3]. The control-flow, organizational, and case perspectives focus on the order of activities, the resources involved, and the characteristics of cases in the process, respectively [37]. Note that while the control-flow perspective mainly focuses on discovering process models or frequent episodes in an event log [36, 43, 47], the case and organizational perspectives define additional views of processes, such as the temporal logic checker [48], i.e., to check automatically the satisfaction of particular logic constraints case by case based on information in the event log or the social network [49, 50], i.e., a graph capturing handovers of work among resources involved in a process. The time perspective is more relevant with performance analysis by considering the timing and frequency of events in a process [37, 51]. Therefore, it can be employed to discover bottlenecks in a process model, monitor performance of originators, and calculate workloads.

2.3 Data Science in Healthcare

This chapter provides a simple introduction for data science in healthcare. We first introduce *hospital information systems* that collect clinical data and *common data model* representing a standardized medical data. Afterward, we describe the reviews on the quantitative methods including process mining, data mining, and operations management in healthcare.

2.3.1 Hospital Information Systems and Common Data Model

According to Winter et al. [52], a *hospital information system (HIS)* can be defined as "*the socio-technical subsystem of a hospital, which comprises all information processing as well as*

the associated human or technical actors in their respective information processing roles". Thus, HIS aims to facilitate a wide range of healthcare-related functions, such as patient care, patient administration, and hospital management [53]. More in detail, it covers the whole artifacts including wards, outpatient units, clinical services (e.g, diagnostics, laboratories, and operations), and hospital administration (e.g., costs and human resources) in a hospital [52].

Hospital information systems consist of numerous applications. The typical examples are *Medical Practice Management System*, *Electronic Health Records*, *Laboratory Information System*, *Computerized Physician Order Entry*, and *Picture Archiving Communication System*. A simple explanation on each system is as follows.

- Medical Practice Management System (MPMS): An application that manages the various aspects of a medical practice to enhance the efficiency and quality of the operation of a medical office [54].
- Electronic Health Records (EHR): An application that collects a comprehensive, cross-institutional, and longitudinal healthcare data of patients [55].
- Laboratory Information System (LIS): An application for collecting, recording, presenting, organizing, and archiving laboratory results [56].
- Computerized Physician Order Entry (CPOE): An application that helps for ordering stage of medications, where most medication errors and preventable adverse drug events occur [57].
- Picture Archiving Communication System (PACS): An application which handles imaging modalities in radiology [58].

Thanks to the developments and applications of these information technologies, a plenty of data have been collected. Such a data enables to implement data-driven approaches including data mining, machine learning, and process mining with an aim to improve a healthcare environment. Thus, they have become a tool to resolve problems occurred in a hospital. As far as the research side is concerned, there have been one-off and gradual developments of new methods. However, in industry, it was hard to apply these methods because data formats are different each other.

To overcome this limitation, *Observational Medical Outcomes Partnership* constructed a *common data model (CDM)* [4, 59, 60] (Hereinafter, we call the OMOP CDM as CDM). They aimed to inform the appropriate use of observational healthcare databases with a standard format. More in detail, CDM was developed to investigate the associations between clinical interventions and outcomes. In such a process, by combining with standardized contents, it facilitates researchers involved in medical informatics to derive meaningful reproducible and comparable results. CDM is designed with 9 basic disciplines: suitability for purpose, data protection, design

of domains, rationale for domains, standardized Vocabularies, reuse of existing vocabularies, technology neutrality, scalability, and backward compatibility [4, 59].

Figure 7 depicts a latest conceptual model for CDM [4]. It consists of six different data groups (e.g., *standardized vocabularies*, *standardized meta-data*, *standardized clinical data*, *standardized health system data*, *standardized health economics*, and *standardized derived elements*), which includes a couple of tables [4, 59]. First, standardized vocabularies include 15 tables which embed the detailed information about clinical concepts used in all of the fact tables. In other words, these tables are used to represent the standardized clinical concepts from different source terminologies. Standardized meta-data has a role to record and manage data sources. Standardized clinical data, i.e., the primary sector within CDM, contains the holistic records of hospital visits of patients and demographic information of them. As such, it can be the main artifact for evidence-based approaches including data mining and process mining. This research also focuses on standardized clinical data to create event logs for process mining. Details are given in Chapter 3.2. Standardized health system data is composed of a collection of tables containing organizations and providers that serve clinical services to patients. Standardized health economics represent cost information of clinical concepts, whereas standardized derived elements describe the rendered data from clinical events.

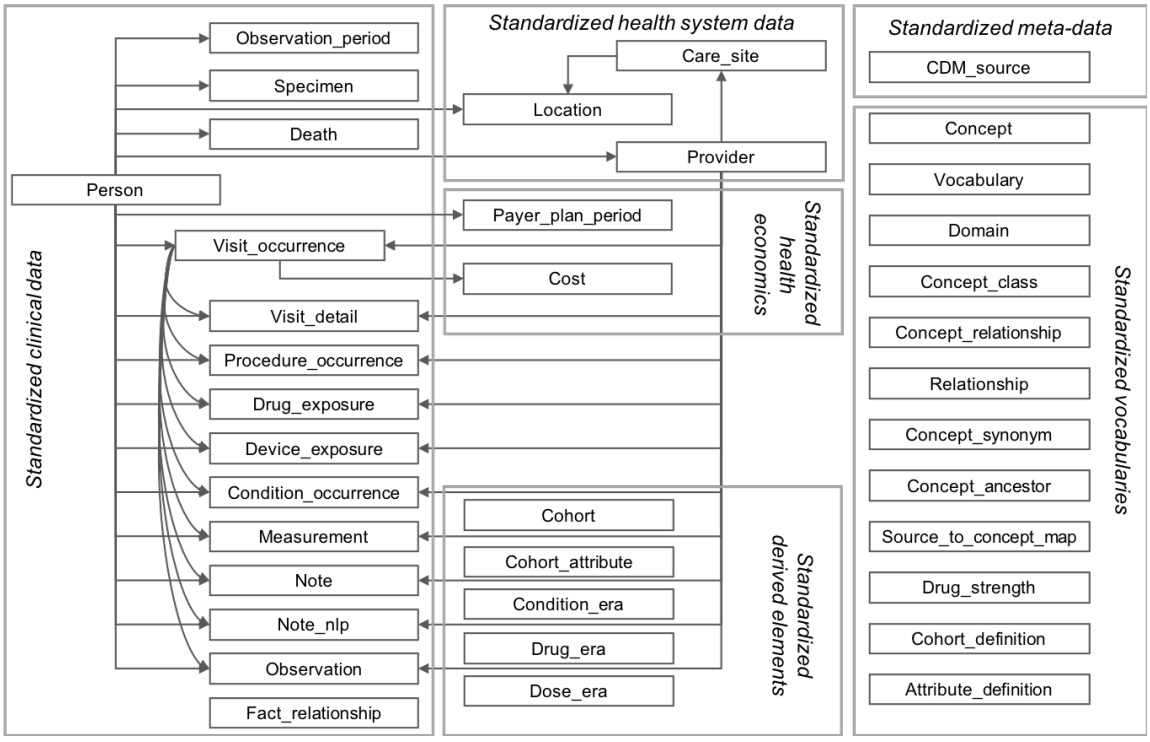


Figure 7. The conceptual model for OMOP CDM [4]

So far, numerous research based on CDM have been conducted. First, researchers have contributed to assess CDM with a focus on flexibility [61–64]. Also, some research have identified that existing data (e.g., *Electronic Health Records*, *Clinical Practice Research Datalink* and *National*

Health Insurance Service-National Sample Cohort Database) can be converted or not [65–68]. Finally, there have been developments of querying and ETL for CDM configurations [69, 70].

2.3.2 Process Mining in Healthcare

Due to the complexity of medical data and increase of demands in data analysis with a clinical process level, there have been many attempts for process mining in healthcare. As a result of searching papers indexed by *scopus* with *"process mining"* and *"healthcare"*, we received 140 relevant papers. Also, there were already eight review papers for process mining in healthcare [16, 71–77]. This part describes the results of literature reviews only for journal articles among numerous studies related to process mining in healthcare. The reasons why we only focus on the journals are as follows: conference proceedings generally include the research in progress, which become a piece of the journal article of the same author; it is unnecessary to review every article since this part aims to introduce the overview of the relevant studies.

After identifying the papers filtered, we analyzed them with types and perspectives of process mining. As far as the type of process mining is concerned, we identified that discovery (59.4%) was the commonly applied, which followed by enhancement (25.0%) and conformance checking (15.6%). In the viewpoint of perspectives, 51.2% of studies were concerned with the control-flow perspective, whereas the performance and organizational perspectives were occupied 34.1% and 14.6% of them, respectively. Hereinafter, we provide a detailed explanation of the relevant research according to the process mining types.

Foremost, regarding the discovery type of process mining, there have been investigations for automatically extracting clinical workflow process models using the whole event logs [40, 78–80]. As an extension of these approaches, Rovani et al. [81] developed a method for the declarative process model that enables to identify the compliance of the clinical guidelines. Furthermore, there have been efforts to create clinical pathways, i.e., standardized clinical process guidelines, from event logs [82–85]. Also, a couple of studies have performed a comparative analysis of discovered multiple processes [86, 87]. To tackle the challenge that clinical processes are generally complicated, some studies have suggested the discovery combined with clustering techniques [9, 88]. Besides, there have been attempts to extend discovery results to other disciplines. For example, discovered models are applied as a source to build a simulation model for redesigns [89–91] and an optimization model to find out the optimal layout [92]. In addition to the clinical workflow data, other data such as medical imaging test data [93] and indoor location system [94] were also applied to construct a process model. Finally, Alvarez et al. [95] have attempted to derive an organizational model in a clinical setting.

For the conformance checking, the relevant literature can be classified into two groups: the process model level and trace level. First, multiple research has applied conformance to assess how accurately the derived process model was created [40, 78, 81, 87]. On the other way, there have been efforts to employ the trace alignment technique to measure the conformance of every trace [96].

Furthermore, a couple of studies have extended the conformance to the outlier detection approach in a clinical process [97, 98].

As far as the enhancement is concerned, there has not been sufficiently conducted yet, compared to other two types. Among those few studies, most of them only focused on the performance perspective in analyzing a clinical process [40, 81, 86, 87]. It is believed that the clinical performances are essential for the outcome of patients. Furthermore, there has been an approach applying enhancement with an aim to improve the clinical process (i.e., process repair) [99].

Also, there have been some proposals to develop a data analysis methodology with process mining. The typical three approaches are as follows: Process diagnostics method [100], L* life-cycle model [3], and The process mining project methodology (PM^2) [101]. Process diagnostics method has focused on quickly understand and diagnosing a given process at a broad level. Compared to this, L* life-cycle model was an extended methodology that covers various aspects of process mining including process improvement and operational supports. The process mining project methodology was the data analysis methodology designed to support process mining projects. These methodologies were not healthcare-oriented but generic methodologies. In other words, they did not present a healthcare reference model for creating clinical event logs or suggest methods and techniques in the healthcare domain.

On the other hand, there have been healthcare-oriented data analysis methodologies in process mining. Process mining in healthcare [1] provided an overview of process mining approaches in a medical setting. In particular, it proposed a comprehensive healthcare reference model outlining all the different classes of data that can be applicable for process mining. Also, the methodology for declarative process mining [9] was proposed to analyze the conformance and deviations rather than process discovery or enhancement by employing a data split method and the existing guidelines. Furthermore, there was a questionnaire-driven data analysis methodology in process mining to improve the efficiency and effectiveness of emergency rooms. In this regard, the authors defined the frequently-posed questions for each step in the methodology. Besides, there have been clustering-based methodologies [9, 82, 102, 103] since clinical process models generally take spaghetti-shaped (i.e., complicated) behaviors. Lastly, there have been other ad-hoc methodologies in process mining [81, 83, 94, 104–106]. These proposals serve as guidelines enabling for non-experts to easily and effectively perform data analysis with process mining techniques. However, several limitations are associated with these approaches, e.g., it provides a broad overview, it does not give any reference models, or it is not extensible to further improvements (i.e., redesigns). To deal with these challenges, we develop a data analysis methodology for process diagnosis and redesign in healthcare. The detailed comparison of the proposed methodology with these existing works are presented in Chapter 2.3.4.

2.3.3 Other Quantitative Approaches in Healthcare

As described before, there have been already numerous studies on *process mining* in healthcare. In addition to this, we also provide the other methods for process diagnosis, redesign, and evaluation with quantitative approaches, i.e., data mining and operations management, in a medical setting.

1) Data Mining in Healthcare

Data mining approaches in healthcare have been applied to enhance the health of patients and outcomes of clinical cares [107]. Based on the classification techniques including *decision trees, k-nearest neighbor (K-NN), support vector machines (SVM), bagging, boosting* and *random forest*, researchers have developed how to predict the change of survival of breast cancer patients [108], classify the activity of chronic disease [109], delineate smoking behaviors from psychological health and distress, and demographic information of patients [110], and characterize skin diseases [111]. Also, research on predicting the clinical outcome with a numerical value has been mainly performed as follows: predicting the number of hospitalization days [112], length of hospital stays with features of inpatients [113], and estimating risks for medical conditions including diabetes and strokes [114] based on the regression techniques. Lastly, for the clustering, there have been following applications: grouping the patients based on the length of stay to provide better services and outcomes with the agglomerative hierarchical clustering [115], applying k-means and agglomerative approaches to analyze large microarray data [116], and conducting hybrid methods with other supervised learning techniques such as incorporating with classification trees to predict healthcare costs [117] and with SVM to classify cancer diseases [118].

As the same as process mining approaches, these research are relevant to the quantitative methods and utilize data. However, it just focuses on solving a specific established problem in clinical processes, while process mining covers a whole process.

2) Operations Management & Research in Healthcare

Operations management & research in healthcare is relevant to planning and controlling of all steps in clinical processes required for providing a care delivery to patients [119]. Therefore, existing works in this discipline can be closely related to the process diagnosis and redesign considered in this research. In this regard, there have been three typical research streams as follows: queuing, simulation (e.g., agent-based model), and mathematical programming [120]. Hereinafter, we describe the literature reviews in each research stream in operations management & research in healthcare.

Regarding the queuing models, queues are formed when it arrives at a service facility and can not be performed immediately to customers. In healthcare, queuing models have been commonly applied since healthcare service systems are characterized by random demand [120]. In this regard, the capacity calculation is one of the typical problems explored with queuing models in healthcare operations management, and followings are the relevant studies: determining the

panel size deriving the optimal number of care providers [121,122], exploring the proper number of beds in emergency departments [123], and discovering nursing levels [124]. Also, queuing theory has been applied for appointment scheduling [125].

Regarding the simulation approach, the relevant studies have focused on a single care site, a department, or a resource and considered patient flow characteristics and scheduling. In this regard, [126] developed a simulation model to minimize the patient delays, and [127] tried to improve the patient flows based on patient waiting times, resource utilization and overtime. Furthermore, the simulation approach has been applied to the emergency department: focusing on the total time spent in emergency rooms by patients [128], exploring appropriate staffing levels for expected arrival rates by patients [129], and balancing economic incentives, workload, and quality of care [130].

Regarding the mathematical programming, optimization has been widely applied in modeling and solving healthcare operations management problems including appointment scheduling, operating room scheduling, capacity planning, and staffing scheduling. For the appointment scheduling, there have been following approaches: exploring the single-server appointment scheduling problem [131, 132], applying the stochastic programming approach [133], and proposing the sampling-based approach [134, 135]. Also, there have been some approaches to derive the optimal operation room scheduling as follows: considering multi-OR scheduling problems [136, 137], allocating surgeries on a given surgical day where the surgery durations are uncertain [138], and providing the stochastic multi-OR scheduling problems [139].

The existing works for operations management (e.g., queuing theory, simulation, and optimization) in healthcare have focused on resolving problems identified in clinical processes. However, these methods were qualitative methods with secondary data; thus, they did not reflect the reality.

2.3.4 Comparison of The Proposed Methodology with The Existing Works

To clarify the distinctive traits of our methodology, we explicitly compare our approach with existing works for quantitative methods or methodologies in healthcare. Table 2 provides the detailed comparison results, which occurs along six criteria: what a research method is mainly applied (Research Method), whether an approach is data-driven (Data-driven), whether it provides the process-level approach (Process-level), whether a particular healthcare reference model is suggested (Healthcare Reference Model), whether it provides detailed methods or data in a medical setting (Detailed Methods or Data), and whether it is extensible to redesign (Extensible to Redesign).

Table 2. Comparison of the proposed methodology with the existing works

Proposal	Research Method	Data-driven	Process-level	Healthcare Reference Model	Detailed Methods or Data	Extensible to Redesign
[100]	Process Mining	✔	✔	X	X	X
[3]	Process Mining	✔	✔	X	X	✔
[101]	Process Mining	✔	✔	X	✔	X
[81]	Process Mining	✔	✔	X	✔	X
[9]	Process Mining	✔	✔	X	✔	X
[82]	Process Mining	✔	✔	X	✔	X
[94]	Process Mining	✔	✔	X	✔	X
[1]	Process Mining	✔	✔	✔(Ad-hoc)	X	X
[106]	Process Mining	✔	✔	✔(Ad-hoc)	✔	X
[107–118]	Data Mining	✔	X	✔/X	✔	X
[119–121, 124, 126, 128, 131, 139]	Operations Management	X	X	X	✔	X
The proposed method	Process Mining	✔	✔	✔(OMOP CDM)	✔	✔

As a result, first of all, all reviewed proposals mostly utilize process mining, data mining, statistical methods, and operations management. Among them, the most of the proposals for operations management in healthcare is not data-driven approaches; thus, there is a lack of accuracy because it does not sufficiently reflect the reality. Also, some proposals do not cover the whole process-level, but the problem-solving for a specific problem in clinical processes is concentrated. In this case, problem identification for the whole process is impossible. In other words, the challenges of a specific part can be resolved, but it is difficult to identify the side-effects of this in the other part. Regarding the healthcare reference model, some proposals do not provide at all, and others provide their ad-hoc methods. Therefore, it is impossible to build systematic and applicable methods for data extraction in healthcare. A subset of existing works only focuses on providing a holistic viewpoint, defining broad and coarse-grained levels for data analysis in healthcare. These approaches tend not to define specific methods or indicators that can be directly be used by practitioners. Lastly, only a couple of approaches covers a further data analysis for redesigns, i.e., connections between data analysis and redesign. Compared to the reviewed existing works, our methodology provides data-driven approaches for the whole process-level. Also, based on the OMOP CDM, i.e., the standardized healthcare reference model, it proposes an explicit and applicable data extraction method. Besides, the proposed methodology suggests the detailed methods including indicators and gives evidence-based support to the redesign phase in the business process lifecycle.

III A Data Analysis Methodology for Process Diagnosis and Redesign in Healthcare

This chapter presents a clinical process analysis methodology with a data-driven approach (i.e., *process mining*). It aims at providing a generic framework that allows non-experts to perform the whole phases including data preparation, preprocessing, analysis and post-hoc analysis. This chapter is organized as follows: Chapter 3.1 presents the overview of the proposed methodology. Also, Chapter 3.2 introduces how to create event logs with the common data model, and Chapter 3.3 explains the data preprocessing phase to make a better quality of data. Furthermore, Chapter 3.4 explains how to perform data analysis with process mining techniques, and Chapter 3.5 explains the last stage, post-hoc analysis, in the methodology. Finally, Chapter 3.6 summarizes this chapter.

3.1 An Overview of The Proposed Methodology

This chapter provides an overview of the clinical process analysis methodology. Figure 8 depicts the overview of the clinical process methodology, which consists of four phases: *data preparation*, *data preprocessing*, *data analysis*, and *post-hoc analysis*.

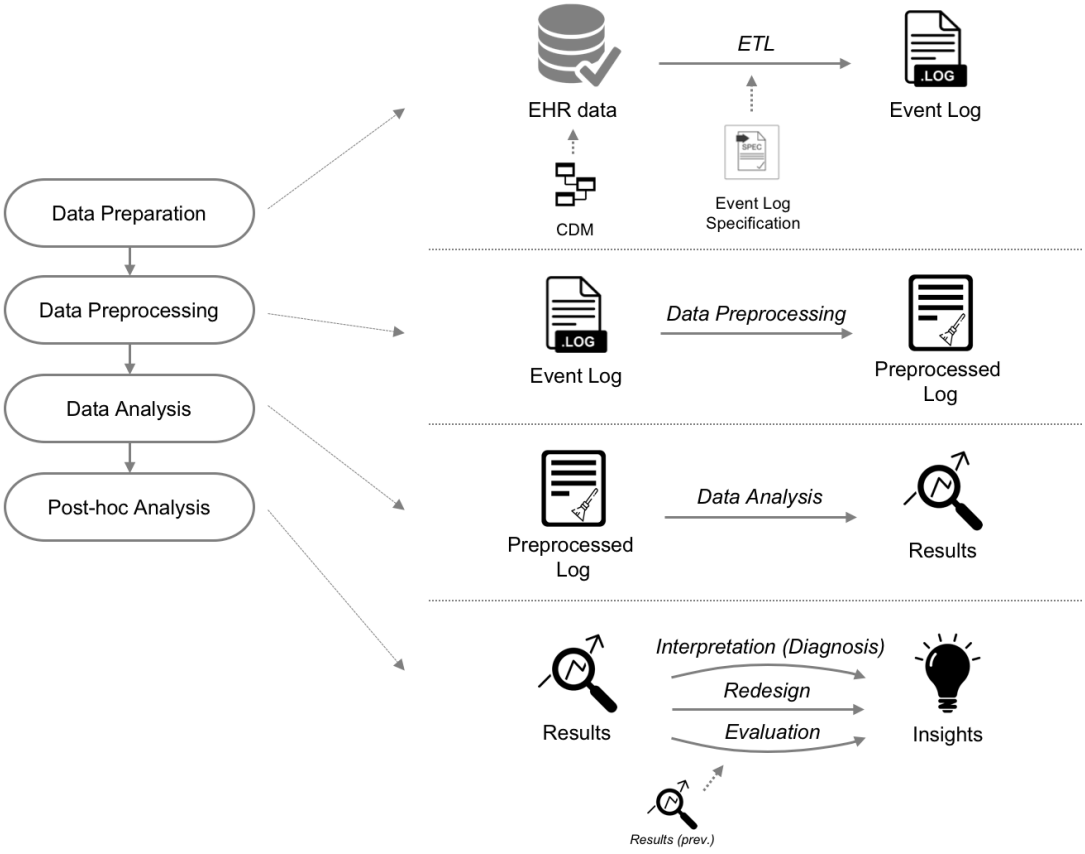


Figure 8. The overview of the clinical process analysis methodology

Each phase, consisting of a series of steps, has a specific goal leading to the meaningful insights. The data preparation phase aims at extracting data with a suitable format (i.e., event logs) for process mining data analysis. In this step, it is required to develop how to obtain clinical event logs from the CDM-mapped EHR data. To this end, we build an event log specification that helps to derive event logs considering the purposes, contents, scopes, and others of the data analysis desired by users. After building event logs, they are preprocessed to improve the accuracy and validity of the data analysis. Afterward, the data analysis phase, the core part in the proposed methodology, presents a couple of aspects to effectively conduct process mining analysis: clinical process types and process mining types. In the last phase, we interpret the results obtained from data analysis with domain experts and perform the post-hoc analysis to improve clinical processes with simulation or to evaluate with the previous data analysis results.

The proposed methodology can be distinguished according to the purpose of the data analysis, and Figure 9 shows the research methods from that. It can have three types of the objectives: diagnosis, redesign, and evaluation. That is, it enables to build the data analysis frameworks for understanding, improving, and evaluating clinical processes based on each type. More in detail, the frameworks for three types have the common streams of the data preparation, preprocessing, and analysis. However, as presented in Figure 8, the post-hoc analysis is differentiated based on the objectives. Research methods for each type will appear in the corresponding chapters.

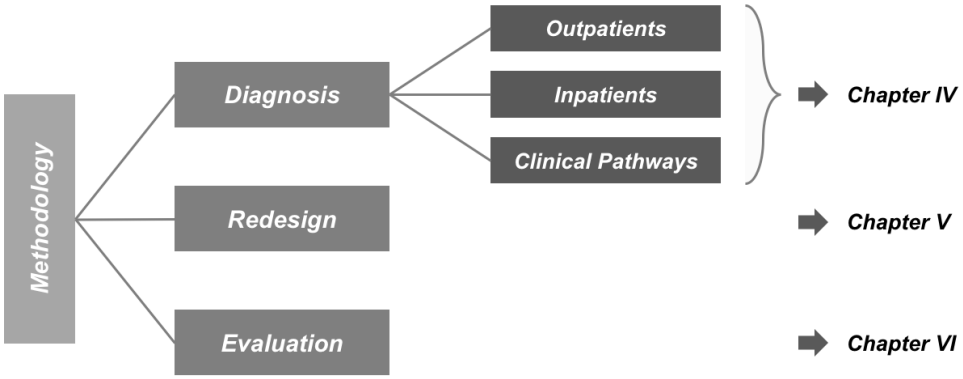


Figure 9. The research methods based on the proposed methodology

3.2 Data Preparation

The data preparation step, i.e., the first step of the methodology, aims at creating event logs that satisfies the goal of the data analysis. This step handles *creating event logs with ETL* from CDM-mapped EHR data based on the event log specification as explained in Figure 8. Here, we do not cover a process of CDM mapping which converts different data from medical institutions that use different hospital information systems into a standard format. This is because it is easily managed with existing works [140] and beyond the scope of this research. On the contrary to CDM (i.e., OMOP CDM) mapping, developing a mechanism to build event logs from CDM-

based EHR data has a straightforward connection with our work. Thus, this chapter introduces the table configurations associated with the event log of all CDM components and extracts event logs based on the suggested *event log specification*. Note that event logs can be constructed with EHR data itself which is not transformed with CDM.

As presented in Chapter 2.3, OMOP CDM is originally composed of 38 tables. Among them, patient-related information or log records are only shown on some tables; others serve as information for the standard of clinical concepts and vocabularies. Figure 10 depicts the only patient-related data in CDM, which consists of 20 tables including standardized clinical data, standardized health system data, and standardized health economics.

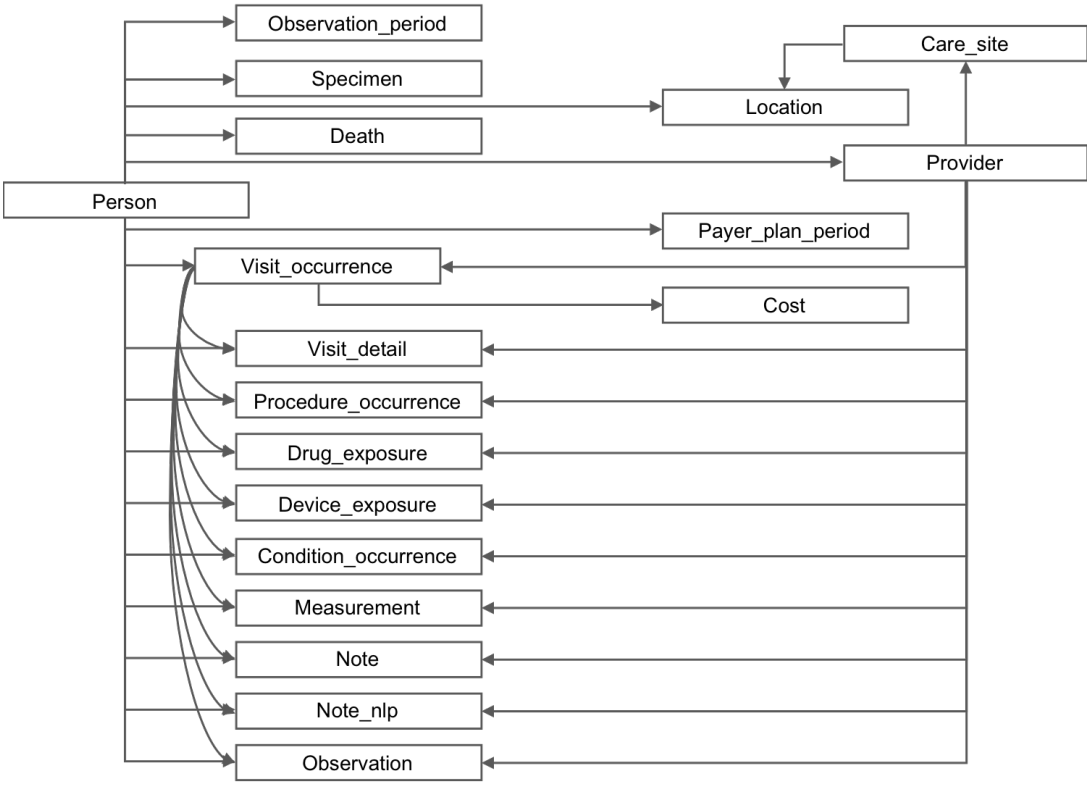


Figure 10. The patient-related data in common data model

The *Person* table at the top left of the figure contains personal information of patients such as the ages, races, and ethnicity. On the basis of the table, four classes are derived including *Observation_period*, *Specimen*, *Death*, and *Payer_plan_period*. Here, *Payer_plan_period* captures a specific health plan benefit structure of patients. *Visit_occurrence*, one of the paramount tables, contains information recorded when patients visit hospitals. For every visit of patients, corresponding records of cares delivered to patients and other information of them are stored in nine different tables: *Visit_detail*, *Procedure_occurrence*, *Drug_exposure*, *Device_exposure*, *Condition_occurrence*, *Measurement*, *Note*, *Note_nlp*, and *Observation*. Furthermore, providers that give cares to patients and the corresponding care sites are also recorded.

The details on patient-related tables are given in Figure 11. For a total of 20 tables, some

tables are excluded and reassembled under the following two conditions: whether a table is required and whether a table can be combined with others. As a result of applying the first condition, *Observation_period* is removed because it includes only the total length of time a patient has received medical cares by manipulating other tables. With the second condition, four tables are removed: *Death, Payer_plan_period, Location, and Cost*. The first two tables, i.e., *Death, Payer_plan_period*, are considered as attributes of patients that can be integrated into the *Person* table if needed. Also, *Location* table is turned out as a property of *Care_site* table. Lastly, four cost-related tables are removed because they could be included as attributes that are dependent on each care table. Thus, it is determined that this data is not required for the event-driven process mining analysis.

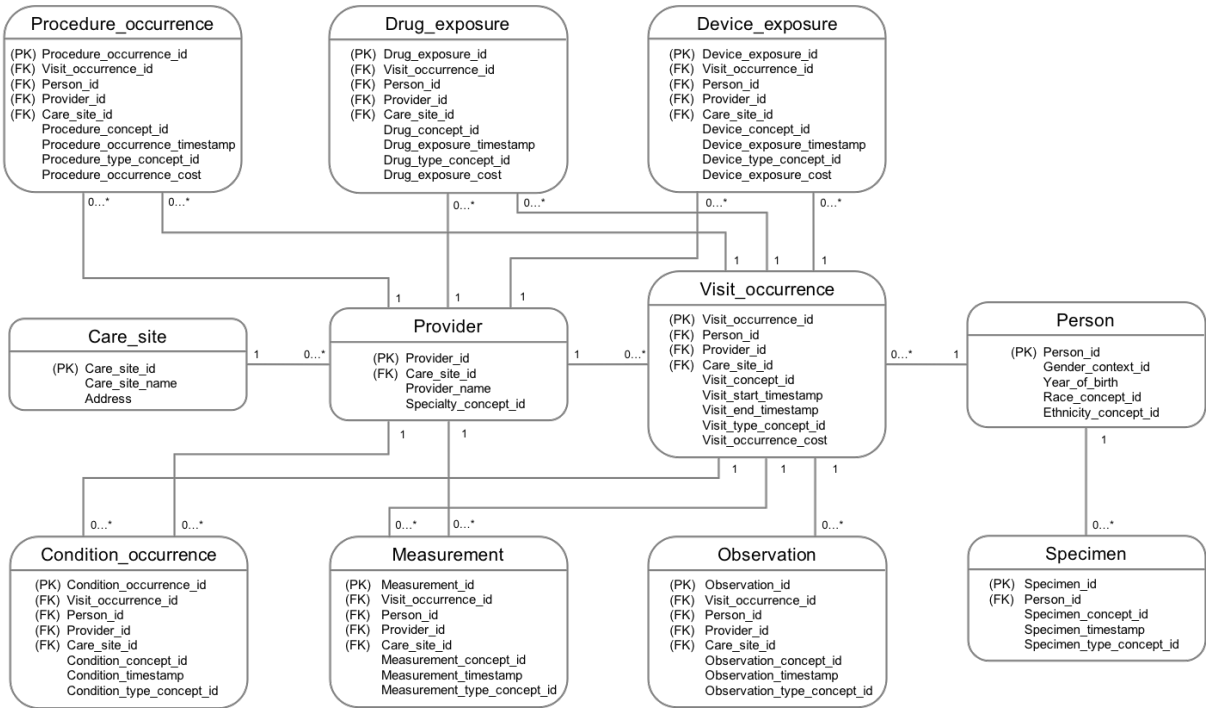


Figure 11. The detailed classes describing care delivery records of patients

For columns of each table, the elements marked as the required in the CDM document are only included; optional attributes also can be added if needed. More in detail for the figure, a patient can have multiple visits (i.e., *Person* class with associated multiplicity $0 \ldots *$), and there is a specific provider for each corresponding visit. Also, each care site can have multiple providers (i.e., *Care_site* class with associated multiplicity $0 \ldots *$). On the basis of visits and providers, the relevant manifold care classes delivered to patients are recorded (i.e., *Visit_occurrence* and *Provider* class with associated multiplicity $0 \ldots *$).

Even though the patient-related data is secured from CDM, it is not straightforward to modify into predefined clinical event logs with following reasons. An event log is a collection of records for patients associated with a particular process, however, CDM-based data is currently

mixed with multiple processes. In other words, it is a combination of the records of outpatients, hospitalized patients, and others; thus, we cannot build an event log with the whole data. Therefore, it is necessary to extract only patients corresponding to a specific process. Also, not all data from CDM patient-related data need to be included in event logs, and its generation is required to suit the user's preferences.

This research develops an event log specification template to cope with these limitations. The application of this specification helps to retain event logs with an easy and convenient approach in such a way that ambiguity disappears, and it acts as a manual to let users know the required and optional elements for building event logs. Table 3 depicts an event log specification template. In a nutshell, users simply fill in the <*something*> part by considering the objectives, scopes, and others for data analysis.

Followings are the descriptions of each field in the template.

- **CEL-<id>**: Every clinical event log is needed to be identified with unique identifier.
- **Description**: It describes the characteristics of the clinical event log. Considering the description, event log specifications can be reused for the next time.
- **Process type**: Users have to determine the process type, i.e., the particular process, to build a useful event log. The options that users can select are four types including the outpatient care, inpatient confinement, emergency room, and long-term care, which came out from CDM document.
- **Period**: A single event log cannot cover the whole period (i.e., infinite time period). That is, users have to specify a certain period of time for data analysis, and it determines whether a particular visit (not an event level) is included within the period.
- **Scope**: Users need to determine which clinical activities should be included in the event log. In other words, the log can be created by selecting some or the whole of six tables (i.e., *Procedure_ occurrence*, *Drug_ exposure*, *Device_ exposure*, *Condition_ occurrence*, *Measurement*, and *Observation*) associated with medical cares in the patient-related data.
- **Fields**: Considering the format of event logs, users have to determine four elements: case identifier, event identifier, case attributes, and event attributes. Here, essentials of event logs including cases, events, activities, timestamps, and originators are specified.
- **Features**: This is the only optional item in the specification form. It considers the additional features to create a limited or filtered event log.

Table 3. Event log specification from CDM

(*Req.*) CEL-<id>	<*Clinical event log identifier/name*>
(*Req.*) Description	<*Clinical event log description*>
(*Req.*) Process type	The clinical event log is specified with one of the following <*process type*> – *Outpatient care* – *Inpatient confinement* – *Emergency room* – *Long-term care*
(*Req.*) Period	The clinical event log is limited to a specific period – Visits (i.e., cases) between <*lower bound*> and <*upper bound*>
(*Req.*) Scope	The <*event tables considered for this clinical event log*> are described
(*Req.*) Fields	The clinical event log must determine case, event, and relevant attributes – Case identifier: <*Table: column name*> – Event identifier: <*Table: column name*> – Case attributes: <*Table: column name*> – Event attributes: <*Table: column name*>
(*Opt.*) Features	The clinical event log can be limited and filtered by other features – Care site: <*Physical or organizational units*> (e.g., hospitals) – Provider: <*Health care providers*> (e.g., physicians) – Patient type: <*Types of patients*> (e.g., ages, races, etc.) – Others

The event log specification template itself cannot be an immediately applicable tool for extracting event logs. That is, an explicit *Extraction, Transformation, and Loading* (ETL) process (i.e., SQL query or relational algebra) is required to get clinical data from a database. To this end, we provide a structured *SQL query* for data extraction from relational databases, which is straightforwardly bridged with the event log specification template. Followings are the basic form of the SQL query. Here, we provide an explanation considering the predefined template for a clear understanding. First, line 1-10 indicates how to settle main entries in the event log, which includes case & event identifiers and case & event attributes predetermined in *Fields* part. Line 11-24 shows the process of determining the tables required to configure the event log and of joining the selected tables. In such a process, a case-related table is created by combining *Visit_ occurrence*, *Person*, and *Provider*. Also, the tables included in *Scope* (line 22) and *Provider* are utilized to construct the event-related tables (line 20-24). Line 27-28 indicates determining the process type as defined in *Process type* presented in the template; *Visit_ type_ concept_ id* in the *Visit* table is relevant for that. Lastly, line 29-31 covers the *Period* in the event log

specification.

```
/* get required elements of event logs */
SELECT    Visit.Visit_occurrence_id as Case,
          ET.[Table included in Scope]_occurrence_id as Event,
          ET.[Table included in Scope]_concept_id as Activity,
          ET.[Table included in Scope]_occurrence_timestamp as Timestamp,
          ET.Provider_id as Originator,
          /* specify case attributes */
          [Visit.column_name],
          /* specify event attributes */
          [ET.column_name]
FROM    ( /* combine Visit_occurrence and Provider tables */
            SELECT *
            FROM    ( /* combine Visit_occurrence and Person tables */
                        SELECT *
                        FROM Visit_occurrence LEFT JOIN Person
                        ON Visit_occurrence.Person_id = Person.Person_id
                      ) AS Visit_temp LEFT JOIN Provider
            ON Visit_temp.Person_id = Provider.Provider_id
          ) AS Visit,
          ( /* combine [Table included in Scope] and Provider tables */
            SELECT *
            FROM [Table included in Scope] LEFT JOIN Provider
            ON [Table included in Scope].Provider_id = Provider.Provider_id
          ) AS ET
WHERE     /* connect Visit table and ET table */
          Visit.Visit_occurrence_id = ET.Visit_occurrence_id AND
          /* filter with a specific process type */
          Visit.Visit_type_concept_id = [Process type] AND
          /* filter with a specific period */
          Visit.Visit_start_date >= [Period.lowerbound] AND
          Visit.Visit_end_date <= [Period.upperbound]
```

An explicit example of using the event log specification template is presented in Table 4. First, the established clinical event log (i.e., *CEL1*) is about an outpatient data for infants at hospital 1 in January of 2018 as presented in *Description*. That is, only *outpatient care* process type is included in the event log, and outpatients' visits within January 2018 are subject to cases in the log. For this log, as an example, we only consider *Procedure_occurrence* table for

events among six different tables. In the *Fields* part, all components required for making up the event log, such as cases, events, activities, timestamps, and originators, are explained in detail. For example, the *Procedure_ concept_ id* column becomes the activities in the log. Lastly, considering the *Features*, only patients who visit *Hospital 1* and whose age is less than three years are contained in the log.

Table 4. An example of event log specification from CDM

CEL-<id>	*CEL1*
Description	*An outpatient care event log for infants at hospital 1 in January of 2018*
Process type	*Outpatient care*
Period	Visits belonging between *01-Jan-2018* and *31-Jan-2018*
Scope	*Procedure_ occurrence*
Fields	– Case identifier —— *Visit_ occurrence: Visit_ occurrence_ id* (Case) – Event identifier —— *Procedure_ occurrence: Procedure_ occurrence_ id* (Event) – Case attributes —— *Person: Gender_ concept_ id* —— *Person: Year_ of_ birth* – Event attributes —— *Procedure_ occurrence: Procedure_ concept_ id* (Activity) —— *Procedure_ occurrence: Procedure_ occurrence_ timestamp* (Timestamp) —— *Provider: Provider_ id* (Originator) —— *Procedure_ occurrence: Procedure_ type_ concept_ id*
Features	– Care site: *Hospital 1* – Patient type: *Infants (age < 3 yrs.)*

Also, the SQL query constructed from the event log specification is as follows. The following example is organized with user-specified values in the template and the fundamental form of the SQL query. In this way, it enables users to create clinical event logs with ease and convenience.

```
1: SELECT   Visit.Visit_occurrence_id as Case,
2:          ET.Procedure_occurrence_id as Event,
3:          ET.Procedure_concept_id as Activity,
4:          ET.Procedure_occurrence_timestamp as Timestamp,
5:          ET.Provider_id as Originator,
6:          Visit.Gender_concept_id as Gender,
```

```
7:              Visit.Year_of_birth as BirthYear,
8:              ET.Procedure_type_concept_id as ActivityType
9:  FROM    (
10:                 SELECT *
11:                 FROM    (
12:                             SELECT *
13:                             FROM Visit_occurrence LEFT JOIN Person
14:                             ON Visit_occurrence.Person_id = Person.Person_id
15:                         ) AS Visit_temp LEFT JOIN Provider
16:                 ON Visit_temp.Person_id = Provider.Provider_id
17:             ) AS Visit,
18:             (
19:                 SELECT *
20:                 FROM Procedure LEFT JOIN Provider
21:                 ON Procedure.Provider_id = Provider.Provider_id
22:             ) AS ET
23: WHERE   Visit.Visit_occurrence_id = ET.Visit_occurrence_id AND
24:             Visit.Visit_type_concept_id = 'Outpatient' AND
25:             Visit.Visit_start_date >= '01-Jan-2018' AND
26:             Visit.Visit_end_date <= '31-Jan-2018' AND
27:             Visit.Care_site_id = 'Hospital 1' AND
28:             Visit.Year_of_birth ≥ 2015
```

3.3 Data Preprocessing

Data preprocessing must be conducted to improve the accuracy and effectiveness of data analyses. Since the quality of data generated by information systems is generally an issue, it is essential to prepare enhanced data through data repair and noise removal. In a healthcare environment, the quality issue of the clinical event logs should be addressed.

There are four kinds of quality issues: *missing data*, *incorrect data*, *imprecise data*, and *irrelevant data* [141]. Missing data indicates that data is missing from logs, while incorrect data signifies that information recorded is not correct. Imprecise data represents that the level of data is too coarse, whereas irrelevant data means that information is not related at all with the log. These four types of quality issues are explicitly connected with the healthcare environment, and it needs to be processed thoroughly.

More in detail, one of the existing studies [142] suggested 11 event log imperfection patterns that can be considered as problems: *form-based event capture, inadvertent time travel, unanchored event, scatter event, elusive case, scattered case, collateral events, polluted label, distorted label, synonymous labels*, and *homonymous label*. The procedure for identifying and resolving

problems with these patterns can be effectively applied in clinical data preprocessing. For example, the *form-based event capture* pattern is one of the frequently occurred issues because multiple orders are recorded at the same time, as a result of the doctor's consultation activity. In such a case, it is required to determine the sequences of concurrent activities for the effective process analysis as a data repair method.

To this end, it is most likely to directly look through data for determining what problems a particular event log has. However, it requires a heavy burden to investigate the whole log by physical eye. As such, dotted chart [51], one of the process mining techniques, can be used to efficiently explore data. Dotted chart analysis provides a helicopter view of processes. Figure 12 shows a dotted chart. In the figure, while the most of cases have a complete sequence consisting of multiple events, it is identified that some of the cases below have an incomplete sequence; thus, it is connected to the *scattered case*. Besides, the distance between the red dots and the purple dots that represent clinical activities is considerable. In this regard, it is likely that data have problems with the *unanchored event*.

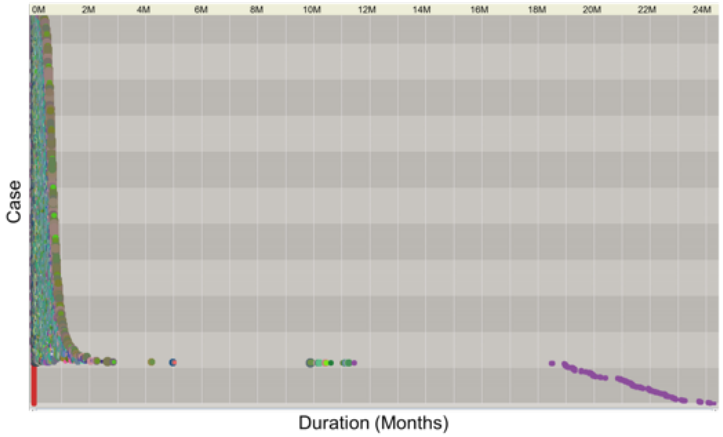


Figure 12. An example of application of dotted chart for data preprocessing

As presented earlier, it is indispensable to conduct data preprocessing that applies effective techniques or thoroughly verifies data with a heuristic approach.

3.4 Data Analysis

The data analysis phase is to understand the clinical processes with the preprocessed event logs. In this regard, it is necessary to derive meaningful analysis results based on numerous process mining algorithms currently developed. Here, it is impossible to elaborate on all process mining techniques because of the limited space and is not necessary because existing relevant studies allow such information to be obtained. Therefore, this chapter describes a structure for effective data analysis with a couple of main aspects associated with analyzing clinical data based on process mining.

The main aspects required for process mining analysis are as follows: clinical process types

and process mining types, as presented in Figure 13. That is, for a specific clinical process type, the data analysis is performed with process mining techniques based on the process mining types. For example, control-flow mining algorithms for outpatients are associated with the discovery of the control-flow.

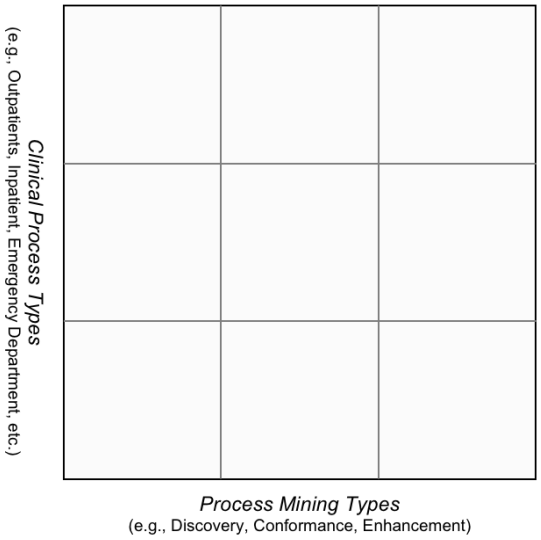


Figure 13. The main factors for process mining analysis in healthcare

First, clinical process types are essential to determine a specific data analysis. As depicted in Table 5, clinical processes in hospitals are composed of outpatients, inpatients, emergency departments, and clinical pathways. Here, the outpatient process is for patients whose all cares including consultation and tests are finished within a day, while the inpatient process is relevant to patients who stay for multiple days in hospitals. The process of emergency departments is the patient flows who visit hospitals when the office is closed. In performing data-driven process analysis, we are not able to conduct heterogeneous data analysis for multiple processes because a single event log is associated with a particular clinical process. Besides, each process type must have a different purpose. For example, one of the significant issues for the outpatient process is consultation waiting time, whereas for inpatients it is the length of stays. Therefore, it is required to determine a clinical process type depending on the goal of the data analysis.

Table 5. Clinical process types in data analysis

Clinical Process Types	Definition
Outpatients	Processes of patients whose all cares are finished within a day
Inpatients	Processes of patients who stay for multiple days in hospitals
Emergency Departments	Processes of patients who visit hospitals when the office is closed
Clinical Pathways	Processes of patients who receive the standardized guidelines

The other one is process mining types. As introduced in Chapter 2.2, process mining consists of three types: discovery, conformance, and enhancement. All three types are available for different purposes to understand and diagnose clinical processes. Discovery includes deriving process models with control-flow mining algorithms as well as social network mining and process pattern analysis. Conformance is associated with evaluating the discovered or reference process with the fitness with log replay or matching rate analysis. Lastly, enhancement includes performance analysis (e.g., bottlenecks) and rule generation with decision mining. Regarding the process mining types, users do not have to take a single type; thus, if needed, users can take multiple types. In other words, it is required to have a comprehensive understanding of clinical processes based on the numerous techniques involved in process mining types.

Table 6. Process mining types in data analysis

Process Mining Types	Relevant Analysis
Discovery	Control-flow mining algorithms, Social network miners, etc.
Conformance	Fitness with log replay, Matching rate analysis, etc.
Enhancement	Performance analysis, Rule generation with decision mining, etc.

3.5 Post-hoc Analysis

The last phase, i.e., post-hoc analysis, to transform the data analysis results into the insights based on the further analysis. In this regard, it is differentiated with three approaches: interpretation (diagnosis), redesign, and evaluation.

The interpretation approach is to clarify the data analysis results with professional knowledge by domain experts. Such a process serves as an essential step for narrowing the gap between the data analysis results and the practical access to the business. For example, hospitals can modify the reference models of cares or the clinical pathways that serve as a guideline with a thorough comprehension of the discovered clinical process models. Also, they may take a new plan for managing inpatients by verifying effectiveness of a derived predictive model. As such, this step is able to create the practical business plans or models based on data analysis results.

Also, the data analysis results can be converted into the experimental model for redesign. The redesign approach is to prepare a method for improving processes, and simulation is primarily utilized. In this regard, the status quo is to build a simulation model based on hand-crafted data. But, process mining enables to build a realistic simulation model by deriving simulation parameters based on the data analysis results. Therefore, we can build a realistic model that reflects the reality, not the ideal model. Chapter V introduces the details.

Lastly, the data analysis results can be utilized for evaluation by comparing them with the previous results. For example, it is required to compare the data analysis results before and after

the redesigns to identify process improvements. Chapter VI introduces the details.

3.6 Summary

This chapter presented the data analysis methodology for clinical processes in process mining. In the proposed methodology, the four phases were suggested including data preparation, data preprocessing, data analysis, and post-hoc analysis. In the data preparation phase, we introduced how to create clinical event logs from the common data model. Also, our framework contained data preprocessing with the dotted chart analysis. For the data analysis part, i.e., the core of the methodology, we provided two factors: clinical process types and process mining types. Finally, the post-hoc analysis proposed three orientations that derive insights from data analysis results: interpretation(diagnosis), redesign, and evaluation. The remaining chapters, i.e., Chapter IV, V, and VI, are presented the data analysis frameworks that cover from the beginning with data preparation to each post-hoc analysis, diagnosis, redesign, and evaluation, respectively.

IV Diagnosing Clinical Processes of Outpatients, Inpatients, and Clinical Pathways

This chapter presents data analysis frameworks for three clinical process types including *outpatients*, *inpatients*, and *clinical pathways*, based on the data analysis methodology for clinical processes presented in Chapter III. This chapter is organized as follows: Chapter 4.1 introduces the background of the research. Chapter 4.2, 4.3, and 4.4 provides the framework for analyzing clinical process processes with process mining for outpatients, inpatients, and clinical pathways, respectively. Chapter 4.5 demonstrates the effectiveness and usefulness of our methodology through in-depth evaluations with four different real-life logs. Finally, Chapter 4.6 summarizes this chapter.

4.1 Introduction

In a healthcare environment, clinical process analysis for diagnosis becomes essential with the data from hospital information systems. Process mining is a promising approach to understand and diagnose business processes with event logs data. As such, as described in Chapter 2.3, process mining has been already applied in healthcare. Those research has only focused on the development of the new algorithm, but it had a limitation of building a guideline for effective clinical process analysis. In other words, there was a lack of connecting the bridge with the practical use of innovative approaches in the research field, even though there has been a growing demand for improving understandability and usability of non-experts.

To overcome these limitations, this chapter proposes systematic clinical process analysis frameworks for *outpatients*, *inpatients*, and *clinical pathways* with a data-driven approach. For each category, we provide a specific goal and suitable analysis methods to achieve it. Furthermore, we show evaluation results with four real-life logs to validate the usefulness of the proposed framework.

In data science, effective data analysis is initiated with answering questions belonging to following four classes: reporting ("*What happened?*"), diagnosis ("*Why did it happen?*"), prediction ("*What will happen?*"), and recommendation ("*What is the best that can happen?*") [3]. The three analysis subjects presented before, i.e., outpatients, inpatients, and clinical pathways, also can be mapped onto the four classes by considering their objectives and analysis methods. Figure 14 provides the scope of the corresponding three categories.

First, the data analysis for outpatients visiting a hospital and staying within a single day is intended to understand their behaviors. Therefore, it focuses on *reporting* and *diagnosis*, and we use discovering a process model, process pattern analysis, performance analysis, and others (see Chapter 4.2 for details). Different from the outpatients, the goal for analyzing logs of inpatients who stay for a certain period of time in hospitals is more clear; it is required to understand and predict the length of stays (i.e., LOSs) of them. This is because the hospital length of stay

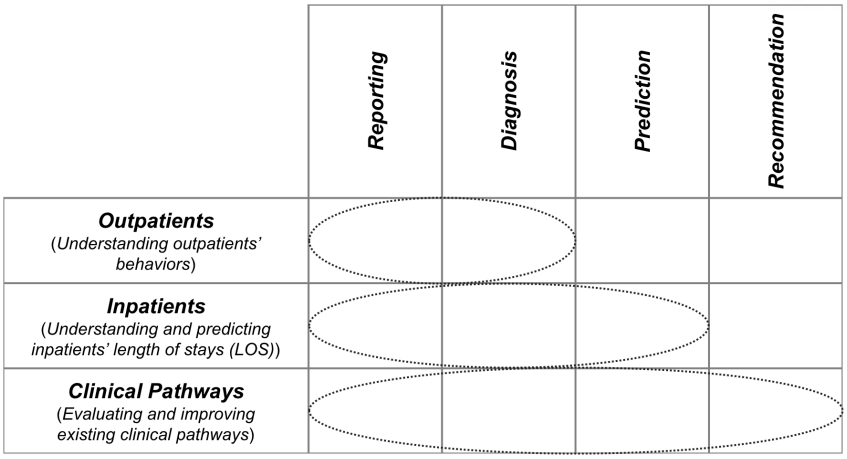


Figure 14. Three classes for clinical process analysis and the corresponding four analysis types

is a key indicator in an inpatient clinical process, which has a strong connection with costs, illness severity, complications, and profit margins [143, 144]. In the four types of data analysis, it is connected with *prediction* as well as *reporting* and *diagnosis*, and we use transfer pattern analysis, LOS performance analysis for identifying the factors affecting the LOS, and predicting inpatients' LOS as well (see Chapter 4.3 for details). Lastly, data analysis for clinical pathways (i.e., CPs) is connected with an aim of building a more standardized clinical process using a data-driven approach. Clinical pathways refer to the treatment guidelines in hospitals that provide standardized treatment methods and procedures for a particular disease or diagnosis, which has several advantages such as decreasing LOS, costs, and complications [145, 146]. Therefore, in this category, it is aimed to develop a quantitative approach to evaluate the existing CPs and to create improved CPs considering clinical data. As such, it covers all types of data analysis from *reporting* to *recommendation* (see Chapter 4.4 for details).

4.2 A Data Analysis Framework for Outpatient Processes

The data analysis for outpatients has an objective to understand the patients' behaviors from the outpatient event logs. Figure 15 depicts the overview of the data analysis framework for outpatients. The proposed framework was developed on the basis of the predefined data analysis methodology with data preparation, preprocessing, analysis, and post-hoc analysis. This chapter provides an explanation of the outpatient event logs, and five specific analysis methods: *discovering a process model, matching rate analysis, process pattern analysis, performance analysis*, and *root-cause analysis*.

4.2.1 Data Preparation & Preprocessing: Outpatient event logs

As introduced before, outpatients represent the patients whose all cares are finished within a day; thus, they do not receive any admission-related orders from doctors. Outpatient event logs are

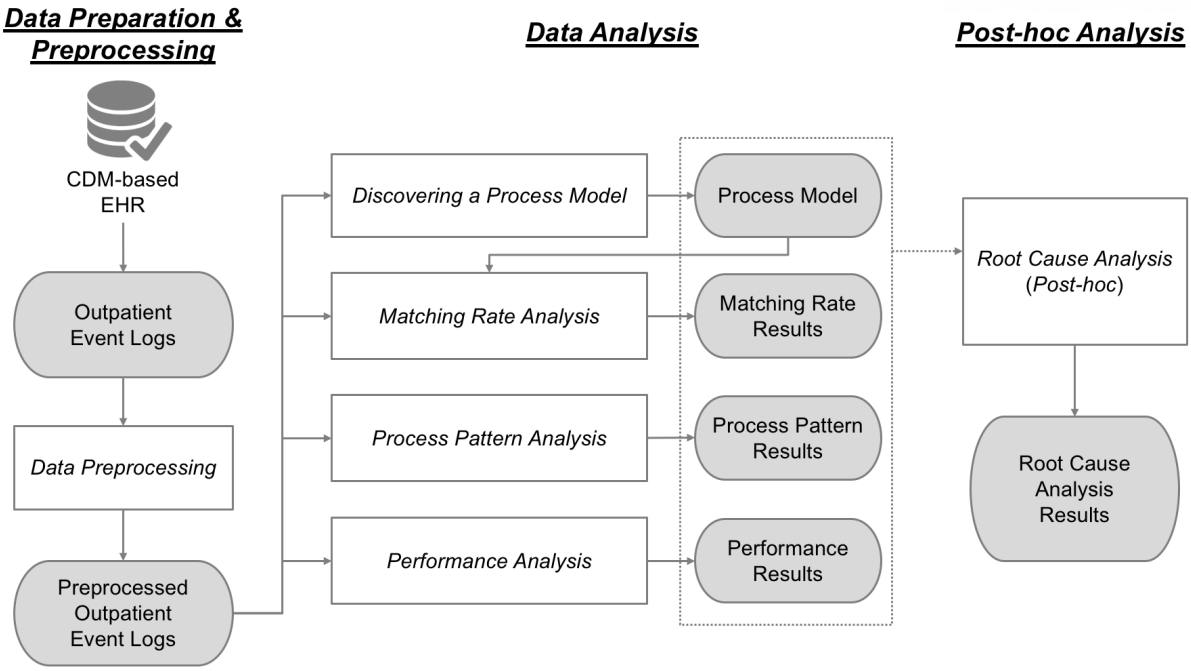


Figure 15. The detailed data analysis framework for outpatients

organized as shown in Table 7. Here, cases represent the visits of outpatients, not the outpatients. In other words, different visits of the same patients become the different cases because each visit is independent. Also, outpatient event logs include specific clinical activities as follows: sign on selective medical service, referral registration, outside image registration, payment, test registration, test, consultation registration, consultation, test scheduling, admission scheduling, outside-hospital prescription printing, in-hospital prescription receiving, certificate issuing, and treatment. Furthermore, care providers or completion time of activities are included in the event logs. As an example in Table 7, we can identify that case 1 who visited at 2018-01-01 received a series of clinical activities from registration to payment.

4.2.2 Data Analysis

Regarding the data analysis, we first tackle the connection between four specific analysis suggested in the framework with two aspects introduced in Chapter III. Figure 16 provides the detailed analysis methods for the outpatient process based on process mining types and clinical perspectives. First of all, deriving a process model is associated with the discovery and the control-flow perspective. In such a method, it aims at investigating orders of outpatient clinical activities, dominant or abnormal flows in the outpatient process. Besides, for the discovery type, the process pattern analysis is conducted with the patients perspective. The matching rate analysis is relevant to the conformance type and the control-flow. It is applied to evaluate the existing reference model by comparing with the discovered model and build an improved model. Lastly, performance analysis with the activities perspective is also included in the proposed framework.

Table 7. A partial example of outpatient event logs

Case	Event	Activity	Originator	Timestamp
Case 1	E1	Registration	Paul	2018-01-01 09:00
Case 1	E2	Consultation Registration	Allen	2018-01-01 09:20
Case 1	E3	Consultation	Mike	2018-01-01 10:00
Case 1	E4	Test Registration	Tim	2018-01-01 10:10
Case 1	E5	Test	Sara	2018-01-01 10:30
Case 1	E6	Consultation Scheduling	Lauren	2018-01-01 10:35
Case 1	E7	Payment	Mason	2018-01-01 10:40
Case 2	E8	Registration	Paul	2018-01-01 09:05
Case 2	E9	Test Registration	Tim	2018-01-01 09:10
Case 2	E10	Test	Sara	2018-01-01 09:25
Case 2	E11	Consultation Registration	Allen	2018-01-01 09:35
Case 2	E12	Consultation	Chris	2018-01-01 10:00
Case 2	E13	Treatment	Emily	2018-01-01 10:20
Case 2	E14	Payment	Mason	2018-01-01 10:30

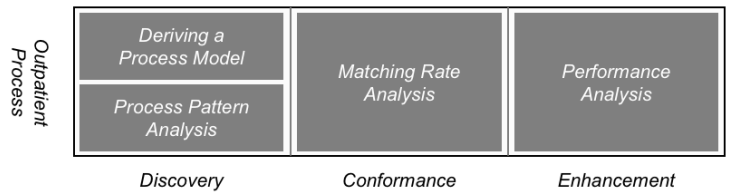


Figure 16. The detailed methods of the data analysis for outpatients

4.2.3 Data Analysis: Discovering a process model

As explained before, discovery aims to produce a model from event logs without a-priori information. Here, we suggest *frequency mining*, which produces a process map (A_L, R_L) based on directly-follows relationships between activities in event logs [40]. Frequency mining is defined as follows.

Definition 2 (Frequency Mining) *Note that frequency mining constructs a process map from an event log L, which is described by a tuple (A_L, R_L). Frequency mining is defined as follows:*

- A_L is the set of activities in a log L;

- R_L is the set of relations between two activities in a log L. A relation $r_{ij} = \{(a_i, a_j)|a_i, a_j \in A \wedge a_i > a_j\}$ is an element of R_L, where $a_i > a_j$ represents a notable directly-followed relationship (i.e., a_j is the direct successor of a_i) that has a higher frequency than a pre-determined threshold value.

According to this technique, if there is a relationship between activity A and activity B, then node A and B are connected by an arc in the process model. Frequency mining has the compelling advantage that it is able to include all paths in a process model (e.g., with zero threshold value). Therefore, it is a better way than other mining methods to apply in the healthcare domain, because all patient paths can be practical and meaningful in a hospital.

4.2.4 Data Analysis: Matching rate analysis

In hospitals, there can exist a reference model considered as a standard outpatient process model developed by domain experts [78]. The example of the reference model is presented in Figure 17, taken from [40]. Even though the reference process model is constructed from clinical professionals, the actual logs may not be matched well with the reference in general. Therefore, we need to compare the discovered process model and reference model. In such a process, we develop a quantitative approach (i.e., *matching rate*) that measures the difference between the model and log.

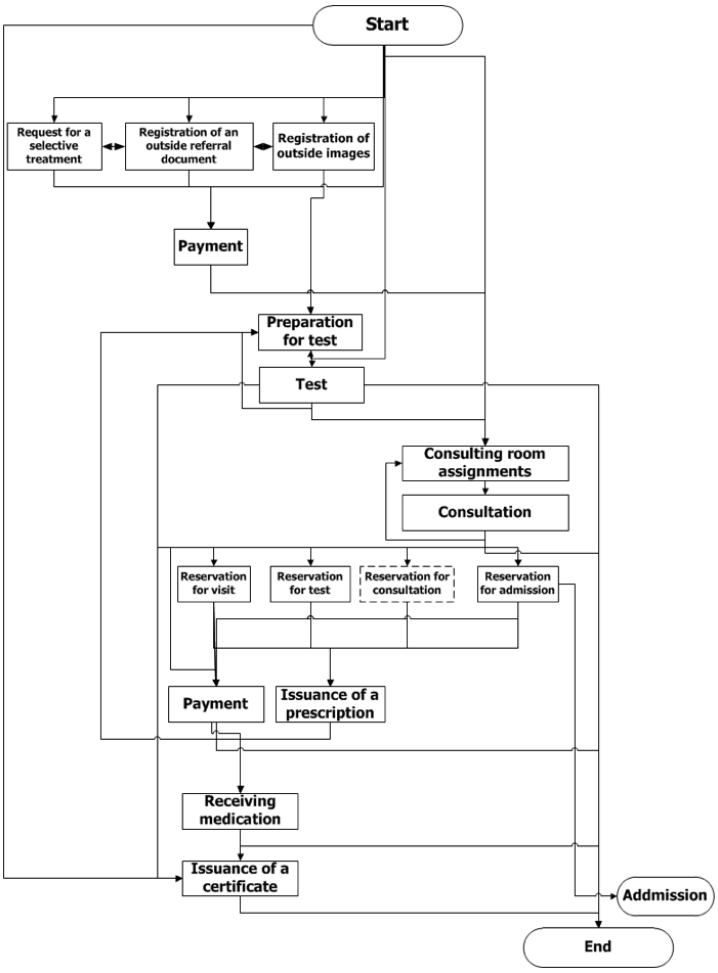


Figure 17. An example of the clinical reference process model

Before introducing how to measure the matching rate, we first define *standard activity relations* and a *matching function* provided in Definition 3. Assume that there exists a reference model of the clinical process in an organization. Standard activity relations are defined as the causal activity relations identified in the reference model. The matching function serves as a role to compare relations between the reference model and log. More in detail, the matching function returns true if an activity relation in an event log is involved in standard activity relations of a reference model, and false otherwise. Figure 18 provides a matching example of a standard model and a log. In the figure, the reference process is $A \rightarrow B \rightarrow C \rightarrow D$, which includes three relations: $(A, B), (B, C)$, and (C, D). The event log contains three variants, 18 cases, and four types of activity relations: $(A, B), (B, C), (C, D)$, and (D, A). Among the activity relations from the log, only first three items accord with the standard relations, while (D, A) has no counterpart in the reference model.

Definition 3 (Standard Activity Relation (SAR), matching) *Let Standard Activity Relation(SAR)* $\subseteq A \times A$ *be a set of standard activity relations where two events have a causal relation. Let* $M_{ar} = \{matched, nonmatched\}$ *be a set of matching results of activity relations.*

- matching: $ar_k \rightarrow M_{ar}$ *is a function testing whether each activity relations are mapped onto the standard activity relations.*

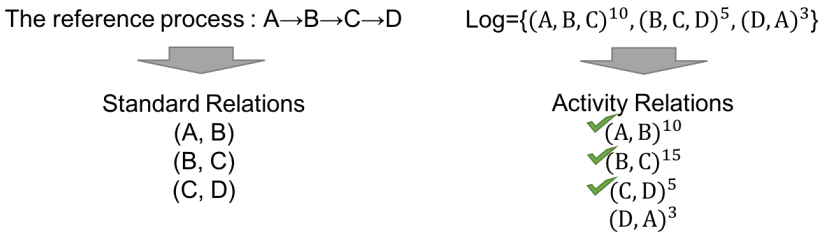


Figure 18. A matching example of relations between the reference model and log

Based on the predefined function, we define the matching rate (MR_{ar}) as described in Definition 4. It represents the number of activity relations for which the matching function evaluates to true, divided by the number of activity relations.

Definition 4 (Matching rate) *Let* MR_{ar} *be the matching rate of the process model.*

$$\text{- } MR_{ar} = \frac{\sum_{0<k<|c|} \sum_{0<i<j\leq n} \begin{cases} 1 & c_k \in L \wedge e_{k,i}, e_{k,j} \in c_k \wedge e_{k,i} > e_{k,j} \\ & \wedge matching(ar_{k,ij}) = matched \\ 0 & otherwise \end{cases}}{\sum_{0<k<|c|} \sum_{0<i<j\leq n} \begin{cases} 1 & c_k \in L \wedge e_{k,i}, e_{k,j} \in c_k \wedge e_{k,i} > e_{k,j} \\ 0 & otherwise \end{cases}}$$

According to Definition 4, in the above example, among 33 activity relations, 30 activity relations are matched with the standard relations; thus the matching rate becomes 0.91 (i.e., 30/33).

4.2.5 Data Analysis: Process pattern analysis

In general, the extracted outpatient process takes the form of "*spaghetti*" [1, 147]. Unlike a "*lasagna*" process, the spaghetti process is unstructured and complicated, thus it is hard to identify the frequent patterns. For this reason, the process pattern analysis is required to comprehend major workflows. The process pattern analysis starts from making groups of event traces of patients. Here, an event trace represents a sequence of activities which are performed to a specific patient in the hospital. After grouping the traces, patients who belong to each group are counted to find the preeminent patterns. In addition, a couple of statistics are applied to get fundamental performance information such as average, median, and standard deviation for the length of stay, and the number of events. Through the statistical information, we can get several findings including which pattern is considered as the major or abnormal workflow.

By conducting the process pattern analysis from the whole event log, we can identify frequent or abnormal patterns among all of them. Moreover, it can be applied as one of the comparative data analysis by dividing the whole event log according to the other attributes such as resource, timestamp, and additional information for patients. For example, we can use the process pattern analysis result for each department to make a reference model for each one. Besides, we can provide a proper guideline for patients by extracting major workflows using attributes for patients. As such, the process pattern analysis is essential in the healthcare process field.

4.2.6 Data Analysis: Performance analysis

The performance analysis using process mining can be used to measure various process performance indicators (*PPIs*) of the clinical process for outpatients. PPIs are a set of measures that become critical to the success of an organization [148]. It covers a wide range of perspectives such as patients, activities, and care providers. More in detail, PPIs can be defined and calculated with entity types, entity identifiers, measures, and aggregation functions. The entity types are features to measure the performance, which include activities and originators. The entity identifiers signify the possible values that belong to the entity type. For example, *test*, *consultation*, and *payment* become the entity identifiers of the entity type *activities*. Based on these two elements and the log (i.e., the event log, the entity type, and the entity identifier), required events are filtered and extracted through the ψ function. After that, extracted events are calculated based on measures such as count, working time, and waiting time. A PPI ($A(M(\mathcal{E}))$) is defined as applying aggregation functions (A) from computed measure values for the filtered events. The PPI is defined as follows.

Definition 5 (Process Performance Indicators) *Let T and V be the universe of entity types and universe of possible values, respectively. For each entity type $t \in T$, V_t denotes the set of possible values, i.e., the set of entity identifiers of type t. Let $\psi \in L \times T \times V_T \Rightarrow \mathcal{E}$ is a function that finds out the set of events from an event log for a given entity type and an entity identifier (where, $\mathcal{E}$ is the set of events). M represents the measures such as count (count), working time (working), waiting time (waiting), duration (duration), etc. $A(M(\mathcal{E}))$ is the process performance indicator from an event log for a given entity type and an entity identifier, and a measure (where, A be the aggregation function). Note that A be the average, median, standard deviation, quantile, and percentile.*

Here, we provide a couple of examples of the patient, activity, and originator perspectives. Table 8 presents the examples of PPIs and their formalization. Note that they do not consider the aggregation functions.

Table 8. The examples of PPIs for the performance analysis

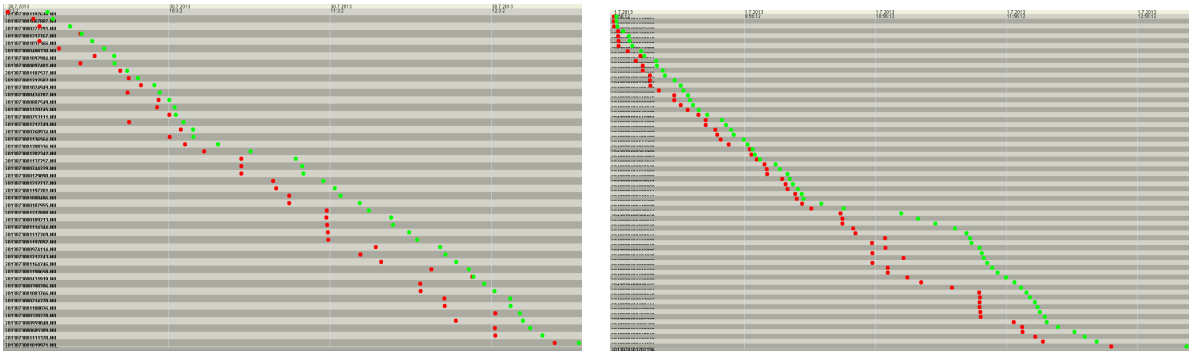
Perspective	PPI	Formalization
Patient	Length of stay for outpatients	$duration(\psi(L, \emptyset, \emptyset))$
	Number of events for a specific patient (c_1)	$count(\psi(L, Case, c_1))$
Activity	Waiting time for an activity (a_1)	$waiting(\psi(L, Activity, a_1))$
	Frequency of an activity (a_1)	$count(\psi(L, Activity, a_1))$
Originator	Working time of a specific originator (o_1)	$working(\psi(L, Originator, o_1))$
	Frequency of an originator (o_1)	$count(\psi(L, Originator, o_1))$

4.2.7 Post-hoc Analysis: Root Cause Analysis

The next step of the data analysis for outpatients is required to interpret the results and investigate the insights through discussions with own or domain experts. In such a process, if needed, it is necessary to identify the root-cause, i.e., post-hoc analysis. One of the typical tools, dotted chart analysis [51], provides a helicopter view of processes and has a strength that can diagnose the data of event logs at a glance.

The tool can be utilized as follows. Assume that we identify the consultation waiting time of a particular care-provider is significantly high through the performance analysis. To this end, the dotted chart with only events that relate to consultation and its previous event can be a solution to investigate such a cause. Figure 19 provides the examples of the dotted chart, where red and green dots represent consultation registration and consultation, respectively. In the figures, the y-axis and the x-axis are configured as patients and actual time, respectively, and the rows are sorted by consultation. In case of Figure 19a, the reversal of consultation is

considered as a reason for the increase of waiting time. It means that the patient who completed the previous work earlier is treated later than the latter. Different from Figure 19a, the sudden increase of working time gives rise to the long delay for consultation in Figure 19b. As such, post-hoc analysis with dotted chart can help to derive meaningful insights.



(a) Reversal of consultation (b) Sudden increase of working time

Figure 19. An example of the dotted charts for post-hoc analysis

4.3 A Data Analysis Framework for Inpatient Processes

The data analysis framework for inpatients aims at comprehending and predicting length of stays (LOS). The data analysis part is composed of six analysis methods: *LOS performance analysis, LOS analysis in terms of transfer patterns, LOS analysis in accordance with diagnosis, analysis for long-term hospitalization patients, deriving correlated factors on LOS*, and *building a predictive model of patient's LOS*. Figure 20 depicts the overview of the data analysis framework for inpatients.

4.3.1 Data Preparation & Preprocessing: Inpatient event logs

Inpatient event logs are associated with the data for patients who stay for multiple days in hospitals by receiving high degree of care, e.g., surgery. Table 9 provides a partial example of inpatient event logs. In the example, each case represents a single patient; thus, 10 clinical events for four days are relevant to the single process instance, i.e., Case 1. Also, inpatient clinical activities include admission, treatment, surgery, procedure, transfer, antibiotics, and discharge. Compared to the outpatient process logs, most of activities is different. Also, it can have department information because one of the key activities for inpatients is transfers of departments.

4.3.2 Data Analysis

Figure 21 provides the detailed analysis methods for the inpatient process based on process mining types. The proposed framework for analyzing inpatient data has five detailed data anal-

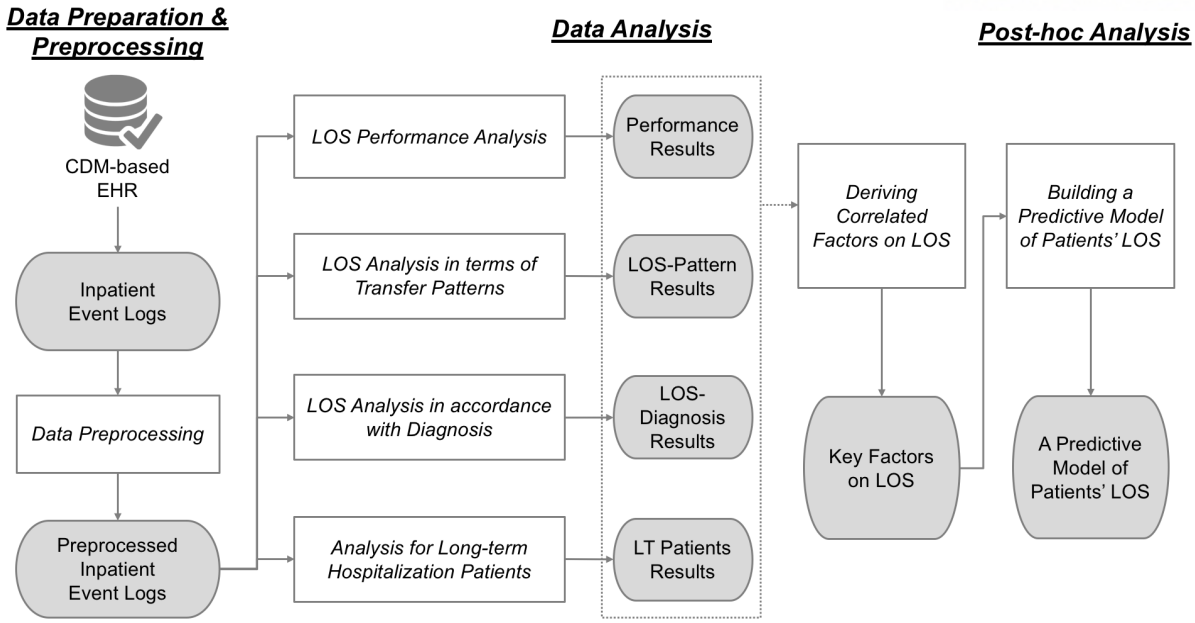


Figure 20. The detailed data analysis framework for inpatients

Table 9. A partial example of inpatient event logs

Case	Event	Activity	Originator	Department	Timestamp
Case 1	E1	Admission	Paul	Dept A	2018-01-01 15:00
Case 1	E2	Treatment	Allen	Dept T	2018-01-02 11:00
Case 1	E3	Treatment	Mike	Dept T	2018-01-03 10:00
Case 1	E4	Surgery	Tim	Dept S	2018-01-03 11:00
Case 1	E5	Antibiotics	Sara	Dept A	2018-01-03 13:00
Case 1	E6	Antibiotics	Lauren	Dept A	2018-01-03 14:00
Case 1	E7	Treatment	Mason	Dept T	2018-01-03 19:00
Case 1	E8	Antibiotics	Paul	Dept A	2018-01-04 09:00
Case 1	E9	Treatment	Tim	Dept T	2018-01-04 10:00
Case 1	E10	Discharge	Sara	Dept D	2018-01-04 11:00

ysis methods. In this regard, all methods are associated with the performance analysis of the whole clinical perspectives; thus, it is engaged in the enhancement type and all different clinical perspectives.

4.3.3 Data Analysis: LOS performance analysis

As far as the performance analysis for the length of hospital stays of inpatients, we employ the proposed performance analysis method (i.e., Definition 5) in Chapter 4.2. As described in

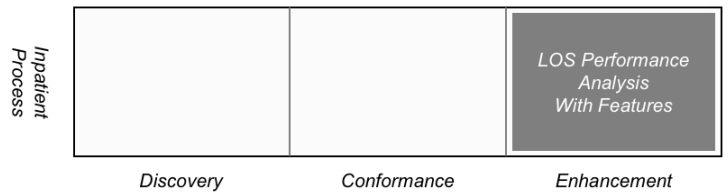


Figure 21. The detailed methods of the data analysis for outpatients

the Definition 5, the performance analysis for LOS also can be performed using the ψ function extracting relevant events, the measure function (M), and the aggregation function (A). A typical example is an indicator for understanding distribution of LOS for whole patients in a log ($duration(\psi(L, \emptyset, \emptyset))$). Also, it is necessary to measure LOS by the department because it is one of the essential factors associated with LOS ($duration(\psi(L, department, d_1))$). As such, it is required to define and measure a couple of proper user-specified metrics such as LOS by patient type or time-specific LOS that are appropriate with the objectives of data analysis for inpatients.

4.3.4 Data Analysis: LOS analysis in terms of transfer patterns

The transfer is defined as a change of department required by the patient's condition and one of the factors associating with the processing of hospitalized patients. In such a process, there are complex factors that increase the days of hospitalization including transfer waiting time and additional lab tests. Therefore, LOS-related delineated analysis for each pattern as well as the comparative analysis based on transfer are required.

To this end, it employs the process pattern analysis provided in Chapter 4.2. That is, it uses a process mining technique to extract the transfer pattern and measure the performance information, such as the frequency of each pattern, the average time required, and the median time required. Here, dissimilar to the process pattern analysis for outpatients, the transfer pattern is constituted based on the department information associated with care providers.

4.3.5 Data Analysis: LOS analysis in accordance with diagnosis

This analysis method aims at identifying a difference in LOS by diagnosis. In this regard, we employ the *Z-score*, i.e., standard score [149]. It is postulated that analyzing the absolute LOS for all diagnoses is of limited value because there are considerable differences in the required LOS according to the specific diagnoses. In order to overcome this challenge, the relative LOS needs to be measured and compared, by deriving the standard score of each diagnosis based on the mean and standard deviation of the LOS per department. This is demonstrated as follows.

Definition 6 (Standard score (Z-score)) *Let LoS signify the length of stay of patients.* o_{LoS}, μ_{LoS}, *and* σ_{LoS} *represents the observed value, the expected mean, and the standard deviation of*

LoS, respectively.

- $Z\text{-}score = \frac{o_{LoS} - \mu_{LoS}}{\sigma_{LoS}}$

4.3.6 Data Analysis: Analysis for long-term hospitalization patients

Long-term care is one of the process types according to the definition of the common data model. Thus, long-term hospitalization patients can be separately analyzed as one of the analysis categories. But, generally, long-term patients are defined as inpatients who have been hospitalized for over 30 days. Therefore, there is no significant difference in the process, sub-process, and even individual activity levels between inpatients and long-term patients. For this reason, analysis for long-term hospitalization patients is included as one of the analysis methods for inpatients.

This analysis conducts a comparative analysis of long-stay patients and others with an aim to identify the characteristics of them. Here, we employ the hypothesis testing method with a statistical approach. In such a process, following comparative items are considered for an effective comparison: average LOS, the rate for surgical patients, the number of surgeries per patient, the rate for transferred patients, the rate for patients on antibiotics treatment, the number of antibiotics per patient (total or in a day), and the number of procedures per patient (whole or in a day).

4.3.7 Data Analysis: Deriving correlated factors on LOS

Before constructing a predictive model for LOS, it is necessary to identify fundamental factors on LOS for better prediction. Thus, this analysis method aims to derive directly-correlated factors on LOS, which serves as inputs for a further step. In this step, a hypothesis testing approach including the *student's t-test* and *analysis of variance (ANOVA)* are employed for investigating relationships between LOS and patient-related shreds of evidence including transfer time, discharge delay time, surgery frequency, diagnosis frequency, severity, bed grade, and insurance type. For each analysis, it has the same null (H_0) and alternative hypotheses (H_1) as follows.

- H_0: The means of all groups under consideration are equal
- H_1: The means are not all equal

If the null hypothesis is accepted, the relevant information is removed from the inputs for building a predictive model, whereas it is connected to the next phase when rejected.

4.3.8 Post-hoc Analysis: Building a predictive model of patient's LOS

The last stage of data analysis for inpatients is to build a prediction model for their LOS. In this regard, we employ a couple of machine learning techniques such as data partitioning, prediction and classification algorithms, and evaluation methods. First, data is partitioned into the training and test set. Similar to the conventional methodology in data mining, the training set is used to build a model, whereas the test set is applied to validate the model and measure the accuracy

of that. As far as constructing a model is concerned, we can have two options to predict the LOS of patients or to classify whether patients belong to the long-term patient group or not. As such, we apply evaluation measures including mean absolute percentage error (MAPE) [150] as presented in Definition 7 for the prediction model or a confusion matrix for the classification model.

Definition 7 (Mean Absolute Percentage Error) $MAPE = \frac{1}{n}\sum_{t=1}^{n}|\frac{A_t - F_t}{A_t}|$, *where* A_t *is the actual value calculated from the logs, and* F_t *is the predicted value derived from the model.*

4.4 A Data Analysis Framework for Clinical Pathways

The data analysis methodology for clinical pathways (CPs) pursues a goal of evaluating and improving existing CPs. To this end, we employ two different features: clinical pathways and CP event logs. Also, a series of four detailed analyses are included in the proposed methodology: *comparing CP orders and logs, CP matching rate analysis, feature-based CP analysis*, and *building an improved CP*. Figure 22 presents the proposed framework.

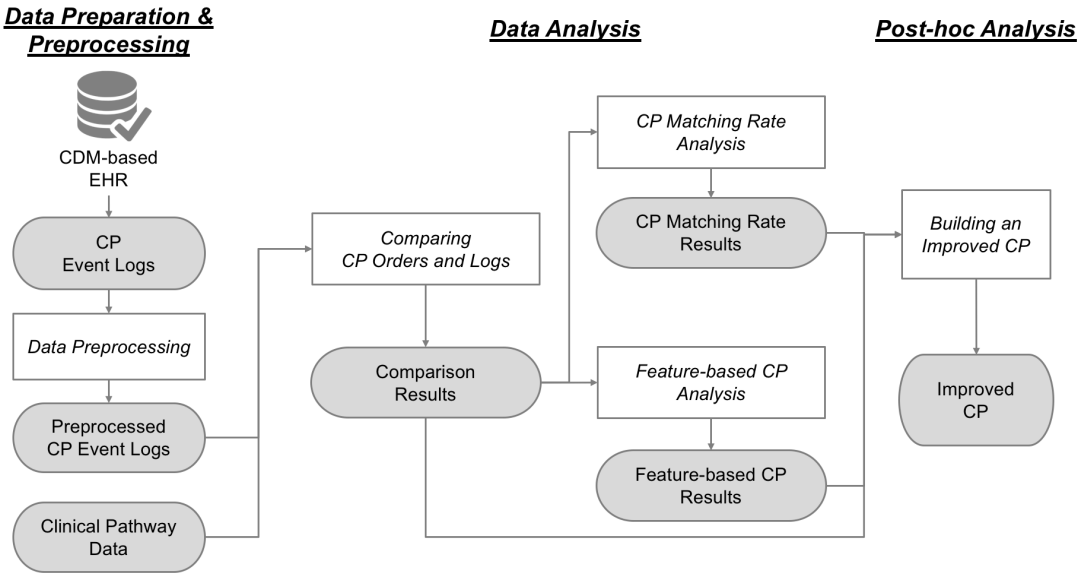


Figure 22. The detailed data analysis framework for clinical pathways

4.4.1 Data Preparation & Preprocessing: Clinical Pathway data and event logs

First, we define clinical pathways as an initial process. The formal definition is given as follows.

Definition 8 (Clinical Pathway, Orders) *Let CP and O be the clinical pathway and the order universe, respectively. Let AN be a set of attribute names, then* $o \in O, cp \in CP$ *and name* $n \in AN$*:* $\pi_n(o)$ *is the value of attribute n for order o. Orders can have following attributes: cp, code, day, stage, and order type;* $o_i = \{cp_i, code_i, day_i, stage_i, type_i\}$*. Then,* $\pi_{cp}(o) =$

$cp, \pi_{coder}(o) = code, \pi_{day}(o) = day, \pi_{stage}(o) = stage,$ *and* $\pi_{type}(o) = type$ *for some order* $o = (cp, code_i, day, stage, type)$.

As defined in Definition 8, CPs are comprised of orders which have five attributes such as CP code, activity, stage, day, and activity type. Table 10 provides a simple artificial CP. The CP with regard to Code A1 is comprised of a total of 8 orders and takes a total of four days. It processes two activities on the first day, three activities on the second day, one activity on the third day, and two activities on the fourth day. It is comprised of a total of five order types, which are *Treatment*, *Test*, *Medication*, *Injection*, and *Diet*. Besides, the stage in which each activity is processed is divided into *Pre OP*, *OP*, *Post OP*, *Normal Order*, and *Discharge*.

Table 10. An partial example of clinical pathways

CP ID	Order Code	Day	Stage	Order type
A1	T1	1	Pre OP	Treatment
A1	Te1	1	Pre OP	Test
A1	M1	2	OP	Medication
A1	T2	2	OP	Treatment
A1	Te1	2	Post OP	Test
A1	I1	3	Normal Order	Injection
A1	Di1	4	Normal Order	Diet
A1	M2	4	Discharge	Medication

Also, we prepare a CP event log, which represents a collection of inpatients who receive a specific CP. As such, the overall format is quite similar to the inpatient event logs. For example, following activities are commonly included in the logs: treatment, test, medication, admission, and discharge. But, more fine-grained clinical events are also included in the logs to perform mapping with orders in clinical pathways. That is, the detailed codes are utilized such as T1, Te1, and M1.

Table 11. A partial example of CP event logs

Case	Event	Activity	Originator	Order type	Timestamp
Case 1	E1	T1	Paul	Treatment	2018-01-01 (day 1)
Case 1	E2	Te1	Allen	Test	2018-01-02 (day 2)
Case 1	E3	M1	Mike	Medication	2018-01-03 (day 3)
Case 1	E4	T2	Tim	Treatment	2018-01-03 (day 3)
Case 1	E5	Te2	Sara	Test	2018-01-03 (day 3)
Case 1	E6	T3	Lauren	Treatment	2018-01-03 (day 3)
Case 1	E7	T4	Mason	Treatment	2018-01-03 (day 3)
Case 1	E8	T5	Paul	Treatment	2018-01-04 (day 4)
Case 1	E9	M2	Tim	Medication	2018-01-04 (day 4)
Case 1	E10	T6	Sara	Treatment	2018-01-04 (day 4)

4.4.2 Data Analysis

Figure 23 provides the detailed analysis methods for the clinical pathways based on process mining types. First, CP matching rate analysis is associated with the conformance and the control-flow perspective. It aims at evaluating the existing clinical pathways with CP event logs. The feature-based CP analysis is also engaged in the conformance type, but it is applied with the activities perspective. Finally, building an improved CP is a method with the discovery and the control-flow perspective.

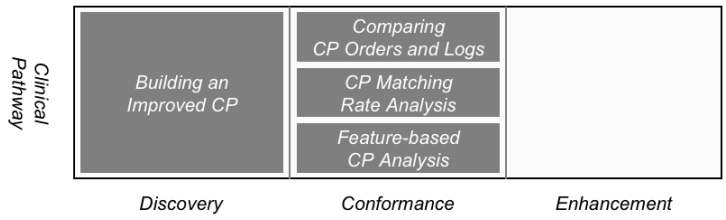


Figure 23. The detailed methods of the data analysis for outpatients

4.4.3 Data Analysis: Comparing CP orders and logs

On the basis of CP orders and event logs, first of all, we compare the activities of events in the event log with orders in the CP. Afterwards, we also identify whether the three attributes (i.e., day, stage, and activity type) of events are identical with them of events. If two conditions are satisfied, we consider them as "*Matching*"; otherwise they become "*Non-matching*". By applying this approach to events of each patient, we are able to build the matching and the non-matching sets for every patient.

These results can be expressed as numerical values with a quantitative approach as defined in Definition 9.

Definition 9 (Indicators for comparing CP orders and logs) *Let N_{cp}, M_{cp}, N_{log}, and R_{log} be the number of orders and non-matched orders in a specific clinical pathway (e.g., cp_1), and the number of events and non-matched events in the CP event log, respectively.*

- $N_{cp} = \sum_{0 \le i \le |o|} \begin{cases} 1 & \text{if } \pi_{cp}(o_i) = cp_1 \\ 0 & \text{otherwise} \end{cases}$

- $M_{cp} = N_{cp} - \sum_{0 \le k \le |c|} \sum_{0 \le j \le |e|} \begin{cases} 1 & \text{if } \exists_{0 \le i \le |o|} \pi_{cp}(o_i) = cp_1 \\ & \quad \wedge \pi_{cp}(e_i) = cp_1 \\ & \quad \wedge \pi_{code}(o_i) = \pi_{act}(e_j) \\ & \quad \wedge \pi_{day}(o_i) = \pi_{ctime}(e_j) \\ & \quad \wedge \pi_{stage}(o_i) = \pi_{stage}(e_j) \\ & \quad \wedge \pi_{type}(o_i) = \pi_{type}(e_j) \\ 0 & \text{otherwise} \end{cases}$

- $N_{log} = \sum_{0 \le j \le |e|} \begin{cases} 1 & \text{if } \pi_{cp}(e_i) = cp_1 \\ 0 & \text{otherwise} \end{cases}$

- $R_{log} = N_{log} - \sum_{0 \le k \le |c|} \sum_{0 \le j \le |e|} \begin{cases} 1 & \text{if } \exists_{0 \le i \le |o|} \pi_{cp}(o_i) = cp_1 \\ & \quad \wedge \pi_{code}(o_i) = \pi_{act}(e_j) \\ & \quad \wedge \pi_{day}(o_i) = \pi_{ctime}(e_j) \\ & \quad \wedge \pi_{stage}(o_i) = \pi_{stage}(e_j) \\ & \quad \wedge \pi_{type}(o_i) = \pi_{type}(e_j) \\ 0 & \text{otherwise} \end{cases}$

First of all, N_{cp} signifies the number of orders included in the CP. M_{cp} refers to the number of events that are included in the CP, but that does not show up in the event log. In other words, M_{cp} refers to *non-matched* orders in the CP. Also, N_{log} is the number of events included in the event log, whereas R_{log} refers to the number of events that are recorded in the event log but not included in the CP. It can be considered as the number of *non-matched* events in the event log.

4.4.4 Data Analysis: CP matching rate analysis

Based on the four numerical matching results defined in the previous chapter, we specify the application rate of orders in the CP (AR_{cp}) and matched ratio of events in the event log (MR_{log}). The AR_{cp} signifies the percentage of orders in the CP that is used in the event log, where uses N_{cp} and M_{cp}. The MR_{log} refers to the percentage of the events in the event log that are included in the CP. In other words, the MR_{log} shows how small the number of additional three orders

are. To calculate the MR_{log}, we use N_{log} and R_{log}. Definition 10 elaborates on the AR_{cp} and MR_{log} respectively.

Definition 10 (Application rate, Matched ratio of events) *Let AR_{cp} and MR_{log} be the application rate for orders in the CP and the matched ratio of events in the log, respectively.*

- $AR_{cp} = 1 - \frac{M_{cp}}{N_{cp}}$
- $MR_{log} = 1 - \frac{R_{log}}{N_{log}}$

We also suggest how to calculate the CP order matching rate from the application rate of orders in the CP (AR_{cp}) and matched ratio of events in the event log (MR_{log}) that were defined earlier. Definition 11 elaborates on the matching rate. CMR signifies the CP order matching rate, here, users have to specify the weights for AR_{cp} and MR_{log}. In this regard, the standard matching rate ($SCMR$) is defined in determining two weights as the same values (i.e., 0.5).

Definition 11 (CP order matching rate) *Let CMR be the CP order matching rate, and w_1 and w_2 are the weight for AR_{cp} and MR_{log}, respectively. Here, the CP order matching rate becomes the standard (i.e., SCMR) if both w_1 and w_2 are 0.5.*

- $CMR = w_1 \times AR_{cp} + w_2 \times MR_{log}$ *(where* $w_1 + w_2 = 1$*)*
- $SCMR = \frac{1}{2} \times AR_{cp} + \frac{1}{2} \times MR_{log}$

4.4.5 Data Analysis: Feature-based CP analysis

This chapter gives how to conduct a performance analysis (e.g., application rates and matching rates) based on features included in the CP. Thus, it has a goal for identifying potential improvement points by evaluating them with each feature. More in detail, users determine the connection of features to high or poor performances and the causes of the poor performance. For instance, the followings may be derived, and these findings are considered in improving the existing CP.

- Orders in the preliminary stage cannot be higher because whether or not to perform is flexible.
- Medication orders in the discharge stage are exposed depending on the patient's condition.

4.4.6 Post-hoc Analysis: Building an improved CP

The last step for CP data analysis is to build an improved CP for better cares. Algorithm 1 provides how to create a new CP with the application rate. The suggested algorithm is relatively simple to use. Among the orders in CP, first of all, when the application ratio (ar_{o_i}) falls below the threshold (eth), those orders are deleted from the newly-built CP (ICP). On the contrary,

additional events that are applied to many patients ($ar_{a_i} \geq ith$) are included in the improved CP (ICP). Then, the new CP becomes a user-centric guideline that bases on the behaviors of patients.

Algorithm 1 *BuildingImprovedCP*(L, CP, eth, ith)

Input Event Log L;
Clinical Pathway CP;
Excluding Threshold eth;
Including Threshold ith;
Output The improved CP ICP;

Let $appliedR : o_i$ or $a_i \rightarrow \mathbb{N}$ be the function calculating the application rate for an order (i.e., o_i) or an activity (i.e., a_i).

1: $ICP \leftarrow CP$
2: **for all** orders o_i in the clinical pathway CP **do**
3: $ar_{o_i} \leftarrow appliedR(o_i)$
4: **if** $ar_{o_i} \leq eth$ **then**
5: $NCP \leftarrow NCP \setminus \{o_i\}$
6: **end if**
7: **end for**
8: **for all** activities a_i in the log L **do**
9: **if** a_i does not exist in CP (i.e., $a_i \notin CP$) **then**
10: $ar_{a_i} \leftarrow appliedR(a_i)$
11: **if** $ar_{a_i} \geq ith$ **then**
12: $NCP \leftarrow NCP \cup \{a_i\}$
13: **end if**
14: **end if**
15: **end for**
16: **return** ICP

4.5 Evaluation

This chapter shares the evaluation results for validating the proposed data analysis methodology. Chapter 4.5.1 and 4.5.2 provides the case study results for outpatients, and Chapter 4.5.3 is relevant with the inpatients. Also, Chapter 4.5.4 considers the validation of the data analysis methodology for clinical pathways.

4.5.1 Evaluation for outpatients data analysis 1

In this chapter, we validate our outpatient process analysis methodology. To this end, a one-month event log for outpatients was prepared from hospital information systems at a tertiary general hospital in Korea. The event log contained 15 types of activities: sign on selective medical services, referral registration, outside image registration, payment, test registration, test, consultation registration, consultation, consultation scheduling, test scheduling, admission scheduling, outside-hospital prescription printing, certificate issuing, and treatment. The summary information of the prepared event log is as follows:

- Approximately 120,000 cases (patients that were treated)
- Approximately 700,000 events (activities performed for outpatients)
- 15 different activities (e.g., consultation, test, payment, and others)

This evaluation focuses on discovering a process model and identifying process patterns for outpatients.

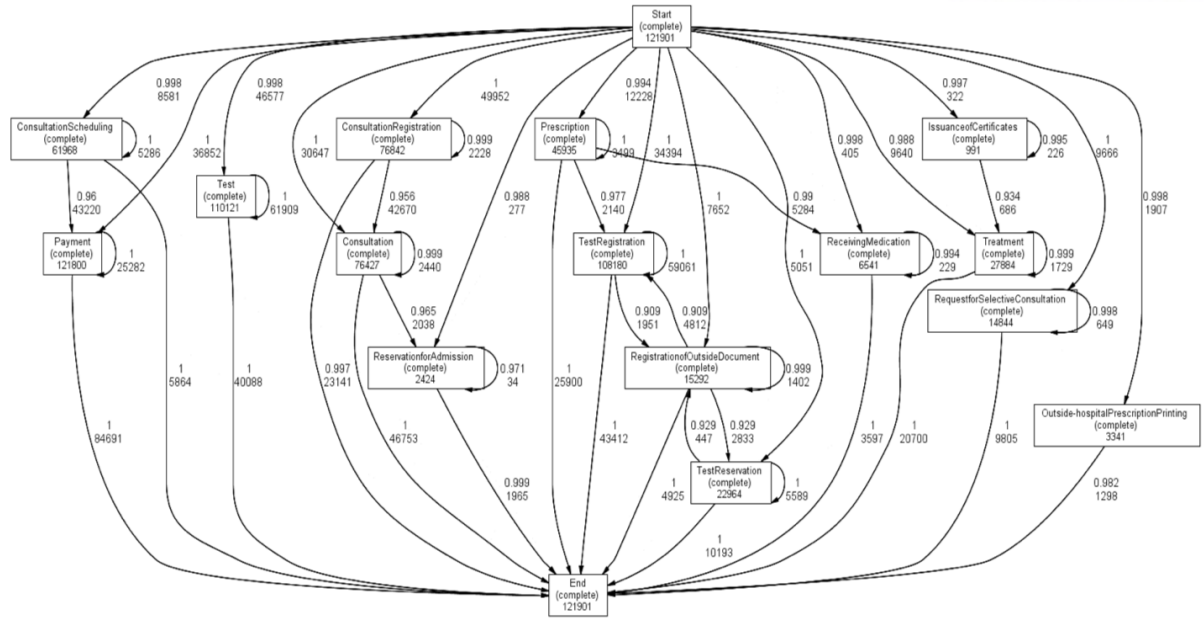
1) Discovering a process model and matching rate analysis

We first discovered a couple of process models from the log using several control-flow discovery techniques and compared them with the standard one. In this regard, we applied three different discovery algorithms including *heuristic mining*, *fuzzy mining*, and *frequency mining*. Figure 24 presents the discovered process models that indicate actual behaviors of outpatients in the hospital.

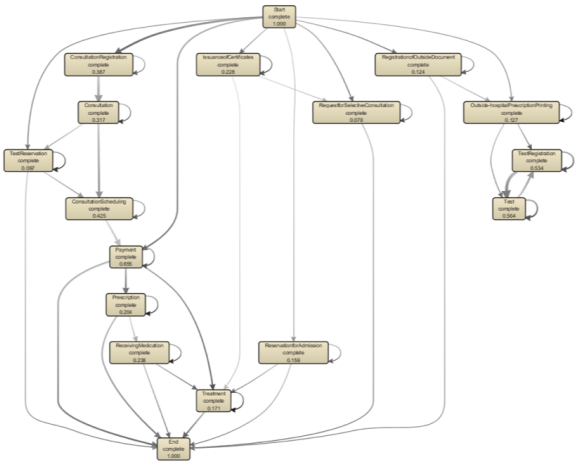
The heuristic mining result (Figure 24a) and fuzzy mining result (Figure 24b) showed major behaviors in the outpatient process, while the frequency mining result (Figure 24c) showed all possible paths among the activities including less-frequent flows. The most frequent flow is from *consultation registration* to *consultation*, which occurred about 64,000 times. The flows that happened more than 10,000 times are as follows.

- *Consultation registration* → *Consultation*
- *Test* → *Consultation registration*
- *Consultation* → *Payment*
- *Consultation scheduling* → *Payment*
- *Payment* → *Payment*
- *Test* → *Test registration*
- *Payment* → *Test registration*
- *Test registration* → *Test*
- *Payment* → *Outside-hospital prescription printing*
- *Payment* → *Treatment*

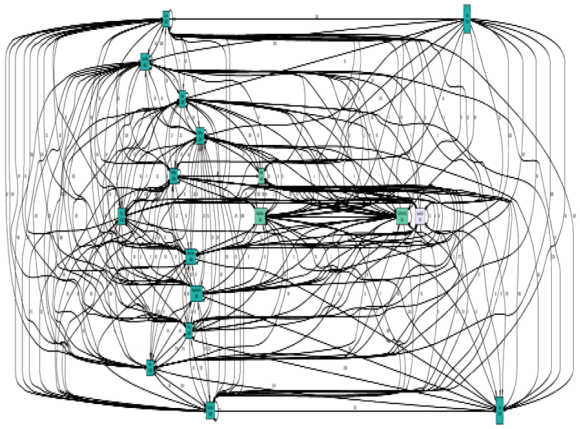
Figure 25 shows the difference between the standard model and frequency mining result. In the figure, the red and green lines represent non-matched and matched flows, respectively. Based on the comparison result, we calculated the matching rate between two models, which



(a) Heuristic mining approach



(b) Fuzzy mining approach



(c) Frequency mining approach

Figure 24. The discovered clinical process models using different discovery algorithms

was measured as 89.01%. As depicted in the figure, the discovered process model was significantly complicated, whereas the reference model was straightforward. Nonetheless, the calculated matching rate was considerably high (i.e., almost 90%). This is because most of the clinical process patterns followed the main flows presented in the standard model. Furthermore, through the discussion with domain experts (e.g., medical professionals), we were able to conclude that there was no any undesirable flows in the discovered model, i.e., the hospital has well managed the clinical process for outpatients.

2) Process pattern analysis

We first identified the most frequent patterns for the whole and each patient type (i.e., *new*

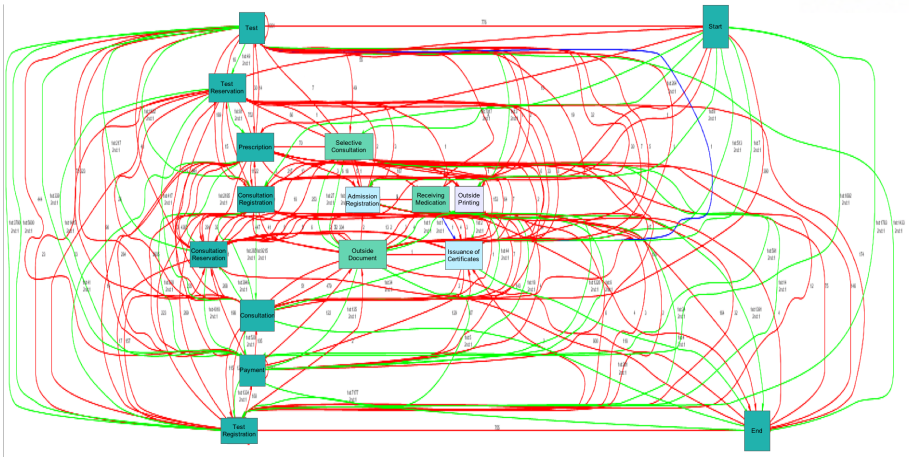


Figure 25. Comparison with the reference model and discovered model

and *returning*) from the whole event log. Figure 26 represents the most frequent patterns on the reference model. In the figure, the black-solid line refers to the most frequent pattern for all patients. In general, most patients firstly registered consulting room after visiting the hospital. Then, they received a consultation from a doctor and paid the money for services provided for them. Finally, patients left the hospital. In simple, the pattern was *Start* → *Consultation registration* → *Consultation* → *Payment* → *End*.

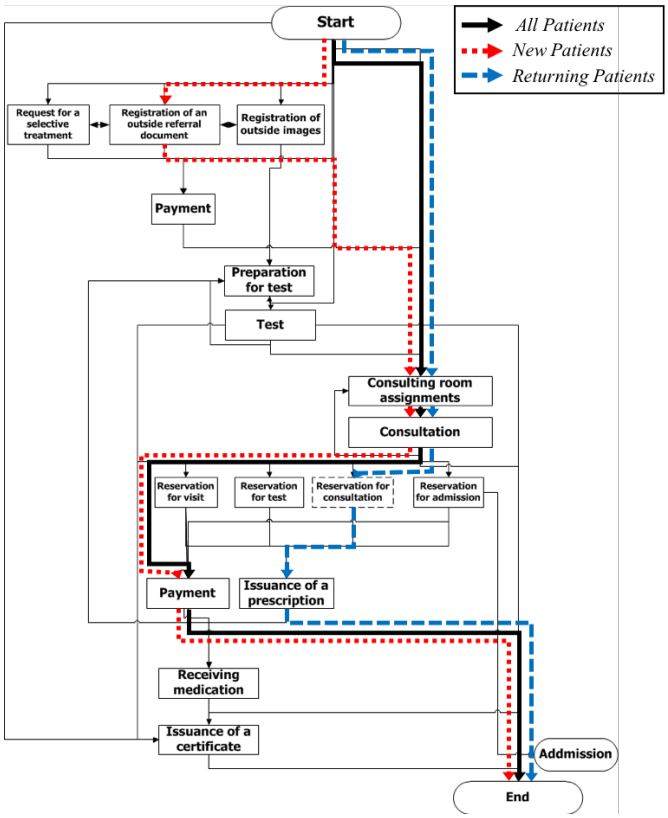


Figure 26. The most frequent patterns

With regard to the pattern, we conducted further analysis using a dotted chart to visualize the pattern. Figure 27 represents the dotted chart result, where red, green, and blue dots refer to *consultation registration*, *consultation*, and *payment*, respectively. In the figure, we identified that the distances between red and green dots are distant, while between green and blue dots are relatively close. That is, in most of the cases, it was turned out that consultation had a problem with a long duration. Furthermore, there were some exceptional cases (i.e., green dots are closer to red dots than blue ones) on the bottom of the chart. We discussed with domain experts, and it was turned out that some patients did not get a proper guideline about following activities, and they waited for a long time.

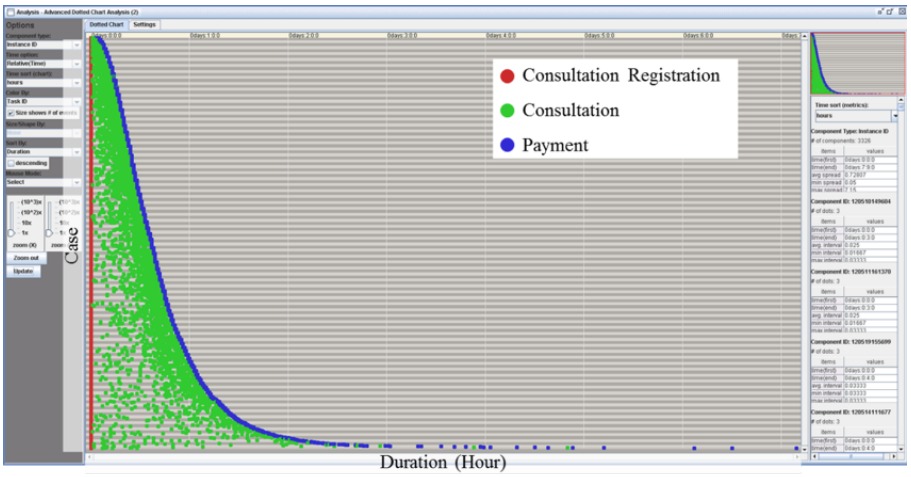


Figure 27. The dotted chart for the most frequent pattern

Also, we found out that the derived process patterns have a discrepancy according to the patient type as depicted in Figure 26. The '*returning patients*' enrolled a consultation room and got a consultation as soon as they visited the hospital, whereas the '*new patients*' were registered an outside referral document for the first task. The pattern analysis showed that '*new patients*' stayed longer than '*returning patients*' in the hospital. Such a data analysis result was used to build the smart healthcare system in the hospital. In other words, patients had the ability to find their own route with a smartphone application developed based on the pattern analysis results.

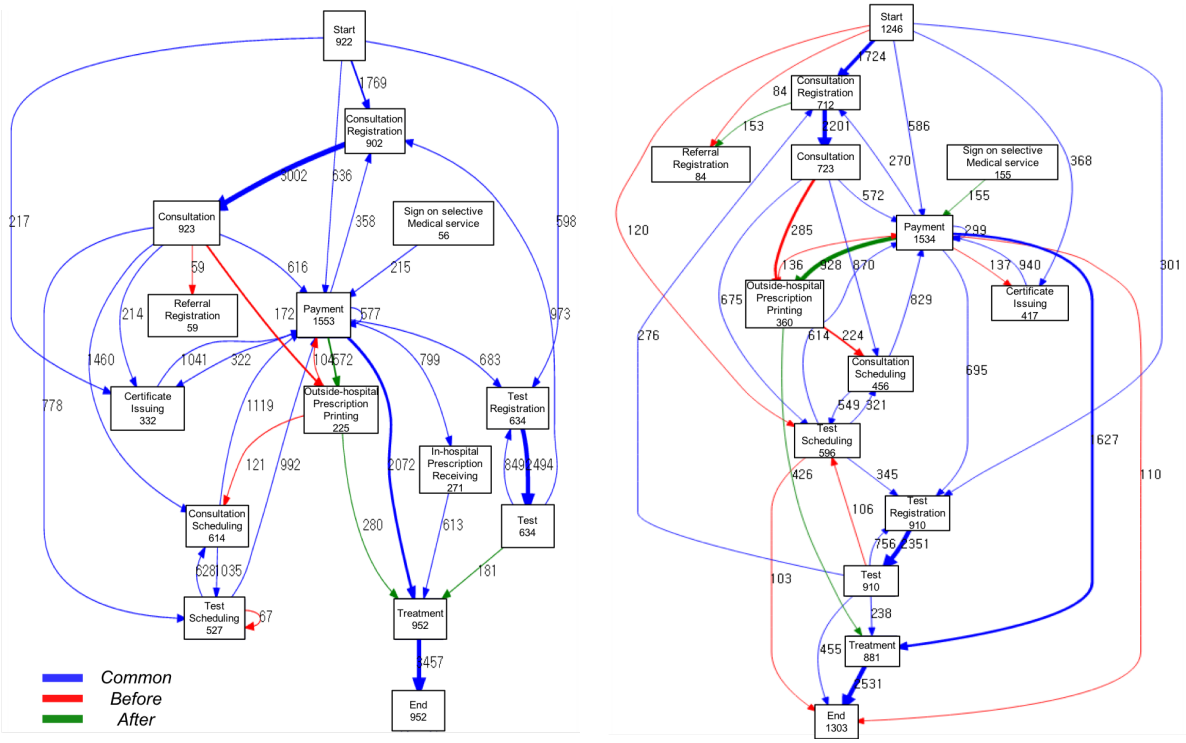
4.5.2 Evaluation for outpatients data analysis 2

This chapter also provides a validation result for outpatient's data analysis methodology with a real-life event log. Different from the first evaluation case, however, this case study focuses on evaluating the effectiveness of the changes in the hospital facility environment before and after the establishment of the new building in terms of the process. To this end, event logs are collected with one month of data prior to the establishment of the new building in July of 2012 and one month of data after the establishment of the new building in July of 2013. Similar to

the first case, the event log contained 15 different activities.

1) Discovering a process model and matching rate analysis

Using process mining technology, we analyzed the patterns of patients in the outpatient care clinics of cancer and clinical neuroscience centers. Figure 28 depicts the discovered process models for two clinical centers.



(a) Cancer center (b) Clinical neuroscience center

Figure 28. The discovered process models for the cancer center and clinical neuroscience center

As shown in Figure 28, we derived the most frequent outpatient care processes before and after the establishment of cancer and clinical neuroscience centers at the new building and confirmed the frequency of each process. We found that there were no remarkable changes in the most frequent outpatient care processes before and after the establishment of the new building. Based on a comparison of the processes used at cancer and clinical neuroscience centers, the cancer center appeared to have more patients with severe, rare, or incurable disease due to a higher number of in-hospital prescriptions issued and a higher rate of self-injection prescriptions. The clinical neuroscience center was hypothesized to have more patients en route from other hospitals due to the higher rate of consultation referral registrations and registrations for video resources of other hospitals compared to the cancer center.

The matching rate, which is the ratio of matches with the expert-driven model in the total flow frequency, increased from 87.0% before the establishment of the new building to 88.9% after the establishment of the new building at the cancer center. However, the matching rate decreased

from 86.8% to 85.2% after the establishment of the new building at the clinical neuroscience center.

2) Performance analysis

To evaluate the efficiency of operating the outpatient clinic after the establishment of the new building, we analyzed a couple of key performance indices, such as the total time of the outpatient care process, the consultation wait time, and the test wait time, while considering changes in the number of patients. The total time of the outpatient care process indicates the time from when the process was logged after a patient's visit to the hospital to the completion of the final process stage. The consultation wait time refers to the duration from the time after consultation is registered until the time when the consultation begins. The test wait time refers to the time after the test is registered until the time when the test begins. Table 12 presents the changes before and after the construction.

Table 12. Changes before and after the construction of the new building

	Cancer center				Clinical neuroscience center			
	Before	After	Growth(%)	p-value	Before	After	Growth(%)	p-value
Total number of outpatients	1000	2546	154.6	–	1337	2243	67.8	–
Total time for outpatient cares	116.97	127.33	8.9	0.023	88.23	90.73	2.8	0.520
Consultation waiting time	23.24	22.18	-4.6	0.271	27.08	23.72	-12.4	0.005
Test waiting time	11.94	19.43	62.7	0.001	7.17	6.61	-7.8	0.112
Number of tests per patient	0.68	0.77	4.4	0.027	0.75	.68	-9.3	0.177

We found that the number of outpatients increased by 154.6% (approximately 2.5 times) in the cancer center and 67.8% (approximately 1.7 times) in the clinical neuroscience center compared to the number of outpatients before the new building was established. However, the total time required for outpatient care increased by 8.9% (10.36 minutes) in the cancer center and by 2.8% (2.5 minutes) in the clinical neuroscience center. The total time required for outpatient care did not increase significantly considering the growth rate of the number of patients. Rather, the consultation wait time decreased by 4.6% (1.06 minutes) in the cancer center and by 12.4% (3.36 minutes) in the clinical neuroscience center compared to the consultation wait times before opening the new building. The test wait time increased by 62.7% (7.49 minutes) in the cancer

center but decreased by 7.8% (0.56 minutes) in the clinical neuroscience center. Moreover, the number of tests per patient increased by 4.4% in the cancer center and decreased by 9.3% in the clinical neuroscience center. For a detailed analysis of the wait time for each test, we categorized the tests by considering the characteristics of each test. The tests were divided into five groups, namely, a specimen test, medical imaging test, special test, departmental test, and miscellaneous tests. For the specimen tests, we included tests such as tests by laboratory departments, including the department of nuclear medicine and the department of pathology. For the medical imaging tests, we included tests from the departments of radiology and medical imaging. A special inspections category included tests conducted by the department of special inspections, such as endoscopy and spirometry tests. The departmental inspections included various specifics tests that were performed in each clinical department. For the miscellaneous section, we included test data logs remaining in groups other than the four aforementioned categories, such as the medical support services team, education training support team, nursing unit, and office of radiation safety management.

Table 13. Changes in the test waiting time and the number of tests

	Test waiting time growth(%)				Growth in the number of tests(%)	
	CC	p-value	CNSC	p-value	CC	CNSC
Specimen test	53.6	0.000	-23.1	0.024	172.9	53.4
Medical imaging test	43.8	0.206	25.9	.310	120.4	52.5
Special test	0.0	0.999	-58.9	0.267	146.4	8.4
Departmental test	-35.5	0.692	297.4	0.054	881.8	55.3
Misc.	-61.7	0.252	-60.7	0.028	87.9	40.7

Based on a comparison of the changes in the test wait time and the number of tests in the cancer center and the clinical neuroscience center by the types of tests, the number of specimen tests for the cancer center increased (Table 13). The number of specimen tests conducted by the laboratory department in the cancer center increased approximately 2.5-fold (380 to 960 tests). However, tests conducted by this department have limitations. For example, patients' test times are not absolute because these tests are not scheduled, which consequently affects the number of patients tested. In addition, tests conducted by the laboratory department did not accurately reflect the patients' wait time because these tests measured the time that elapsed from the printing of the specimen test label to the time when the specimen was recorded as being taken at the testing site. In contrast, for the scheduled tests, the time was recorded from when a patient was registered for a test to when the test was initiated.

4.5.3 Evaluation for inpatients data analysis

The data analysis methodology for inpatients was validated with the log data recorded between January and December 2013 were extracted from the EHR of a tertiary general hospital. For the full year of 2013, we have collected 53,965 subjects except for 745 and 1,029 subjects who were in the hospital at the first and last day of the year, respectively. Also, there was a lack of a discharge date for two subjects, and 8,295 patients were received the day surgery which does not have to be hospitalized. They were also removed from the set of target subjects being analyzed. In a nutshell, out of a total of 53,965 subjects, 8,419 subjects were excluded due to repeat admission for unexpected events (122), lack of a discharge date (2), and day surgery (8,295). Finally, data from 45,546 subjects were analyzed. For accurate data analysis, the following data were excluded: data presumed to have been wrongly entered, such as transfer note dates recorded before the admission date or after the discharge date; transfer completion dates recorded before the admission date; and procedures performed beyond the extracted date.

1) Performance analysis for LOS

Examining the data from 2013, the hospitalized patients were averagely discharged around 7 days, and the range of the length of hospital stay was quite extensive (i.e., interquartile range: 2.0–8.0). Also, as far as the distribution of the LOS was concerned, approximately 55% (25,228) of hospitalized patients were discharged within four days, and out of these patients, approximately 20% (8,969) were left the hospital on the second day of hospitalization.

Furthermore, a granular analysis on the length of stay was carried out by departments. Figure 29 presents the boxplot of the length of hospital stay for each department. Unexpected records, i.e., outliers, were removed in the graph for the meaningful comparative analysis.

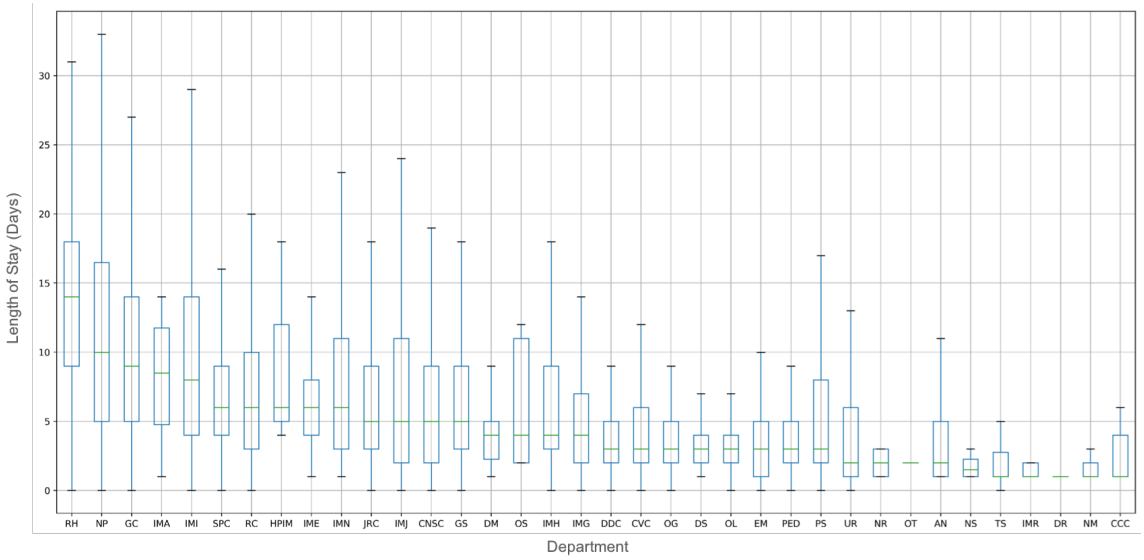


Figure 29. The length of stay by department

The median length of hospital stay was 14 days in rehabilitation medicine; 10 days for neuropsychiatry; 9 days for geriatric center admissions; and 8 days for internal medicine, infectious diseases. Also, the IQR of hospital stay was 11.50 (i.e., 5.0–16.50) days for neuropsychiatry; and 10 (i.e., 4.0–14.0) days for internal medicine, infectious diseases.

Based on the analysis of average and interquartile range (i.e., IQR) of the LOS within each department, patients were divided into three groups. Note that IQR signifies the statistical dispersion of the distribution. Figure 30 shows the results of the analysis according to the average and IQR of LOS per department.

Based on the analysis for the average and IQR of LOS, we identified that there was a positive correlation between two measures. As it were, the average of LOS was higher, and the overall disperse of LOS was higher as well. Considering this trend, we classified the departments into three groups as follows.

- *Group A* – Average & IQR of the LOS: Low
- *Group B* – Average & IQR of the LOS: High
- *Group C* – Average & IQR of the LOS: Considerably High

First, *Group A* was the group with the relatively lower IQR and average LOS than other departments. The group included those being treated under radiology (DR), ophthalmology (OT), or obstetrics and gynecology (OG) among others. These departments within the group A were seen to be doing well in keeping their patients with the short and low dispersed stay. Therefore, it was judged to be a group with a considerably low need for improvement. *Group B* included relatively higher IQR and the average of LOS than other departments. Clinical neuroscience center (CNSC), internal medicine nephrology (IMN), and internal medicine allergy (IMA) exhibited this trait and were noted as the departments to be improved their inpatient management. It was identified that these departments had the average LOS close to the average of the whole departments, i.e., 6.01 days. However, some patients had remarkably higher LOS than others; thus, it resulted in the slightly high IQR value. Therefore, we concluded that there is a need to systematically manage the medical care process of the specific patients. Finally, *Group C* was characterized by significantly higher average and IQR of the length of hospital stay. Rehabilitation medicine (RH), neuropsychiatry (NP), internal medicine infectious disease (IMI), Geriatric Center (GC) were included in Group C, and detailed analysis of patient characteristics was required to identify issues that may cause prolonged LOS.

2) LOS in accordance with diagnosis

Diagnosis was a significant factor correlating with the number of days of care [18]. LOS is determined by different variables and depends on specific diagnoses. The average LOS for each ICD-10 diagnosis issued by each department was converted to a Z-score, and then analyzed. An ICD-10 code consists of the first three characters for designating diagnosis category, the next three characters (characters three through six) for representing further details including

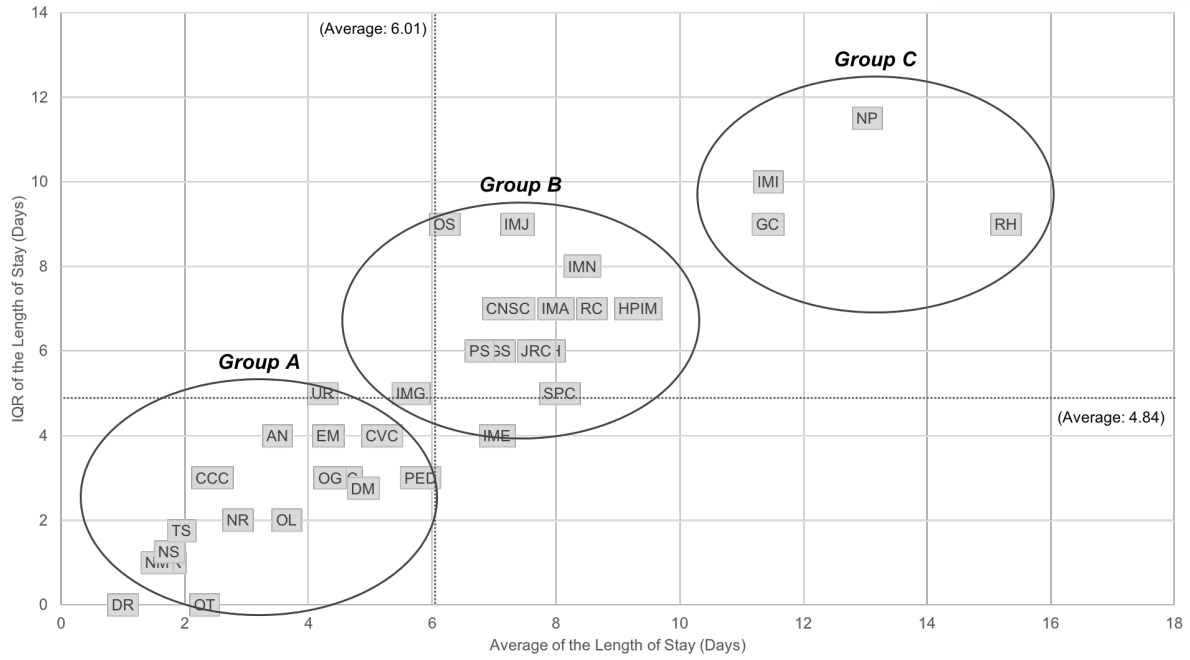


Figure 30. The results of the analysis according to the average and IQR of LOS per department

the related etiology, anatomic site, or severity, and the seventh character for expansion. Figure 31 shows the distribution of diagnostic standard scores according to each department. Diagnoses such as J44.9(chronic obstructive pulmonary disease), T82.7(vascular graft infection), T04.3(crushing injuries involving multiple regions of lower limb), M00.99(septic arthritis site unspecified), Z93.8(jejunostomy state) often yield greater standard scores compared to other diagnoses, even within the same area of medicine.

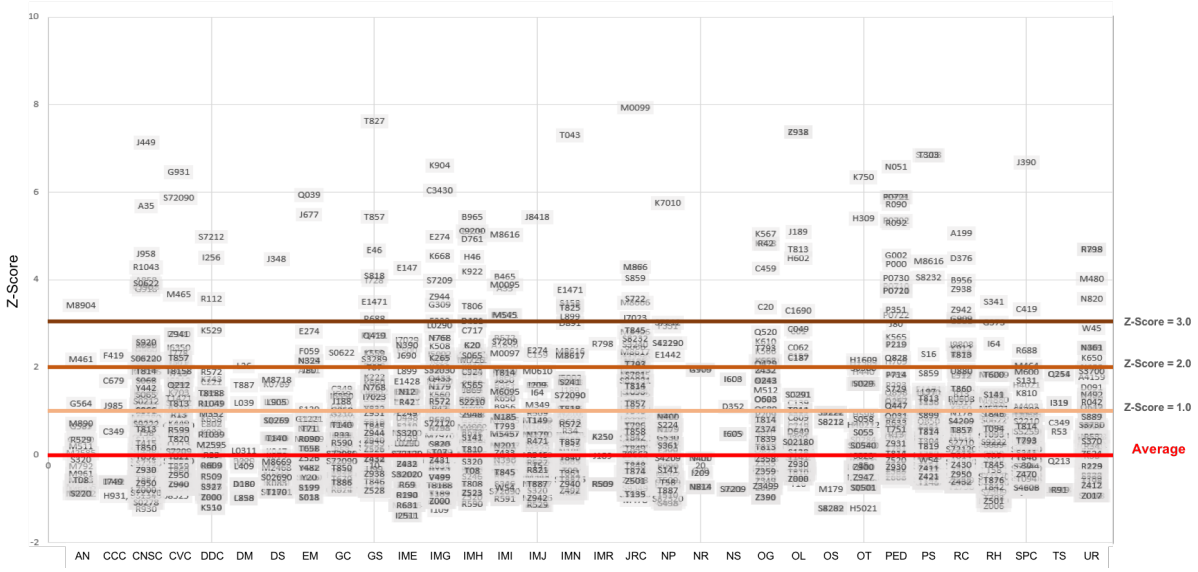


Figure 31. Diagnosis standard deviation distribution by department

3) Analysis for long-term hospitalization patients

Discharge of long-term inpatients is one of the main indicators actively managed by the hospital because shorter hospital stay is directly associated with an increase in hospital income, by increasing hospital turnover rate as well increasing the daily average cost of medical care. Usually, "*long-term inpatients*" are defined as patients who have been hospitalized for over 30 days.

Patients were divided into three groups, according to LOS: A (under 7 days), B (7 days or more and under 30 days), and C (30 days or more). Approximately 3% (1,327people) of all patients were long-term inpatients in Group C. Table 14 shows that, compared to patients with shorter LOS, long-term inpatients included a significantly higher rate of surgical patients (54.26%), transferred patients (41.97%), and patients on antibiotic treatment (92.31%) as well as a greater number of surgical interventions (2.01 cases), antibiotics (116.55 cases), and procedures (385.99 cases) per patient, and a greater number of treatments per person per day (7.96 cases).

Table 14. Comparison between differing length of hospital stay

	A	B	C
Items	(under 7 days)	(7 $\sim$ 30 days)	(30 days or more)
Number of Patients	31250	12969	1327
(Percentage, %)	(69.00)	(28.00)	(3.00)
Average LOS (days)	3.03	12.23	48.51
Surgical patients (%)	38.72	50.75	54.26
Surgery per patient (cases)	1.01	1.13	2.10
Transferred patients (%)	0.75	12.35	41.97
Patients on antibiotics treatment (%)	54.98	77.86	92.31
Antibiotics per patient (cases)	7.57	24.78	116.55
Antibiotics per patient in a day (cases)	2.50	2.03	2.40
Procedures per patient (cases)	22.87	84.54	385.99
Procedures per patient in a day (case)	7.55	6.91	7.96

With increased LOS, patients are exposed to a higher risk of infections and the use of broad-spectrum antibiotics increases accordingly. The use of broad-spectrum antibiotics may lead to the development of resistance to drugs and other serious side effects. For this reason, a number of antibiotics are managed as Restricted Antibiotics and subject to limited prescription. In this study, the ratio of restricted antibiotics administered to 1,000 randomly selected patients (12.79%) was higher than that in group A (0.7%) or group B (2.99%) (P-value < 0.001).

4) LOS analysis in terms of transfer patterns

Transfer is defined as a change of department required by the patient's condition and one of the

factors associating with the processing of hospitalized patients.

According to the analysis of hospital days based on transfer pattern, it was found that out of all patients, 5.25% (2,392) those who have been transferred on average spent 17 more days stay the hospital than those who were not transferred. There were highest number of incidents of patients being transferred to the departments of RH and IMH, and out of these, those being transferred from CNSC (Mean: 29.56 and IQR: 21.25–34.00) and SPC (Mean: 34.08 and IQR: 23.25–42.00) to RH had the highest interquartile range and also the average LOS. LOS by transfer pattern is shown in Table 15.

Table 15. Length of hospital stay by transfer pattern

		Length of stay (days)					
	Items	# of patients	Mean	Med	IQR	Min	Max
Transfer	Patients who were not transferred	43154	6.08	4.0	2-7	0	243
	Patients who were transferred	2392	23.12	17.0	10-29	1	213
Transfer	CNSC → RH	294	29.56	27	21.25-34	7	148
patterns	IMG → GS	251	16.73	12	8-20	2	110
	RC → IMH	169	11.33	9	7-13	3	67
	JRC → RH	71	25.08	22	18-28.5	4	87
	SPC → RH	62	34.08	28	23.25-42	11	88
	IMG → IMH	55	11.55	9	8-12	3	44
	GS → IMH	46	15.33	11	8-19.75	3	59
	CNSC → IMH	41	14.93	13	9-18	3	56
	GS → PS	37	14.16	10	8-12	4	61
	IMN → UR	22	13.59	10.5	7-20	6	31
	IMH → GS	22	18.27	16.5	10-24.75	5	43
	CVC → RC	21	19.48	17	10-24	7	63
	GS → IMG	20	17.35	12	9-21.25	5	46

5) Deriving correlated factors on LOS

The results of the analysis between LOS and different hospitalization variables are presented as follows: time required for transfer, discharge delay, surgery frequency, diagnosis frequency, severity, bed grade, and insurance type.

Patients requiring 2 or more days to transfer had a greater number of hospital days (Mean: 23.99 and IQR: 11.00–27.75) than patient with a LOS of under 2 days (Mean: 14.01 and IQR: 7.00–17.00). Patient discharge time showed that patients requiring 1-2 days (Mean: 7.12 and IQR: 3.00–8.00) had a higher LOS than patients requiring 1 day or less (Mean: 6.96 and IQR:

2.00–8.00) or 2 days or more (Mean: 5.54 and IQR: 3.00–5.00). Hospital stay based on incidence of surgery showed that patients undergoing 3 times or more surgical interventions had the longest LOS (Mean: 50.30 and IQR: 23.00–64.00) compared to patients undergoing no surgery (Mean: 6.17 and IQR: 2.00–7.00), 1 intervention (Mean: 6.97 and IQR: 3.00–8.00) or 2 interventions (Mean: 21.25 and IQR: 10.00–24.00). In terms of diagnosis, patients with 3 or more diagnoses had the longest hospital stay (Mean: 38.24 and IQR: 16.00–50.00) compared to patients with no diagnosis (Mean: 3.33 and IQR: 2.00–4.00), 1 diagnosis (Mean: 6.07 and IQR: 2.00–7.00), 2 diagnoses (Mean: 14.53 and IQR: 4.00–20.00).

Patients receiving critical care (Mean: 7.94 and IQR: 3.00–9.00) were more likely to have longer LOS than those who were not (Mean: 6.56 and IQR: 2.00–7.00), and patients on general wards (Mean: 5.84 and IQR: 2.00–7.00) were more likely to remain in hospital longer than patients on upper grade wards (Mean: 2.90 and IQR: 1.00–3.00). Analysis of hospital stay according to insurance type indicated that admissions involving industrial accidents, medical assistance, medical research, and automobiles occurred less frequently than admissions on health insurance, although the LOS was relatively higher. All variables were statistically significant. ($P<0.05$)

6) Building a predictive model of patient's LOS

This chapter presents a model for the prediction of the number of days in hospital based on the significant variables analyzed above. Multiple regression analysis was performed to develop the model. The following five variables were used as independent variables: frequency of surgery, frequency of diagnosis, frequency of patient transfer, severity, and insurance type. LOS was used as a dependent variable. Also, we partitioned data into the training and test dataset to measure the accuracy of the model; 80% and 20% of data became the training and test data, respectively. As a result, all five variables were statistically significant and, therefore, correlated with the prediction of the length of hospital stay. In the regression model from the training dataset, R^2 was 0.267, and duration of hospitalization was calculated as followed: LOS (days) = 2.10 + 2.62 × (frequency of surgery) + 3.04 × (frequency of diagnosis) + 11.13 × (number of transfer) + 1.76 × (severity) - 1.03 × (insurance type). As a result of measuring the accuracy with the test dataset, we identified that the mean absolute error of the model is 4.68.

4.5.4 Evaluation for clinical pathways analysis

In this study, we primarily analyzed the appendectomy CP, which has been in use since 2009 in a tertiary general hospital. Because the appendectomy CP has been continuously improved since its development, this study focused on analyzing the most updated version of the CP. Thus, we analyzed the appendectomy CP based on patients who were enrolled in the appendectomy CP (out of all hospitalized patients) between August in 2013 and June in 2014. The CP had eight stages including pre-operation, pre-operation (preliminary), intra-operation, post-operation, post-operation (preliminary), normal order, normal order (preliminary), and dis-

charge. Also, there were three types of orders: inspection, treatment, and medication. As far as the data was concerned, it was extracted from 164 hospitalized patients (9,296 events) in the appendectomy CP, which was applied for a total of three consecutive days. To ensure the accuracy of the analysis, we used data that had a structured value in the pre-processing of the extracted data and excluded order data that were categorized as '*diet*'. In addition, we used only working orders and excluded those medication orders that were entered by anesthesiology because anesthesiology orders are not targeted toward making improvements in the CP.

1) Comparing CP orders and logs

Using the CP and the collected event log, we conducted the matching analysis using the matching algorithm. Figure 32 shows the visual results of the matching analyses. The rows indicate each patient, and columns show all events in the event log. In the figure, dark gray boxes signify matching events since they were conducted for each patient and included in the CP. Light gray boxes signify non-matching events since they were conducted for each patient, but not included in the CP. Thus, the collection of dark gray boxes for each row represent the matching set for each patient; the collection of light gray boxes for each row represent the non-matching set for each patient. Lastly, white boxes signify events, which were not performed for each patient. When looking at the matching results of the events included in the CP, some events were not conducted in all or almost all patients. These orders became subjects of elimination while editing the CP. Conversely, when looking at the events not included in the CP, some events were conducted by almost all patients. These events became subjects of additional inclusion.

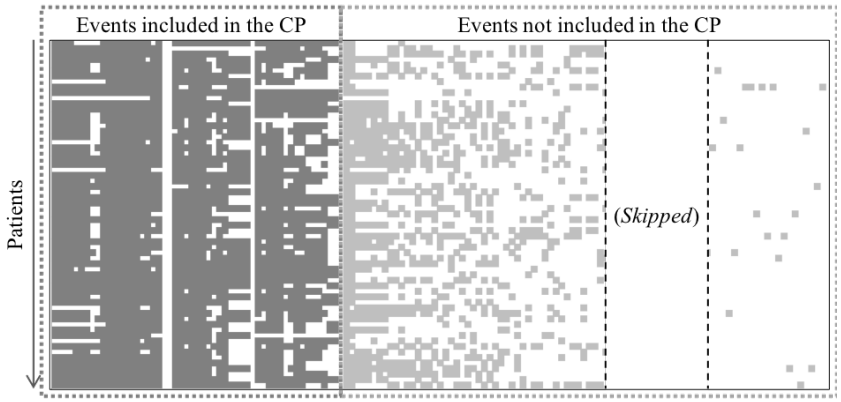


Figure 32. The result of the matching process

2) CP matching rate analysis

Based on the analyzed matching results, we calculated each patient's N_{cp}, M_{cp}, N_{log}, R_{log}, AR_{cp}, MR_{log}, and the standard matching rate ($SCMR$). Table 16 shows the statistical result of them. N_{cp}, which signifies the number of orders in CP, was 53 for all, because the same CP was applied to all patients in the study. Out of these 53 activities, on average of 14 turned out that were not performed on patients. In other words, 14 orders were missing (M_{cp}). Also, on average, about

67 activities were performed to patients (N_{log}), and among them, 28 events were analyzed as additional activities that were not included in the CP. In other words, 28 events were remaining (R_{log}). Patients' average value of the application rate of orders in the CP (AR_{cp}) was 0.74, and that of the matched ratio of events out of the event log (MR_{log}) was 0.61. Finally, the standard matching rate was about 0.68 averagely. For AR_{cp}, it was applied to the maximum of 92% and, for MR_{log}, 1.00 was recorded. In addition, the minimum value of matching rate was 0.44, and the maximum value was 0.91.

Table 16. The statistical result for measuring matching rate

	N_{cp}	M_{cp}	N_{log}	R_{log}	AR_{cp}	MR_{log}	$SCMR$
Average	53.00	14.02	66.56	27.57	0.74	0.61	0.68
Median	53.00	13.00	65.00	24.00	0.75	0.62	0.68
Minimum	53.00	4.00	28.00	0.00	0.47	0.22	0.44
Maximum	53.00	28.00	190.00	148.00	0.92	1.00	0.91

3) Building an improved CP

In the case study, based on discussion with domain experts, we deleted orders where the application ratio was lower than 0.5, and included some orders that had the application ratio of 0.75 or above, and created a new CP. As a result, a total of eight orders were deleted out of existing orders in the CP, and two events were included. Table 17 shows the results of the matching rate calculation. After the improvement, AR_{cp} and MR_{log} increased 0.1 and 0.02 respectively, which resulted in a 0.05 increase in the standard matching rate.

Table 17. The matching rate analysis result according to the CP change

	N_{cp}	M_{cp}	N_{log}	R_{log}	AR_{cp}	MR_{log}	$SCMR$
Existing	53.00	14.02	66.56	27.57	0.74	0.61	0.68
Revised	47.00	7.75	64.89	25.63	0.84	0.63	0.73

4.6 Summary and Discussion

This chapter presented data analysis frameworks for three clinical process types: outpatient, inpatients, and clinical pathways. These frameworks focused on diagnosing and understand each clinical process on the basis of the particular goals. More in detail, we provided specific analysis methods for three different types: outpatients, inpatients, and clinical pathways. In such a step, several concepts (e.g., frequency mining, matching rate, process performance indicators, length of

hospital stays performance analysis, and CP matching rate) were defined with a formal approach. Also, the in-depth evaluation results with four real-life logs (i.e., two for outpatients, one for inpatients, and one for clinical pathways) demonstrated the usefulness of our approach.

The frameworks have a distinctive contribution that narrows the gap between process-oriented research and the practical field. In other words, our frameworks answer the call for research in process mining for improving usability and understandability of process mining techniques and results for non-experts (see challenges 10 and 11 in the *process mining manifesto* [151]). Also, this research initiates with the common data model that has a role of shared resources for research communities and healthcare organizations. Therefore, it will act a standardized guideline for analyzing a clinical process for everyone. Furthermore, organizations with well-established healthcare information systems often do not make sufficient use of data. These frameworks will be a signal enabling such discarded data to be used for improving the clinical process. Finally, these frameworks were developed by collaborating with experts of process mining, health informatics, and clinical domain experts; thus, it has a strong confidence compared to other existing frameworks.

Despite these contributions, the frameworks has also several limitations. First, the proposed approaches may be a versatile solution, but they are not perfect. In other words, the objectives of a particular analysis may be modified or become new, and additional analysis items for them may be added accordingly. Therefore, it should be continuously developed and extended on the basis of domain experts of professionals, developments of state-of-the-art research methods, and experiences with numerous case studies. Furthermore, this research does not contribute to tool support for implementing the framework. Although several skillful process mining open-source tools including ProM Framework [152] are already implemented, it is highly supported by the development of software that can cover the entire process of this framework.

V Redesigning Clinical Processes with the Simulation-based Approach

In this chapter, we propose a novel approach to conduct a clinical process redesign with discrete event simulation based on process mining. The main benefits of this approach are that it is automatically implemented without any qualitative methods and highly precise thanks to a data-driven approach. This chapter is organized as follows: Chapter 5.1 presents the background of this research including the illustration of the problem. Chapter 5.2 provides the comprehensive explanations of the proposed methodology, and Chapter 5.3 gives its application in a case study to demonstrate the usefulness. Finally, Chapter 5.4 concludes this chapter by provisioning the summary and discussion.

5.1 Background

A typical approach for clinical process redesigns is *Discrete Event Simulation (DES)* [44, 153–157]. In the healthcare field, multiple studies have been using DES where clinical activities are considered as the crucial events in clinical processes. However, it requires generally much time and effort to build an accurate simulation model. This is because the status quo is that simulation models are created by manually recorded data, which may be inaccurate, or interviews, which are time-consuming. To overcome these limitations, Rozinat et al. [44] proposed to combine simulation with process mining as to extract process-related knowledgeable information from so-called event logs [3, 37, 44, 49, 150]. Process mining uses such automatically recorded logs to automatically derive the specific operations in a particular context, which is one of the leading components of a simulation model. The authors explained how to make a Colored Petri Net (CPN) model using four kinds of analyses [44].

Unfortunately, it still has a couple of challenges to straightforwardly apply this method for clinical processes.

1) The collected data from Electronic Health Record (EHR) system is not sufficient to derive accurate simulation parameters. Three main elements for building a healthcare simulation model are a process of medical activities, service times, and arrival rates. Out of them, it is hard to find out actual values of the service times and arrival rates from EHR data due to the following reasons:

- Service times: EHR systems, which typically, only record completion time of clinical activities.

- Arrival rates: Patients visit a hospital with a scheduled appointment; thus, the reservation system needs to be considered.

2) Even if the derived simulation model fully reflects the reality, there is no systematic approach to deriving effective improvements, i.e., experimental scenarios. The next step of building a simulation model is to identify all the possible alternatives and determine the best option for

the optimal decision making with simulation analysis. However, the existing methods presented in this regard are all heuristic-oriented and unstructured approaches. Therefore, a bridge that connects the simulation model analysis and useful scenarios is still missing.

To overcome these challenges, this paper proposes a novel decision support framework for simulation-based redesign analysis. To this end, a data-driven simulation model is constructed using process mining analysis, which includes process discovery, patient arrival rate analysis, and service time analysis. Also, improvement alternatives are investigated from data analysis for making experimental scenarios. Then, the validated optimal redesign methods are determined from the simulation analysis.

To specify the research subject of this chapter, we focus on the long waiting time considered as a key challenge in the outpatient clinical process [158, 159]. It is recognized as the critical problem since the longer patients have to wait before their consultation can take place, the less satisfied they are, which may lead to decreasing profits [160]. A confounding factor is that there are significant differences with respect to quality delivery and efficiency among clinicians. In order to handle this problem, it seems worthwhile to consider how the personal appointment schedules of clinicians can be optimized as to improve the overall efficiency of patient management.

5.2 A Discrete Event Simulation Approach based on Process Mining

5.2.1 Overview

The proposed decision support framework for the optimized medical scheduling is composed of four phases: *data preparation & preprocessing, data analysis, post-hoc analysis (simulation modeling)*, and *further analysis (experiments)*. Figure 33 represents the overview of the proposed framework. First, an appropriate format of data, i.e., an event log, is collected from EHR system's log data of the hospital and preprocessed for effective data analysis in the data preparation phase. After that, three kinds of process mining analysis are performed to derive the inputs for creating a simulation model: process discovery, arrival rate analysis, and service time analysis. Based on these results, a simulation model is constructed, and then the model is evaluated to validate whether the model reflects the behaviors observed from the data (i.e., As-Is analysis). Here, a couple of Key Performance Indicators (KPIs) are employed. Lastly, in the experiments & decision support phase, improvement alternatives are investigated by performing further data analysis. To this end, best practices for business process redesign [17] are utilized as candidates for process improvement. Then, several scenario-based simulation analyses are performed to identify the optimal redesign method (i.e., To-Be analysis).

5.2.2 Data Preparation & Preprocessing

As stated before, we employ process mining approaches for deriving simulation parameters. That is, we need to utilize event logs, which represent the behaviors recorded by an information system

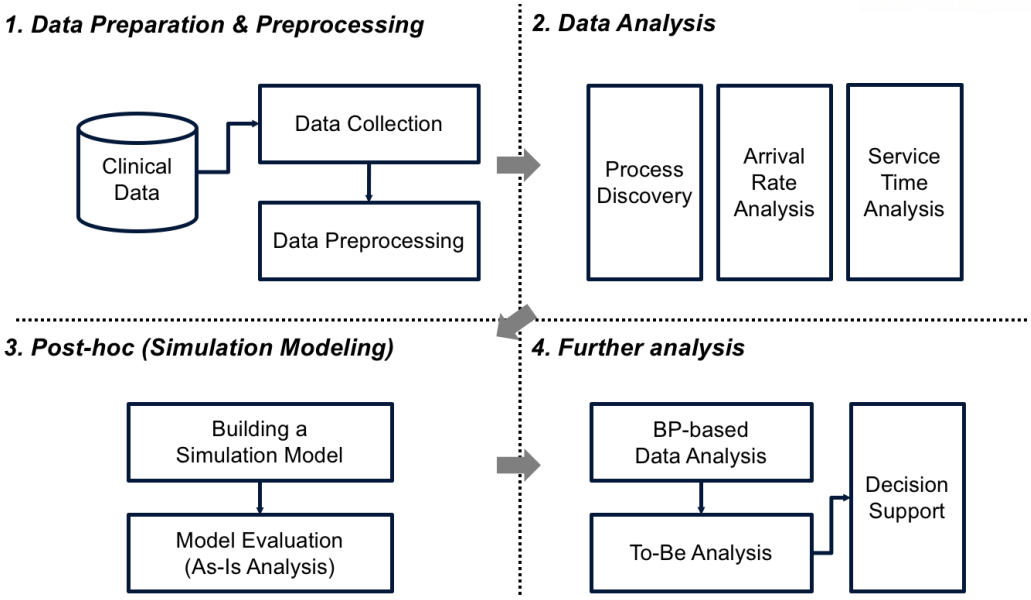


Figure 33. The proposed decision support framework for medical scheduling

and are used in process mining approaches. To this end, we employ Definition 1 as introduced before.

After collecting clinical event logs, the data preprocessing step is conducted to improve the accuracy and effectiveness of the data analysis. It includes removing noisy data, identifying outliers, and handling incomplete or error data.

5.2.3 Data Analysis

1) Process discovery

Process discovery aims at extracting process models from event logs [3, 40–45]. Through seminal research, many kinds of discovery algorithms have been developed, such as alpha-mining [3], heuristic mining [45], genetic mining [41], fuzzy mining [42] and inductive mining [43]. In this research, we apply frequency mining introduced in Definition 2.

2) Arrival rate analysis

As we stated earlier, patients visit a hospital by a specifically scheduled appointment. To arrive at an accurate simulation model, it is essential to build it based on the characteristics of such schedules, including information on slot capacity and intervals between slots. Here, we propose a method to analyze a realistic arrival rate by applying two sorts of information related to the reservation system.

(1) The number of appointments for each reservation slot

(2) The patients' visiting time (the actual) compared to the reservation time (the planned)

The pseudo-code in Algorithm 2 explains the proposed approach in detail.

By computing how many patients visited the hospital in each slot, we can derive the visiting

Algorithm 2 *DerivingArrivalRate*(L, TS)

Input Event Log L;
Time slots for reservation TS (a time slot $ts_k \in TS$)

Output The number of appointments for each time slot N;
A collection of patients' visiting time compared to the reservation time D

$N \leftarrow$ an initialized array with size $|TS|$
$D \leftarrow \{\}$
for all traces σ_i in the log L **do**
  $visitT_{\sigma_i} \leftarrow 0$
  for all events e_j in the trace σ_i **do**
    for all time slots $ts_k \in TS$ **do**
      if $\pi_{rtime}(e_j)$ is involved in ts_k **then**
        $N[k] \leftarrow N[k] + 1$
        break
      end if
    end for
    if $visitT_{\sigma_i} = 0$ **or** $visitT_{\sigma_i} < \pi_{ctime}(e_j)$ **then**
      $visitT_{\sigma_i} \leftarrow \pi_{ctime}(e_j)$
    end if
  end for
  $D \leftarrow D \cup visitT_{\sigma_i}$
end for
return N, D

distribution of patients. After that, we figure out the actual arriving time by applying the second type of information.

3) Service time analysis

Working time is one of the indispensable components in making a simulation model. However, it is hard to get the accurate working time because most EHR systems in hospitals record only the completed time for each activity [87]. In other words, the duration of activities cannot be divided into waiting and working time because of the absence of the start time of the consultation. For this reason, the status quo is that service time is derived from manual checking [161]. To avoid this laborious and imprecise step, we suggest a new method to estimate the working time for consultation from event logs. The pseudo-code in Algorithm 3 explains the proposed approach in detail.

This method has a realistic assumption that patients visit the consultation room where a doctor works in the consecutive order. In other words, the doctor sees patients one at a time,

Algorithm 3 *DerivingServiceAndWaitingTime*(L, r_i)

Input Event Log L;
A resource r_i

Output A collection of service time for a resource (r_i) S_i;
A collection of waiting time for a resource (r_i) W_i

Note that $\pi^{CR}_{ctime}(e_i)$ is the completion time for consultation registration, and $\pi^{C}_{ctime}(e_i)$ is the completion time for consultation.

1: $S_i \leftarrow \{\}$
2: $W_i \leftarrow \{\}$
3: **for all** events e_j in the log L **do**
4: **for all** resources $r_i \in R$ **do**
5: **sort by** $\pi^{C}_{ctime}(e_j)$
6: $E_i \leftarrow \{\}$
7: **if** $\pi^{C}_{res}(e_j) = r_i$ **do**
8: $E_i \leftarrow E_i \cup e_j$
9: **end if**
10: **for all** events $e_k \in E_i$ **do**
11: **if** e_{k-1} does not exist (i.e., k=1) **or** $\pi^{C}_{ctime}(e_{k-1}) < \pi^{CR}_{ctime}(e_k)$ **then**
12: $S_i \leftarrow S_i \cup \{\pi^{C}_{ctime}(e_k) - \pi^{CR}_{ctime}(e_k)\}$
13: **else**
14: $S_i \leftarrow S_i \cup \{\pi^{C}_{ctime}(e_k) - \pi^{C}_{ctime}(e_{k-1})\}$
15: $W_i \leftarrow W_i \cup \{\pi^{C}_{ctime}(e_{k-1}) - \pi^{CR}_{ctime}(e_k)\}$
16: **end if**
17: **end for**
18: **end for**
19: **end for**
20: **return** S_i, W_i

one after another. To explain the principle clearly, we provide a graphical example in Figure 34. In this figure, for each patient the end times for the *consultation registration* and *consultation* are shown. Note that all records are sorted by the end time for consultation. The method we propose to measure the consultation *service time* can be divided into two ways. First, we can distinguish those patients who either get a consultation as the first patient in a time slot (e.g., *P1*) or those patients whose end time for consultation registration is later than the previous patient's end time for the actual consultation (e.g., *P4*). Neither of these types patients has to wait at all since no people are waiting. For these patients, the actual service time of their

consultation equals the difference between the registered end times for consultation registration and consultation. By contrast, for the rest of them (e.g., *P2*, *P3*, and *P5*), the end time for the consultation registration of each patient is earlier than the end time for consultation of the previous patient. In such case, it is likely that at least one previous patient is waiting at the registration desk or consulting the doctor. Therefore, such patients have to wait before their consultation. The service time for such patients then equals the difference between their own consultation end time and the consultation end time of the previous patient. In absolute terms, a slight error may be introduced in that the preparation time for consultation might be included in the extracted service time. However, we believe that the advantages of such an automated approach outweigh those of manually measuring service and waiting times.

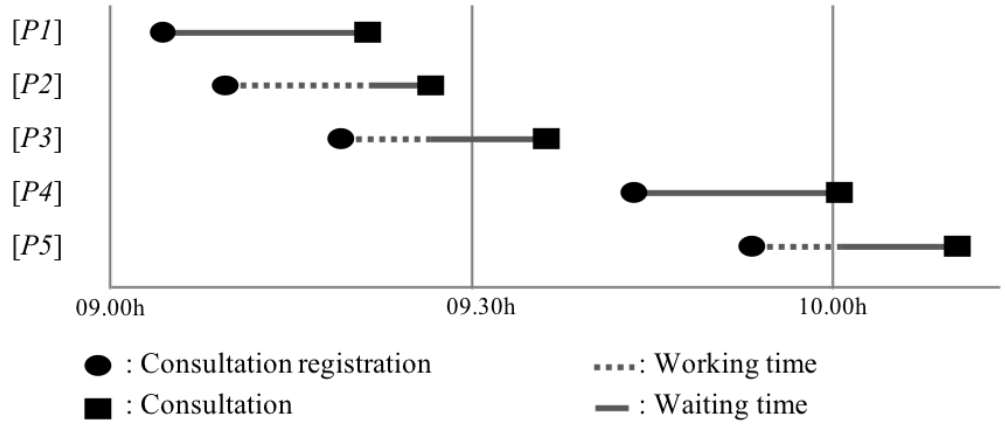


Figure 34. Measuring working time for consultation

5.2.4 Post-hoc Analysis (Simulation Modeling)

Based on the results of the three process mining analyses, we can now easily derive a simulation model. This is because we can get every input for the model from the process mining results—the extracted model, the arrival rate, and the service time. The model is then used to carry out a so-called as-is analysis, which allows for a comparison of the results generated by the simulation model with the observed data as present in the actual logs (i.e., records). To evaluate the model thoroughly, we define a couple of KPIs. In this study, we set three measures based on an in-depth discussion with domain experts in the hospital: the waiting time for consultation ($wt(L)$), the controllable waiting time for consultation ($cwt(L)$), and the end time of a clinic session for a specific doctor ($et(L)$). These are given in Definition 12. Note that events are sorted by the execution time for consultation, and KPIs are calculated for each doctor. Also, $\pi^{CR}_{ctime}(e_i)$ and $\pi^{C}_{ctime}(e_i)$ is the completion time for consultation registration and consultation, respectively. Moreover, $\pi^{C}_{rtime}(e_i)$ is the reservation time for consultation, and MAX is a function to find out the maximum value.

Definition 12 (Key performance indicators for assessing the constructed model) *Let*

$wt(L)$, $cwt(L)$, and $et(L)$ be the waiting time for consultation, the controllable waiting time for consultation, and the end time of a clinic session for a specific doctor, respectively.

- $wt(L) = \sum_{0 \le e < |c|} \sum_{0 \le i < |e|} \begin{cases} \pi^{C}_{ctime}(e_{i-1}) - \pi^{CR}_{ctime}(e_i) & \text{if } \pi^{C}_{ctime}(e_{i-1}) > \pi^{CR}_{ctime}(e_i) \\ 0 & \text{otherwise} \end{cases}$
- $cwt(L) = \sum_{0 \le e < |c|} \sum_{0 \le i < |e|} \begin{cases} \pi^{C}_{ctime}(e_{i-1}) - \pi^{C}_{ctime}(e_i) & \text{if } \pi^{C}_{ctime}(e_{i-1}) > \pi^{CR}_{ctime}(e_i) \\ 0 & \text{otherwise} \end{cases}$
- $et(L) = \sum_{0 \le e < |c|} \sum_{0 \le i < |e|} \{MAX(\pi^{C}_{ctime}(e_i))\}$

Among them, the second one signifies the difference between the start time of consultation and the reserved time, which excludes the waiting time which is incurred due to patients who registers ahead of the reserved time. Also, the third one represents the timestamp when consultation of the last patient in a session is completed. Based on these KPIs, we evaluate whether the simulation model accurately reflects the real situation or not. This may provide the required confidence to use the simulation model for alternative scenarios. Besides, we employ evaluation measurements, e.g., Mean Absolute Percentage Error (MAPE) [150] for comparing KPIs quantitatively, as defined in Definition 7.

5.2.5 Further Analysis (Experiments)

As mentioned earlier, prior to the scenario-based experimental simulation analysis, it is necessary to conduct preliminary data analysis for identifying a useful scenario which is applicable to improvements of the relevant process. This is because process improvement methods are quite diverse and have a broad range. In this chapter, 29 heuristic best practices by Reijers and Mansar [17] are employed as an available set of process improvement alternatives. It covers practical redesign methods such as activity elimination, case types, and case assignment. Based on these best practices, we suggest a series of steps to derive evidence-based simulation scenarios. First, we obtain the applicable best practices with the following conditions.

(1) Whether or not a best practice satisfies the goal of the simulation analysis

(2) Whether or not a best practice is already applied to the process

(3) Whether or not a best practice is more needed opinions from domain experts than data analysis

(4) Whether or not information related to a best practice is stored in the log

After that, we define indicators to identify the availability of each best practice determined through data analysis. The existing work [20] has developed the relevant indicators for each best practice, and they are applied immediately or slightly modified. For example, it is required to measure the number of occurred events within a small unit of time (e.g., 1 minute) for an activity to identify the availability of the order-based work best practice, i.e., removing batch-processing and periodic activities in a process. Then, as the measured value exceeds the predefined threshold, the relevant best practice is considered as one of the experimental simulation

scenarios.

The discovered experimental scenarios are tested based on the data-driven simulation model. To this end, we employ the waiting time for consultation and controllable waiting time among the KPIs already presented. Then, the extent of improvements is evaluated for all the experiments. Finally, the optimal scenario-based medical scheduling for a clinician is derived.

5.3 Evaluation

5.3.1 Context

We used the real-world data from EHR system at a fully-digitalized tertiary general university hospital in Korea. Using the clinical event logs, we tried to cooperate with many medical staffs for the effective clinician-specific care scheduling. To this end, we conducted classifying and grouping clinicians by the number of patients and the waiting time for consultation. Among the clusters, we set the target as a group which includes doctors who have a large number of patients and the long waiting time. In the target group, even though we were able to choose several clinicians, but the characteristics of patients for each clinician was too different to make a single simulation model. As a result, we selected only one doctor at a department of the hospital, and the data was a collection of patients who got a consultation from the doctor in May of 2012. The event log contained 15 tasks: consultation registration, consultation, consultation scheduling, test registration, test, test scheduling, payment, sign on selective medical service, referral registration, outside image registration, admission scheduling, outside-hospital prescription printing, in-hospital prescription receiving, treatment, and certificate issuing. Also, it included several attributes such as completion time, resources, departments, patient types, and reservation time. After extracting the data, we cleaned it using several preprocessing steps applying existing methods [141]. In summary, the preprocessed logs had about 8,000 events which were performed for about 1,300 patients. To conduct process mining analyses, we applied *ProDiscovery* [162] which were developed by our research group. Furthermore, we used *Automod* [163] to create a simulation model and received the further simulation analysis results using *Autostat* [163].

5.3.2 Process mining analysis results

We performed the three process mining analyses as described—process discovery, arrival rate analysis, service time analysis. First, we derived a process model using the frequency mining; Figure 35 represents the whole outpatient process from the event log. The derived model was very complicated and resembled a *'spaghetti process'*. That is to say, we were able to discover the flows of all outpatients in the hospital using the frequency mining. However, the simulation analysis of this case study aimed to decrease the waiting time for consultation of individual patients. In other words, we had to focus on the major flow which relates to consultation. As a consequence, we tried to find out the major flow by controlling the threshold value to make

a simulation model for personal clinician scheduling. Figure 36 describes the main flow of the discovered outpatient process with a pre-established threshold.

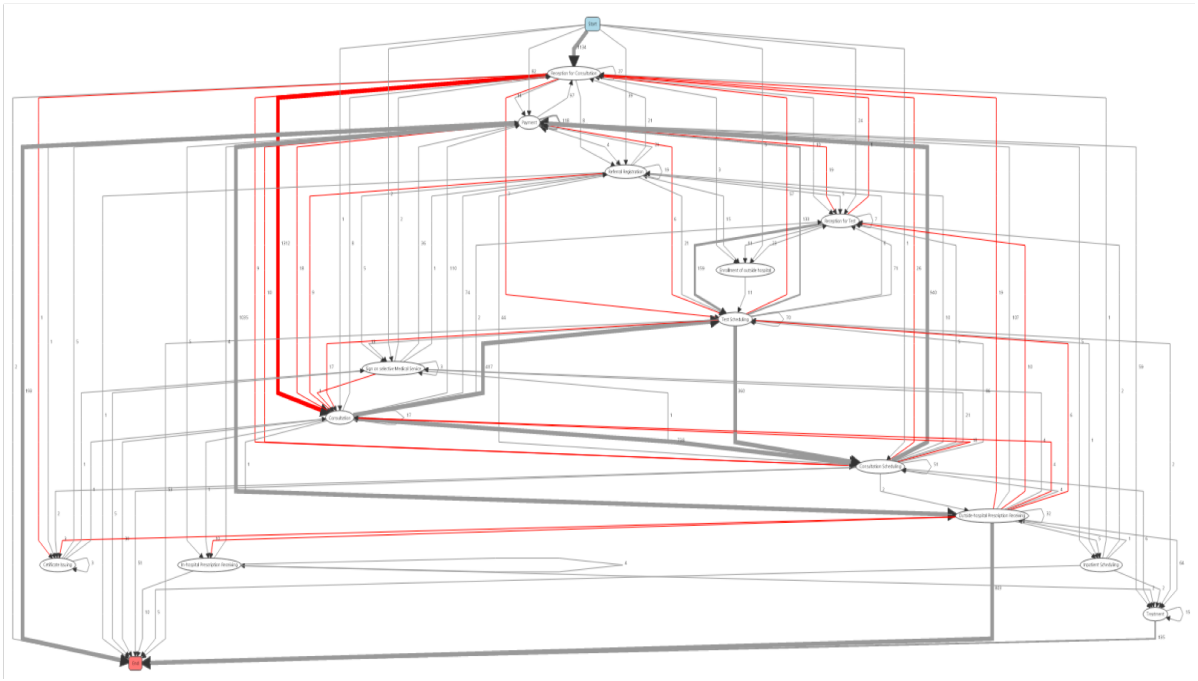


Figure 35. The discovered outpatient process of doctor A

Figure 36. The discovered major outpatient flow of doctor A

Second, we calculated the average number of appointments for each reservation slot and the patients' visiting time compared to the reservation time for the arrival rates of patients. In the event log, there were 19 slots in a session which were set up every 10 minutes from 9 a.m. to 12 p.m. Table 18 shows the number of appointments per each slot depending on the patient types: new patients and follow-up patients. The average number of the new and the follow-up patients per slot was 3.71 and 0.55 respectively. In total, about 81 patients visited the hospital to get a consultation from the doctor A in a session on average.

After that, we calculated the visiting time compared to the reservation time for deriving the arrival rates. Figure 37 depicts the result that patients visited the hospital 7.52 minutes earlier than the booked time on average. Also, 837 patients (65%) arrived early at the hospital, and 451 patients (35%) were late compared to the reservation time. These two results were applied as the arrival rates in the simulation model.

Lastly, we calculated the service time for consultation using the suggested approach. The average and median of the consultation service time were 3.35 and 2.68 minutes, respectively. More specifically, there was a difference according to the patient type as 3.33 minutes for follow-

Table 18. The average number of appointments of each reservation slot

Reservation slot	New Patients	Follow-up Patients	Sum
9:00:00	0.25	5	5.25
9:10:00	0.75	4.42	5.17
9:20:00	0.75	4.75	5.5
9:30:00	0.59	0.33	0.92
9:40:00	0.41	4.58	4.99
9:50:00	0.66	4.42	5.08
10:00:00	0.75	4.25	5
10:10:00	0.83	4.42	5.25
10:20:00	0.92	4.25	5.17
10:30:00	0.58	0.5	1.08
10:40:00	0.92	4.17	5.09
10:50:00	0.83	4.5	5.33
11:00:00	0.33	4.75	5.08
11:10:00	0.5	4.33	4.83
11:20:00	0.25	5.08	5.33
11:30:00	0.66	8.08	8.74
11:40:00	0.42	0.75	1.17
11:50:00	0.08	0.75	0.83
12:00:00	0.16	1.08	1.24
Average	0.55	3.71	4.26
Sum	10.64	70.41	81.05

up patients and 3.56 minutes for new patients on average. To make a precise simulation model, we applied the service time depending on the types of patients.

5.3.3 Simulation modeling & evaluation results

Based on the process mining analyses results, the simulation model was created, which covers 19 reservation slots from 9 a.m. to 12 p.m. for each session. To validate the model, we performed the as-is simulation analysis. Table 19 shows the evaluation results between the calculated KPIs from logs and simulation analyses with 500 runs. First, the average of the consultation waiting time (*KPI1*) and the controllable waiting time for consultation (*KPI2*) from the simulation analysis was 36.41 and 31.36 minutes, respectively. Also, the average of the end time of the clinical session for the doctor (*KPI3*) from the simulation analysis was 12:54:28 PM, which displayed a 6-minute time difference with the logs. After that, we conducted a further evaluation using MAPE, and

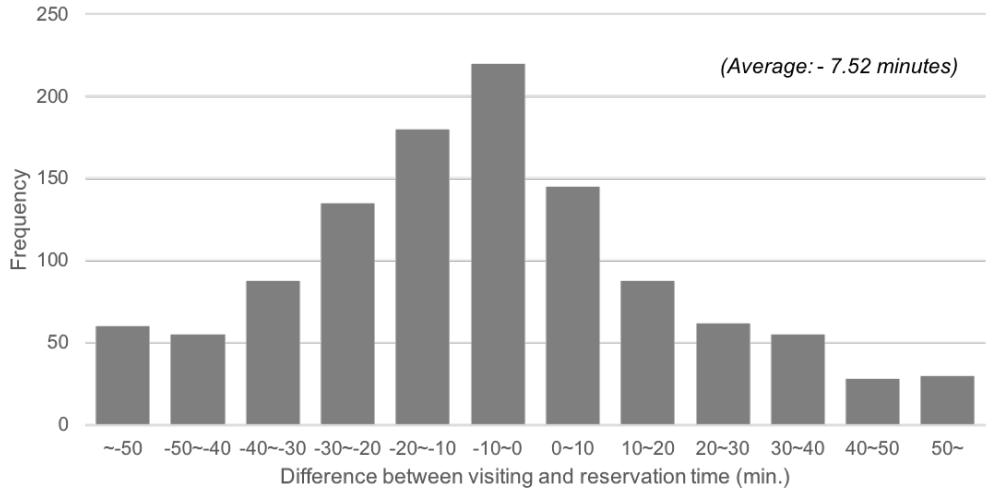


Figure 37. The distribution of the difference between visiting time and reservation time

it showed that MAPE values of three KPIs were less than 3.5%, which we take as an indication that our simulation model closely resembles the real patient process. In other words, the model is suitable to conduct the to-be simulation analyses.

Table 19. The evaluation results between event logs and simulation models using KPIs

Runs: 500 times	KPI1: Waiting time time for consultation		KPI2: Controllable waiting time for consultation		KPI3: End time of the clinical session	
(Unit: min.)	Logs	Simulation	Logs	Simulation	Logs	Simulation
Average	35.09	35.04	31.13	30.08	12:48:52	12:54:28
UCL(95%)	33.71	37.33	29.53	28.09	12:40:09	12:51:11
LCL(95%)	36.47	32.75	32.73	32.66	12:57:35	12:57:44
MAPE	0.14%		3.37%		0.73%	

5.3.4 Experimental simulation analysis results

As a result of the BP-based data analysis, we prepared four scenarios to decrease the waiting time: decreasing the number of appointments per reservation slot, making a break time in the middle of the clinic session, rearranging patients' reservation, and subdividing reservation intervals. A graphical explanation is provided in Figure 38. Among them, the first two scenarios were relevant with *Extra Resources*, increasing the number of resources in a process, of 29 best practices. Also, the third and the fourth scenario were constructed based on *Case Types* (i.e., distinguishing the process considering a type of cases) and *Order-based Work* (i.e., eliminating batch-processing and periodic activities), respectively. For each scenario, we give the detailed

explanation of how it was created.

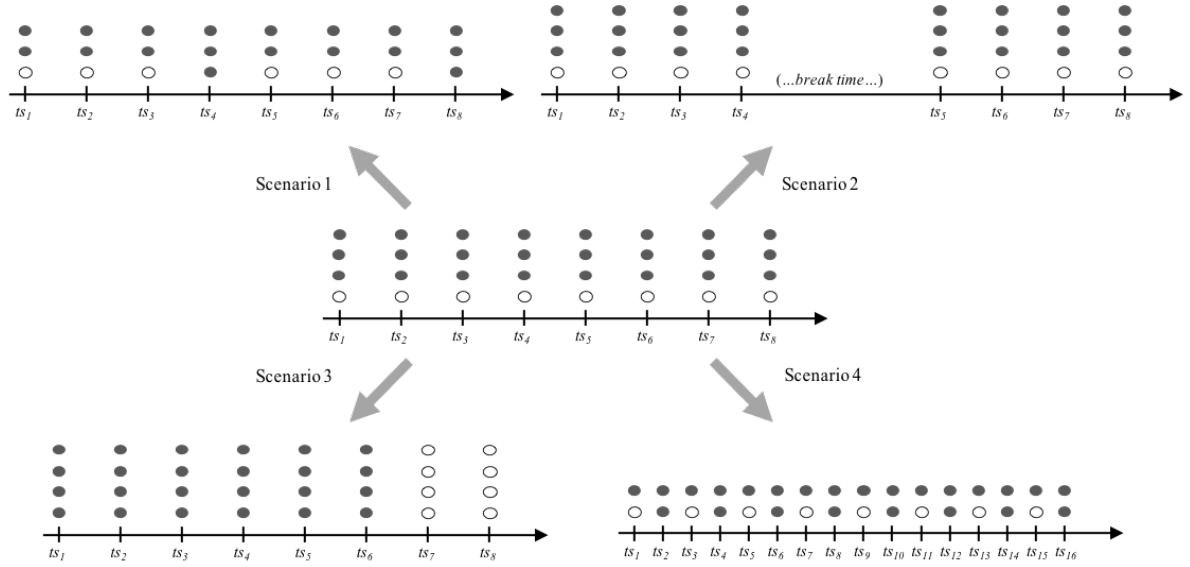


Figure 38. Four graphical to-be simulation scenarios

1) Decreasing the number of appointments per reservation slot

One of the most influenceable factors to waiting time is the number of appointments per slot, that is to say, the number of patients. Assuming that other conditions are same, it is evident that the fewer patients assigned to a doctor, the fewer time patients have to wait. In the left upper example in Figure 38, we give a graphical explanation of the first scenario. In scenario 1, we tried to figure out how much waiting time is decreased as the number of patients declines from 5% to 25% at intervals of 5%.

2) Making a break time in the middle of clinical session

From the event logs, we discovered a trend that the waiting time is on the rise as the time closer to the end of the clinic session. Figure 39 represents an example of an average waiting time of patients who involved in each slot in a clinic session. The figure shows that the average waiting time becomes around 80 minutes at the ending session, while the value is less than 20 minutes within the first 20 time slots. One of the solutions would be to decrease the service time, but it is not realistic because hospitals have to consider patients' satisfaction and there is only so much we can decrease it. In scenario 2, to reduce the number of waiting patients, we implemented an alternative solution which creates a break time in the middle of the clinic session. In the right upper example in Figure 38, we provide a graphical explanation of the second scenario. Simulation analyses were performed as we inserted a break time of 5 to 25 minutes at intervals of 5 minutes. We tried to figure out how much waiting time is decreased as the break time increases from 5 to 25 minutes at intervals of 5 minutes.

3) Rearranging patients' reservation

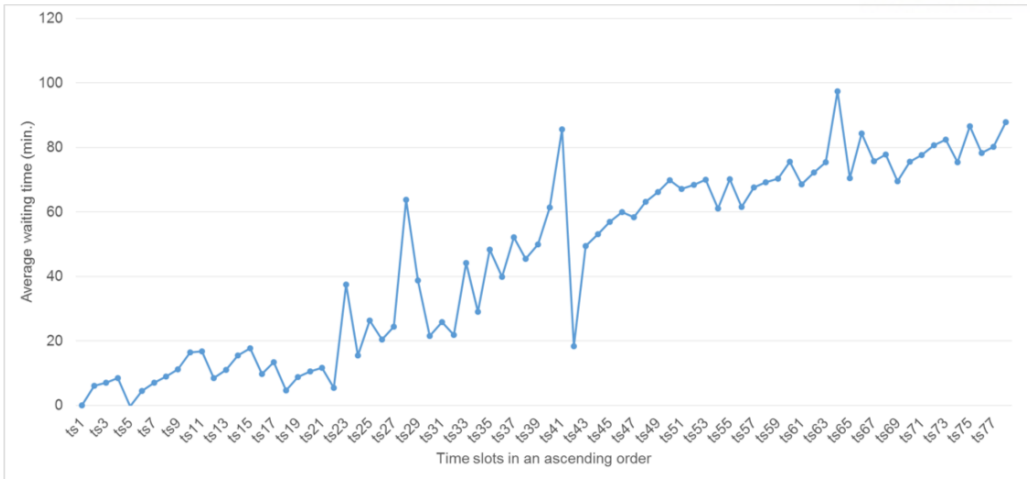


Figure 39. An example of average waiting time of each time slot in a clinical session

The third plan is also a solution to cope with the problem of scenario 2, the cumulated waiting patients. From the event logs, we checked out that the consultation service time is depending on the patient type. The patients who visited the hospital for the first time had longer service time than the follow-up patients because the patients should be newly observed with more time. Based on the trend, we rearranged the patients' reservation as the follow-up patients in the beginning and the new patients in the ending of the session. For example, in Figure 38, suppose that white dots represent the patients who had longer service time. As shown on the left below example (i.e., scenario 3), we can make a scenario to reduce the cumulated waiting time by rearranging as white dots at the ending and gray colors in the beginning side.

4) Subdividing reservation intervals

The last scenario for decreasing the waiting time is subdividing the number of slots. In the hospital, there was a trend on the batch-shaped consultation registration due to their patient reservation system. Figure 40 is the dotted chart of two tasks, where red and green dots represent consultation registration and consultation, respectively. In the figure, the y-axis and the x-axis are configured as patients and actual time, respectively, and the rows are sorted by consultation. In the black boxes of the figure, we can identify that multiple registrations on consultation were performed within a few minutes (i.e., *batch-processing*). As a result, the patients who registered relatively later than others in the same slot had to wait more to get the consultation. To solve this problem, we changed the system which has reservation slots in every 5 minutes, that is the number of appointments per each slot is decreased as well. In the right below example in Figure 38, we explain the fourth scenario graphically.

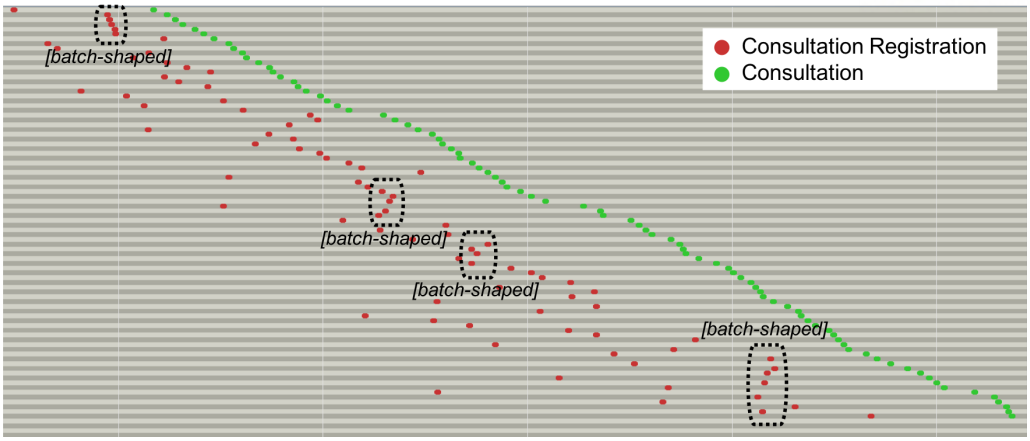


Figure 40. Dotted Chart Analysis – The batch shape of consultation registration

Table 20. Scenario-based simulation analysis results

Scenario 1: Decreasing the number of appointments per reservation slot						
	Current	-5%	-10%	-15%	-20%	-25%
KPI1 (min.)	35.04	32.15	27.81	22.01	17.44	15.70
(Rate of change(%))	(−)	(-8.24)	(-20.62)	(-37.20)	(-50.23)	(-55.20)
KPI2 (min.)	30.38	27.11	22.90	17.23	12.83	10.65
(Rate of change(%))	(−)	(-10.77)	(-24.63)	(-43.30)	(-57.78)	(-64.96)
Scenario 2: Making a break time in the middle of clinical session						
	Current	-5 min.	-10 min.	-15 min.	-20 min.	-25 min.
KPI1 (min.)	35.04	33.73	32.44	29.43	27.12	25.24
(Rate of change(%))	(−)	(-3.74)	(-7.41)	(-16.01)	(-22.59)	(-27.97)
KPI2 (min.)	30.38	29.02	27.60	24.73	22.05	20.09
(Rate of change(%))	(−)	(-4.48)	(-9.16)	(-18.59)	(-27.41)	(-33.85)

Scenario 3: Adjusting patient's reservation		
	Current	*Adjusted*
KPI1 (min.)	35.04	34.04
(Rate of change(%))	(−)	(-2.84)
KPI2 (min.)	30.38	29.53
(Rate of change(%))	(−)	(-2.81)

Scenario 4: Subdividing reservation intervals		
	Current	*Adjusted*
KPI1 (min.)	35.04	33.40
(Rate of change(%))	(−)	(-4.69)
KPI2 (min.)	30.38	28.93
(Rate of change(%))	(−)	(-4.78)

5) To-be simulation analysis

As we explained earlier, we performed the simulation analyses based on four scenarios which decrease the consultation waiting time. To measure the impacts of each scenario, the waiting time for consultation (KPI1) and the difference between the start time of consultation and the reserved time (KPI2) were used among three KPIs which were applied to evaluate the simulation model. Table 20 represents the simulation analyses results in each scenario. First, in scenario 1, both KPI1 and 2 were significantly decreased as the number of appointments per each reservation slot decreased. The reduction of 25% in scenario 1 caused a reduction of about 55% and 65% for KPI1 and 2, respectively. Second, making a break time in the middle of a clinical session (scenario 2) moderately reduced KPI1 and 2; a decrease of 25% led to the reduction of about 28% in KPI1 and 34% in KPI2. Lastly, in scenario 3 and 4, KPI1 and 2 slightly decreased due to the adjustments in reserving patient slots and subdividing the reservation intervals.

5.3.5 Organizational Relevance

Through the simulation analysis of scenario 1, it becomes evident that the number of patients has a significant impact on the consultation waiting time. Also, the measured values also decreased due to inserting a break in the middle of the session (scenario 2). That is, the methods in scenario 1 and 2 can be considered as highly substantial improvements of a clinician's appointment schedule when the goal is to decrease patients' waiting time. However, these methods potentially negatively affect hospital revenue. After all, the average number of consulted patients per day decreases. Therefore, whether these scenarios are attractive to pursue the hospital in question depends on multiple factors.

Interestingly, the arrangement of patient groups according to scenario 3 and subdividing reservation intervals as in scenario 4 do not incur additional expenses. They simply affect the reservation policies without diminishing the number of patients or requiring doctors to spend more time. Even though the effects of scenario 3 and 4 are relatively small compared to those of scenario 1 and 2, they may be worthwhile to pursue.

After discussing with domain experts in the hospital, we received comments on the results of our simulation analyses. They considered that the methods of case 3 and 4 to change the individual clinicians' schedules are indeed applicable and attractive.

5.4 Summary and Discussion

In this chapter, we suggested a decision support framework for optimizing clinician medical scheduling using discrete event simulation approach, which is constructed based on three process mining analysis including process discovery, arrival rate analysis, and service time analysis. Furthermore, it covered how to derive effective improvement methods to decrease waiting time for consultation. In the case study, we applied the real-world data to the proposed framework.

Also, we performed the four scenario-based experiments using the simulation model for the personalized care scheduling. As a result, we showed that not only two cases which need additional costs have a significant effect on the waiting time, but also the changes of reservation systems which do not require more costs decreased the waiting time.

The main strength of the proposed approach is that it is data-driven and highly automated. In that sense, it considerably simplifies the application of DES in a clinical setting. From a perspective of innovation, the arrival rate analysis and the service time analysis based on process mining are innovative new methods. Both of these consider the specific characteristics of hospitals and the data they have at their disposal. Based on our approach, simulation models can be utilized in diverse healthcare settings to determine improved personal schedules for clinicians. Also, our approach provides a systematic method that overcomes the limitations of the existing works for scenario-based simulation analysis. This helps to break away from the traditional rule-of-thumb approach and reduces computing time and power with efficient simulation analysis.

Also, our approach has extensive flexibility. In this chapter, we focused on how to solve the problem of optimizing personal clinical schedules. However, our approach can handle other processes in the healthcare environment such as clinical test or reception processes. In addition, it can support other service processes such as banking and public office task processes similar to the outpatient process.

Our work also has several limitations. As far as the proposed framework is concerned, it still needs further automated approaches. In the framework, for example, it is relevant to create a simulation model from process mining analysis results or prepare an improved simulation model that reflects the scenario. In particular, techniques that automatically reflects the improvements based on redesign best practices in the simulation model can maximize the effectiveness of the simulation analysis. Also, this chapter covers a single case study to validate our framework. Future research should strive to conduct further case studies. Lastly, we are working to develop the decision support system that supports our framework. It will be helpful for practitioners for effective hospital management.

VI Evaluating Effects of Process Redesigns in Healthcare

This chapter proposes a business process assessment framework focused on the process redesign and tightly coupled with process mining as an operational framework to calculate indicators. Specifically, Chapter 6.1 introduces the background and motivation of the research subject in this chapter. Chapter 6.2 presents the overview of the proposed framework. The primary two concepts of this framework, i.e., *best practice implementation indicators* and *process performance indicators*, are introduced in Chapter 6.3 and 6.4, respectively. These chapters include the detailed explanation and formal definition for each indicator. Chapter 6.5 validates the usefulness by providing the case study result with a real example. Finally, Chapter 6.6 concludes this chapter.

6.1 Background

While several frameworks defining measurements for business process evaluation have been proposed in the literature [164–167], they suffer from the following limitations; it is not specified in depth which type of data should be collected to calculate indicators and whether that is feasible or not, and they do not focus on evaluating of business process redesigns.

To overcome these challenges, this chapter proposes a new framework of business process performance indicators. Figure 41 presents the overview of the redesign assessment methodology. Here, it starts with the process redesign heuristics (i.e., redesign best practices) suggested by Reijers and Mansar [17]. The methodology includes two sets of indicators: (i) one to clearly identify the implementation of the best practice, i.e., *Best Practice Implementation indicators* (BPIs), and (ii) one to assess process improvements yielded by its application, i.e., *Process Performance Indicators* (PPIs). In this way, the proposed methodology gives an evidence-based support to the entire business process redesign phase, covering both redesign implementation (with BPIs) and more traditional process improvement evaluation (with PPIs). The proposed framework considers process mining [36,37,40,45,48–51] as the underlying evidence-based process analysis technology. Therefore, for both types of indicators, we define how they can be calculated using process related data, i.e., event logs, using standard process mining functionality [36, 37, 40, 45, 48–51]. In doing so, we also implicitly identify what kind of process data must be collected to calculate BPIs and PPIs.

The proposed framework is relevant both from a research and a practical standpoint. From the research standpoint, having scientific methods to assess the benefits of BPR linked to applications of best practices increases the reliability of the knowledge base about BPR best practices accumulated thus far in the literature. While many studies advocate the use of quantitative and evidence-based mechanisms to assess business process performance [165], the assessment of BPR best practices and their effect on process performance is often qualitative, based on second-hand data, such as executive and user surveys [17, 168]. As recognized by other authors, e.g. [169], a

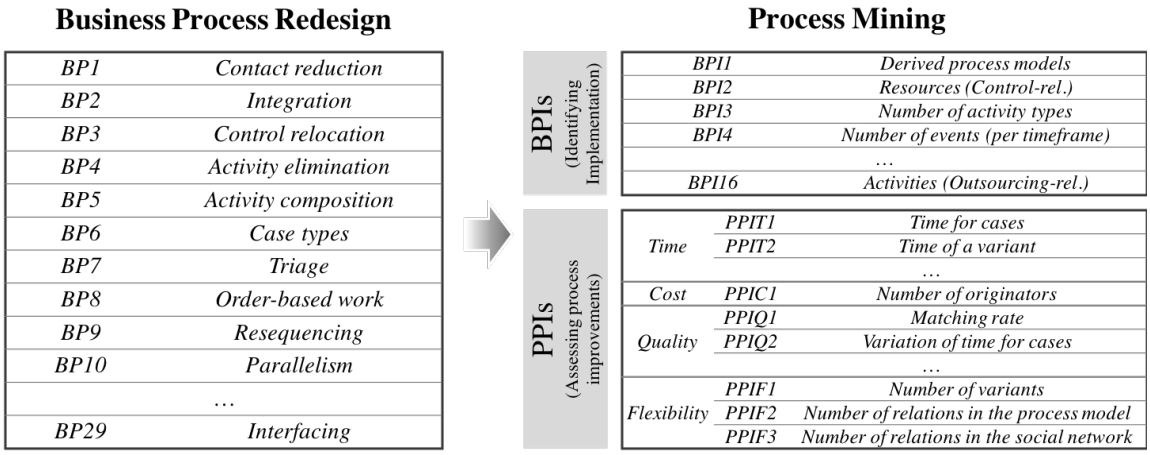


Figure 41. BPIs and PPIs for the redesign assessment

methodology to link BPR best practices to clearly defined, measurable, and repeatable PPIs is currently lacking. From a practical standpoint, the proposed framework gives process analysts and decision makers actionable tools to assess the results of their choices in BPR initiatives.

6.2 An Overview of Indicators for Assessing The Effects of Redesigns

In defining evaluation measures for best practices, our approach has a twofold goal. The first objective is to assess whether a specific best practice has been applied for a redesign. To understand whether a specific effect originates from using the best practice or other factors, in fact, it is important first to be certain that a best practice has been implemented. In this regard, we define BPIs for each of the 29 best practices identified by Reijers and Mansar [17]. The second goal is to comprehend the impact of the application of best practices when redesigning a business process. In this regard, as previously discussed, we consider the performance dimensions: time, cost, quality, and flexibility. A summary of all best practices, BPIs, and PPIs is shown in Table 21. The table provides what PPIs can be applied for each best practice. Also, applicable PPIs (e.g., *PPITs*, *PPICs*, *PPIQs*, *PPIFs*) are defined based on the four dimensions. Here, all PPIs can be employed for each best practice, while only a couple of BPIs is applied. In addition, we give potential effects (e.g., *positive*(+), *negative*(−), *neutral*(•)) of each redesign item in four dimensions suggested by Reijers and Mansar [17].

Table 21. Summary of BPIs and PPIs

Category	BP	BPIs	PPIs			
			Time (*PPIT1∼5*)	Cost (*PPIC1*)	Quality (*PPIQ1∼4*)	Flexibility (*PPIF1∼3*)
Customers	Contact reduction	Derived process models (*BPI1*)	+	−	+	•
	Integration	Derived process models (*BPI1*)	+	+	•	−
	Control relocation	Resources who perform the control-related activity (*BPI2*)	•	−	+	•
Business process operation	Activity elimination	Number of activity types (*BPI3*)	+	+	−	•
	Activity composition	Number of activity types (*BPI3*)	+	+	•	−
	Case types	Derived process models (*BPI1*)	+	+	−	−
	Triage	Derived process models (*BPI1*)	•	−	+	−
	Order-based work	Number of events for each timeframe (*BPI4*)	+	−	•	•
Business process behavior	Resequencing	Derived process models (*BPI1*)	+	+	•	•
	Parallelism	Derived process models (*BPI1*)	+	−	•	−
	Knock-out	Derived process models (*BPI1*)	−	+	•	•
	Exception	Derived process models (*BPI1*)	+	−	+	−
Organization	Case assignment	Number of resources for each case (*BPI5*)	•	•	+	−
	Numerical involvement	Number of resources for each case (*BPI5*)	+	−	•	−
	Split responsibilities	Number of events performed by each resource for activities (*BPI6*)	•	•	+	−
	Flexible assignment	Number of events performed by each resource for activities (*BPI7*) → Allocated resources for each timeframe (*BPI8*)	+	−	•	+

	Specialist-generalist	Number of events performed by each resource for activities (*BPI7*) → Specialist-Generalist ratio (*BPI9*)	+	•	+	−
	Customer teams	Derived social networks (*BPI10*)	•	•	+	−
	Extra resources	Number of resources (*BPI11*)	+	−	•	+
	Empower	Derived process models (*BPI1*) and derived social networks (*BPI10*)	+	•	−	+
	Centralization	Workloads for each resource (*BPI12*)	+	−	•	+
	Case manager	Whether there exist any activities related to subscribing (*BPI13*)	•	−	+	•
Information	Control addition	Derived process models (*BPI1*)	−	−	+	•
	Buffering	Whether there exist any activities related to subscribing (*BPI14*)	+	−	•	•
Technology	Task automation	Whether resources appear in the automated activity (*BPI15*)	+	−	+	−
	Integral technology	Whether there exist any changes from technologies (*BPI16*)	+	−	•	•
External	Trusted party	Whether there exist any activities related to obtaining information from outside (*BPI17*)	+	+	•	−
environment	Outsourcing	Derived process models for internal party (*BPI1*)	+	+	•	−
	Interfacing	Not applicable	+	•	+	−

Figure 42 presents the data analysis framework for evaluating the effects of redesigns. In the data preparation & preprocessing steps, clinical event logs after applied the redesign are collected and preprocessed to make a suitable input for data analysis. After that, performance analysis is conducted based on the best practice implementation indicators and process performance indicators explained above. The last step, the post-hoc analysis, in this framework has a goal of evaluating the benefits of redesigns. In such a process, performance analysis results before BPR are employed. That is, on the basis of performance analysis results before and after the redesign, comparison analyses are performed in the post-hoc analysis phase.

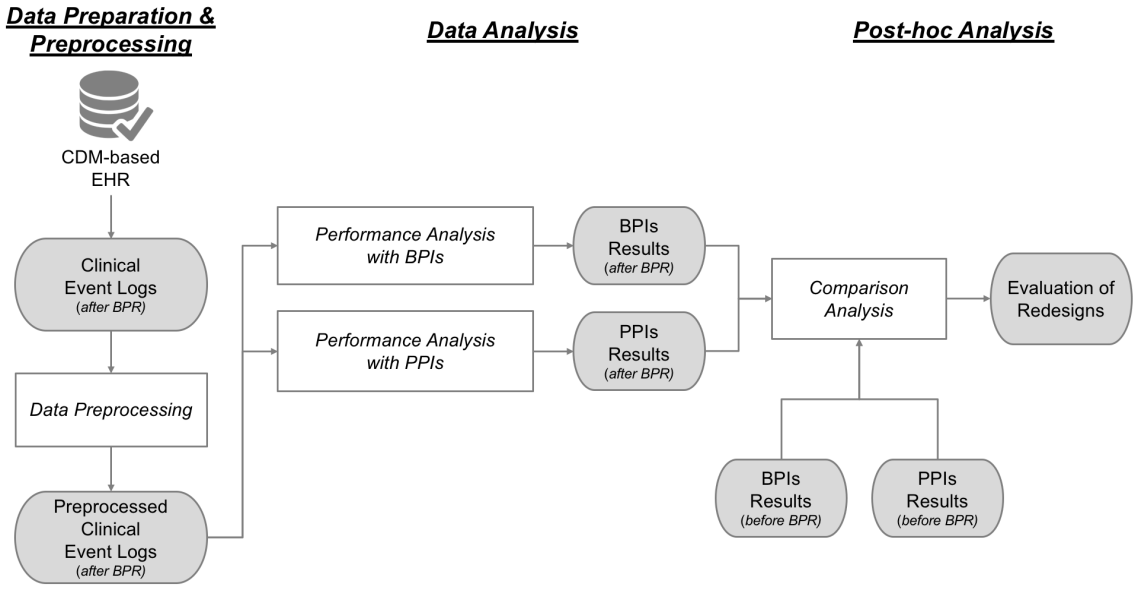


Figure 42. The overview of the redesign assessment framework

6.3 BP Implementation Indicators (BPIs)

As provided in Table 21, we define 17 BPIs for 29 best practices. For each indicator, we also suggest suitable process mining techniques through which it can be calculated. Note that information in event logs for process mining may not be able to cover all possible BPIs. When this is the case, we suggest which additional information is needed to measure the implementation of redesigns.

6.3.1 Customer

Contact reduction concerns decreasing the number of communications with customers and *integration* refers to combining an existing process with a business process of customers. These best practices are related to a change of workflows; thus, they lead to a change of a process model. More in detail, *contact reduction* removes repetitive loops from the process, while *integration* removes customer-related activities or sub-processes from a process model. Therefore, identify-

ing the application of the contact reduction and the integration best practices can be checked by comparing discovered process models (*BPI1*) before and after redesigns.

Control relocation is defined as transferring controls towards customers. The most obvious evidence of the application of this best practice is that customers, instead of internal employees, perform control-related activities in the to-be process. Thus, we need to interrogate the originator information of the control-related activities (*BPI2*). Process mining functionalities provide the *Linear Temporal Logic (LTL) checker* [170] that enables to check the satisfaction of LTL constraints in a process. For *control relocation*, the following constraint can be applied: *eventually* ((*activity* == *"some control-related activity"*) ∧ (*resource* == *"customers"*)). Moreover, other resource perspective techniques such as the *organizational model mining* [50] or the *originator by task matrix* [49] can also be used to check the implementation of this best practice. Note that activities in the event log should be classified in control-related and non-control-related.

6.3.2 Business process operation

Activity elimination implies removing unnecessary activities, while *activity composition* indicates integrating low-level activities into a combined activity. The application of these best practices leads to a change of the number of activity types in the process (*BPI3*). Therefore, the *log summary* [37] can be used, since it gives an overall summary of the behaviors in an event log. The log summary results would provide a decrease of the value for *activity elimination* and an increase of that for *activity composition.*

Case types distinguishes a new process when activities or sub-processes appear for a specific type of cases. Assume that a series of activities in a business process are differentiated based on two types of cases. If this best practice is implemented, it is possible to divide a process into two different processes. Therefore, *control-flow mining algorithms* [36] can be used to check the implementation of this best practice.

Triage separates a common activity into several alternative activities considering the abilities of resources or types of cases. Thus, process instances after redesign can select one of the alternative activities instead of the common activity in the as-is process. As such, the application of this best practice generates changes in the control flow of a process. More in detail, several alternative activities will appear after the redesign and these will be connected by XOR-split/join gateways in the process model. Therefore, comparing discovered process models (*BPI1*) is the key method to determine the implementation of the *triage* best practice.

Order-based work eliminates batch-processing and periodic activities in a process. To check its implementation, the number of batch-processing activities needs to be measured in a process for each timeframe (*BPI4*). For example, if a hospital eliminates a test activity at a specific time window in the as-is setting, e.g., between 9am and 10am, the activity is no longer quite frequent in that time frame in the to-be process model. Process mining provides the *basic performance analysis plugin* [37] that provides information about the frequency of events in every period (i.e.,

day-hour chart). Similar information is also given in the *dotted chart* [51]. In the chart, batch activities can be recognized by time frames crowded with several dots of the same type (e.g., color).

6.3.3 Business process behavior

The application of all the best practices in this category generates variations of process models. Therefore, the implementation can be checked by comparing as-is and to-be process models (*BPI1*).

Resequencing concerns adjusting the ordering of activities. In general, this best practice recommends moving an activity to a more appropriate place in the process, e.g., next to other activities performing similar actions in a process. For instance, once this best practice is applied, in the to-be process we will be able to observe a sequence relationship between the activity and the other activities similar to it.

Parallelism implies to put activities in parallel when possible. Thus, if the *parallelism* is applied, the relationship between activities in the process model changes from the sequence to the parallel. This can be observed in the to-be process model.

Knock-out concerns controlling the order of knock-out activities, i.e., activities that could terminate the execution of a process. In practice, this best practice is similar to the *resequencing* best practice, since it proposes to adjust the position of specific types of activities, e.g., knock-outs. Differently from *resequencing*, however, both the locations of knock-outs in a process model and the termination probability of each knock-out activity should be examined. Based on these measures, it should be checked whether the termination probability is higher as the knock-out activity is put closer to the start.

Exception implies to isolate exceptional cases in a business process. Identifying the application of the *exception* is similar to *integration*, since it makes newly added activities or sub-processes for exceptional cases that do not exist in the as-is process model. Therefore, it requires identifying the presence of newly added activities or sub-processes for exceptional cases in the to-be process model.

6.3.4 Organization

Case assignment concerns making resources perform as many activities as possible in a case. Checking the implementation of this best practice requires measuring the number of resources involved per case. As a result of applying the best practice, a smaller number of resources work together in an individual case. The number of resources involved per case (*BPI5*) can be obtained from the basic performance analysis [37].

Numerical involvement concerns minimizing the number of resources in a business process. Similar to *case assignment*, the number of resources involved per case (*BPI5*) can be calculated

to examine the implementation of this best practice. As such, the number of resources involved per case decreases.

Split responsibilities concerns letting resources perform different activities and have different roles in a business process. Thus, as a result of this best practice, responsibilities in the process will be separated. To check the implementation of this best practice, the number of events executed by each resource for activities (*BPI6*) should be analyzed. This can be done using the *originator by task matrix* [49] in process mining. If resource roles are clearly separated each other, it yields that different resource groups conduct different activities.

Flexible assignment concerns resource allocation so that flexibility can be maximized in the near future. In other words, it represents that it is better to assign works to specialists before considering generalists. Checking the implementation of this best practice requires a prerequisite step that divides originators into specialists and generalists. The *originator by task matrix* [49] can be used to perform this step: in the matrix, specialists will be involved in a limited number of specific activities, whereas generalists will be included in several different activities (*BPI7*). Once the separation between specialists and generalists has been performed, the *dotted chart* [51] can be used to investigate which type of resource is allocated first for the maximum flexibility (*BPI8*).

Specialist-generalist concerns controlling the specialist-generalist ratio in a business process. Thus, in common with the *flexible assignment*, a prerequisite step is to separate specialist from generalist roles or resources (*BPI7*). Then, the specialist-generalist ratio is calculated for the as-is and to-be process and compared (*BPI9*). When the implementation of this best practice is considered, organizations predetermine the proper specialist-generalist ratio based on their situations. Therefore, for this best practice, it should be checked whether or not the calculated value is different from the expected value in planning BPR.

Customer teams concerns composing resource groups from different departments to handle specific types of cases entirely. Checking the application of this best practice requires analyzing the as-is and to-be social networks (*BPI10*). If a working group cooperates to handle a single case, *handovers of works in the social network* [50] occur within the working group only. In other words, as a result of the implementation of customer teams, the derived social network shows separate working groups.

Extra resources entails increasing the number of resources in a process. As a result of the application of *extra resources*, the total number of resources (*BPI11*) in a process increases. The total number of resources involved in a process is shown in the *log summary* [37].

Empower concerns removing middle management by providing decision-making roles to workers at lower levels. The effects of this best practice are twofold. First, middle management decision-making tasks in a business process are removed. Second, as the middle management disappears, handovers of work among resources are changed. More in detail, the handovers of work related to activities executed by middle management-oriented in the to-be social network

decrease. Thus, as-is and to-be process models (*BPI1*) must be compared to detect the elimination of middle management decision steps, e.g., a test or an inspection activity, and as-is and to-be social networks (*BPI10*) should be compared to detect changes in handovers of work.

Centralization concerns considering resources as if they are centralized. Assume that there is a business process where resources in each location can perform limited types of activities. If the *centralization* best practice is implemented, these limitations will be removed. Therefore, checking the implementation of this best practice requires additional information about the location of resources. Then, based on the *originator by task matrix* [49] and the location information, we can check whether the works are distributed regardless of location information after applying the best practice (*BPI12*).

Case manager concerns designating a resource responsible for a particular case type. Checking the implementation of this best practice requires a particular attribute in event logs identifying the case manager belonging to individual cases. If this information is in event logs, then the case manager implementation can simply be checked by using the *LTL checker* [170] as follows: *eventually* (*case-manager attribute* != $\emptyset$) (*BPI13*).

6.3.5 Information

Control addition concerns adding control-related activities to check the completeness of inputs and outputs in a process by adding appropriate activities or sub-processes. To determine the implementation of the redesign, we need to compare the as-is and to-be process models (*BPI1*). In particular, looking for additional control-related activities in the to-be model is required for the *control addition* best practice.

Buffering entails subscribing to updates instead of requesting information when possible. An effective way to check the application of the best practice is to utilize the *LTL checker* [170] considering the following constraint: *eventually* (*activity* == *"some subscribing-related activity"*) (*BPI14*).

6.3.6 Technology

Task automation concerns creating activities automated when possible. The execution of automated activities is not associated with any human resources. Therefore, the implementation of this best practice can be examined using the following constraint in the *LTL checker* [170]: *eventually* ((*activity* == *"automated activity"*) $\wedge$ (*resource* == $\emptyset$)) (*BPI15*). Also, we can assess the implementation of this best practice using the *originator by task matrix*, by investigating resources of automated activities.

Integral technology concerns applying new technology for elevating physical constraints. Given that the implementation of new technology may concern a range of new possibilities, it is impossible to devise a precise way of checking the implementation that accounts for all possible scenarios. However, we argue that technology should at least have an impact on the

information in event logs, introducing, for instance, new activities and/or new and more precise information that can be logged (*BPI16*). Therefore, qualitatively comparing as-is and to-be event logs can at least reveal whether a change has occurred in the process. If the as-is and to-be logs contain the same information, then we can affirm that the new technology has not been implemented or, at least, it is not applied appropriately in the process.

6.3.7 External environment

Trusted party concerns using results from a trusted party instead of determining information oneself when possible in a process. The application of this best practice can be examined by analyzing whether or not there exist activities in a process that obtain information from outside. This can be monitored through *LTL checker* [170] as given: *eventually* (*activity* == *"obtaining outside information-related activity"*) (*BPI17*).

Outsourcing concerns contracting out a (part of a) business process. This can be checked by comparing as-is and to-be process models (*BPI1*). In particular, only events involving internal employees are likely to appear in an event log. Hence, through event logs it is only possible to check whether a process or part of it is no longer executed and assume that this means that it has been outsourced.

Interfacing concerns developing a standardized interface with customers. We argue that the implementation of this best practice cannot be checked using process mining techniques because it just concerns modifying the way in which communication with customers occurs, but it does not change the nature of this communication. Therefore, the event logged by IT systems supporting communication with customers are not likely to change.

6.4 Process Performance Indicators (PPIs)

Here, we suggest 13 PPIs on the basis of four process performance measures explained by Reijers and Mansar [17]. In this chapter, we give a detailed explanation on PPIs including how to measure them.

6.4.1 Time

Most BPR efforts aim at increasing the efficiency of business processes by improving time-related indicators, such as decreasing processing time and waiting time. In the proposed methodology, we suggest 5 indicators in the time perspective as descibed in Table 22. All time-related indicators require a basic measure and can be aggregated using standard aggregation functions. In these indicators, the operation time is the actual process time of an activity, and waiting time is the time between the end of the previous activity and the start of the current activity.

Definition 13 (Sequential time point, status) *Let* $TP_k = \{tp_{k,1}, tp_{k,2}, tp_{k,3}, \ldots, tp_{k,p}\}$ *be the finite set of sequential time points of the k-th case, where* $t_{k,1} = tp_{k,1}, t_{k,n} = tp_{k,p}$. *Let*

Table 22. Process Performance Indicators in Time Perspective

PPI#	Explanation	Measure	Aggregation Function
PPIT1	Time for cases in a log	Cycle/Operation/Waiting Time	AVG, MED, MAX, MIN
PPIT2	Time of a variant (v_1)	Cycle/Operation/Waiting Time	AVG, MED, MAX, MIN
PPIT3	Time of an activity (a_1)	Cycle/Operation/Waiting Time	AVG, MED, MAX, MIN
PPIT4	Time for events performed by an originator (o_1)	Cycle/Operation/Idle Time	AVG, MED, MAX, MIN
PPIT5	Time for events performed by an originator (o_1) for an activity (a_1)	Cycle/Operation/Idle Time	AVG, MED, MAX, MIN

$St_c = \{working, waiting\}$ *be a set of case status.*

- status: $tp_k \rightarrow St_c$ *is a function mapping each time point to a status*

(*e.g.* $status(tp_{k,1})$ *is the status in first-time point of the k-th case*)

According to Definition 13, the time between the minimum and the maximum timestamp in an event log of events belonging to the k-th case is divided into p intervals, identified by the sequential time points $tp_{k,p}$. The status function returns *working* or *waiting* depending on whether, in a given time interval, the case was in a working status or a waiting status, respectively. For example, in Figure 43, case 1(c_1) has three events and 10 time points from the minimum timestamp (i.e., $tp_{1,1}$, the start time of $e_{1,1}$) to maximum timestamp (i.e., $tp_{1,10}$, the complete time of $e_{1,3}$). In this example, case 1 is in status *working* at $tp_{1,1}, tp_{1,2}, tp_{1,3}, tp_{1,5}, tp_{1,6}, tp_{1,7}, tp_{1,8}, andtp_{1,9}$; thus, the status function returns working at the corresponding time. On the other hand, case 1 is in status *waiting* at $tp_{1,4}$ and $tp_{1,10}$ because there is no event in the event logs with timestamp belonging to this time interval. Based on users' preferences, the size of intervals between consecutive time points can be varied. If a large number of time points in a given unit of time is chosen, users can get more accurate values, the expense of higher computational power required to process results. The purpose of defining time points and status of the process instance in the time interval is that we can sample operation time and waiting time for cases without learning process models. In general, most process models have concurrency, i.e., activities may be executed in parallel. When concurrency is defined, if a process model is available, operation time and waiting time for cases considering flows of models can be calculated. Sequential time points and status, however, allow to calculate operation and waiting times even with processes with

concurrency simply from the event log, without considering process models.

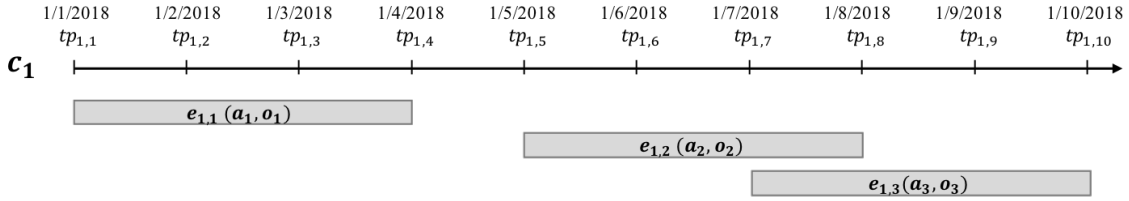


Figure 43. An example of events and time points of case 1 (c_1)

As we discussed earlier, we derive time for cases in the log based on the sequential time points and the status mapping algorithm. Algorithm 4 shows a pseudo code to derive cycle, operation, and waiting time for cases in the log. Based on the pseudo code, *PPIT1* and *PPIT2* are defined.

Definition 14 (PPIT1: Cycle/Operation/Waiting time for cases in the log) *Let CT_c, OT_c, WT_c be cycle time, operation time, and waiting time of cases in the log, respectively.*

- $\{CT_c, OT_c, WT_c\} = DerivingTimeForCases(L, TP, \emptyset)$

PPIT1 covers all cases in the log, thus, we do not put any specific variant in the third input place. If the example in Figure 43 is applied, OT_1 becomes 8 days, obtained by adding 3 days from $tp_{1,1}$(1/1/2018) to $tp_{1,4}$(1/4/2018) and 5 days from $tp_{1,5}$(1/5/2018) to $tp_{1,10}$(1/10/2018), whereas WT_1 is 1 day, i.e., the day between $tp_1, 4$(1/4/2018) and $tp_{1,5}$(1/5/2018). Then, CT_1 becomes 9 day by summing up the operation time and the waiting time.

Definition 15 (PPIT2: Cycle/Operation/Waiting time of a variant) *Let $CT_c(v_1)$, $OT_c(v_1)$, $WT_c(v_1)$ be cycle time, operation time, and waiting time of cases which correspond to a variant (v_1), respectively.*

- $\{CT_c(v_1), OT_c(v_1), WT_c(v_1)\} = DerivingTimeForCases(L, TP, v_1)$

Similar to the previous three metrics, $CT_c(v_1)$, $OT_c(v_1)$, and $WT_c(v_1)$ represent cycle time, operation time, and waiting time for cases involved in a specific variant (v_1). Therefore, we can get the results by putting the third input as v_1 in the *DerivingTimeForCases* function.

Definition 16 (PPIT3: Cycle/Operation/Waiting time of an activity) *Let $CT_c(a_1)$, $OT_c(a_1)$, $WT_c(a_1)$ be cycle time, operation time, and waiting time of cases which correspond to an activity (a_1), respectively.*

- $CT_e(a_1) = OT_e(a_1) + WT_e(a_1)$
- $OT_e(a_1) = \{\pi_t(e_{k,j}) - \pi_t(e_{k,i}) | \forall_{0<k\leq|c|} \forall_{0<i<j\leq n} e_{k,i}, e_{k,j} \in c_k \wedge \pi_{et}(e_{k,j}) = complete \wedge \pi_{et}(e_{k,i}) = start \wedge \pi_a(e_{k,j}) = \pi_a(e_{k,i}) \wedge \nexists_{i<l<j} e_{k,l} \wedge \pi_a(e_{k,j}) = a_1\}$
- $WT_e(a_1) = \{\pi_t(e_{k,j}) - \pi_t(e_{k,i}) | \forall_{0<k\leq|c|} \forall_{0<i<j\leq n} e_{k,i}, e_{k,j} \in c_k \wedge \pi_{et}(e_{k,j}) = start \wedge \pi_{et}(e_{k,i}) = complete \wedge \pi_a(e_{k,j}) = \pi_a(e_{k,i}) \wedge \nexists_{i<l<j} e_{k,l} \wedge \pi_a(e_{k,j}) = a_1\}$

Algorithm 4 *DerivingTimeForCases*(L, TP, v_i)

Input Event Log L;
Timepoints TP;
A variant v_i

Output Records of time for cases in the log $\{CT, OT, WT\}$, *where*
CT is the list of cycle time for cases in the log;
OT is the list of operation time for cases in the log;
WT is the list of waiting time for cases in the log

Let case c_k be an element in the event log L and TP_k be the subset of Timepoints TP. TP_k is composed of p sequential time points of the case c_k, and $tp_{k,i}$ signifies the ith time point of the case c_k.

1: **for each** case $c_k \in L$ **do**
2: **if** $var(c_k) \neq v_i$ **then**
3: **break**
4: **end if**
5: $cycleT, operationT, waitingT = 0$
6: **while** $tp_{k,i+1} \in TP_k$ **do**
7: **if** status$(tp_{k,i})$ = *working* $\wedge$ (status$(tp_{k,i+1})$ = *working* $\vee$ status$(tp_{k,i+1})$ = *waiting*) **then** $operationT \leftarrow operationT + (tp_{k,i+1} - tp_{k,i})$
8: **else if** status$(tp_{k,i})$ = *waiting* $\wedge$ (status$(tp_{k,i+1})$ = *working* $\vee$ status$(tp_{k,i+1})$ = *waiting*) **then** $waitingT \leftarrow waitingT + (tp_{k,i+1} - tp_{k,i})$
9: **end if**
10: $cycleT \leftarrow operationT + waitingT$
11: **end while**
12: $CT \leftarrow CT \cup cycleT$
13: $OT \leftarrow OT \cup operationT$
14: $WT \leftarrow WT \cup waitingT$
15: **return** $\{CT,\ OT,\ WT\}$

Definition 16 defines time-related values of activities. $OT_e(a_1)$ represents the operating time of an activity (a_1), which can be calculated as the difference between start and complete timestamps of an activity in the event log. The waiting time of an activity $WT_e(a_1)$ is calculated as the start timestamp of a_1 and the complete timestamp of the predecessor activity. $CT_e(a_1)$ is the sum of $OT_e(a_1)$ and $WT_e(a_1)$. For example, in the Figure 43, the operation time of activity 2 is 3 days, derived from $to_{1,5}$(1/5/2018) and $to_{1,8}$(1/8/2018), while the waiting time is 1 day, between $to_{1,4}$(1/4/2018) and $to_{1,5}$(1/5/2018). Therefore, the total cycle time of a_2 is 4 days.

Six indicators in Definition 17 and 18, $CT_e(o_1)$, $OT_e(o_1)$, $WT_e(o_1)$, $CT_e(o_1, a_1)$, $OT_e(o_1, a_1)$, $WT_e(o_1, a_1)$, are very similar to previous three indicators in Definition 16. The difference is that the indicators in Definition 17 refer to a specific resource (o_1), while the indicators in Definition 18 refer to both an activity (a_1) and a resource (o_1).

Definition 17 (PPIT4: Cycle/Operation/Idle time of an originator) *Let $CT_c(o_1)$, $OT_c(o_1)$, $WT_c(o_1)$ be cycle time, operation time, and idle time of events which performed by an originator (o_1), respectively.*

- $CT_e(o_1) = OT_e(o_1) + WT_e(o_1)$
- $OT_e(o_1) = \{\pi_t(e_{k,j}) - \pi_t(e_{k,i}) | \forall_{0<k\leq|c|} \forall_{0<i<j\leq n} e_{k,i}, e_{k,j} \in c_k \wedge \pi_{et}(e_{k,j}) = complete \wedge \pi_{et}(e_{k,i}) = start \wedge \pi_a(e_{k,j}) = \pi_a(e_{k,i}) \wedge \nexists_{i<l<j} e_{k,l} \wedge \pi_o(e_{k,j}) = o_1\}$
- $WT_e(o_1) = \{\pi_t(e_{k,j}) - \pi_t(e_{k,i}) | \forall_{0<k\leq|c|} \forall_{0<i<j\leq n} e_{k,i}, e_{k,j} \in c_k \wedge \pi_{et}(e_{k,j}) = start \wedge \pi_{et}(e_{k,i}) = complete \wedge \pi_a(e_{k,j}) = \pi_a(e_{k,i}) \wedge \nexists_{i<l<j} e_{k,l} \wedge \pi_o(e_{k,j}) = o_1\}$

Definition 18 (PPIT5: Cycle/Operation/Idle time with an originator and an activity) *Let $CT_c(o_1, a_1)$, $OT_c(o_1, a_1)$, $WT_c(o_1, a_1)$ be cycle time, operation time, and idle time of events which correspond to an activity (a_1) and performed by an originator (o_1), respectively.*

- $CT_e(o_1, a_1) = OT_e(o_1, a_1) + WT_e(o_1, a_1)$
- $OT_e(o_1, a_1) = \{\pi_t(e_{k,j}) - \pi_t(e_{k,i}) | \forall_{0<k\leq|c|} \forall_{0<i<j\leq n} e_{k,i}, e_{k,j} \in c_k \wedge \pi_{et}(e_{k,j}) = complete \wedge \pi_{et}(e_{k,i}) = start \wedge \pi_a(e_{k,j}) = \pi_a(e_{k,i}) \wedge \nexists_{i<l<j} e_{k,l} \wedge \pi_a(e_{k,j}) = a_1 \wedge \pi_o(e_{k,j}) = o_1\}$
- $WT_e(o_1, a_1) = \{\pi_t(e_{k,j}) - \pi_t(e_{k,i}) | \forall_{0<k\leq|c|} \forall_{0<i<j\leq n} e_{k,i}, e_{k,j} \in c_k \wedge \pi_{et}(e_{k,j}) = start \wedge \pi_{et}(e_{k,i}) = complete \wedge \pi_a(e_{k,j}) = \pi_a(e_{k,i}) \wedge \nexists_{i<l<j} e_{k,l} \wedge \pi_a(e_{k,j}) = a_1 \wedge \pi_o(e_{k,j}) = o_1\}$

6.4.2 Cost

To conduct cost-related analyses, event logs should include cost information as an event attribute (i.e., cost-enhanced event logs). If cost-enhanced logs are available, it is possible to assess the effects of redesigns by defining more direct cost-related PPIs, such as the changes in direct/indirect costs. However, it is often unfeasible to obtain cost-enhanced event logs [21]. Thus, we need to develop a cost-related PPI which can be calculated from information commonly available in event logs. In this chapter, we suggest an alternative indirect cost-related PPI, i.e., the total number of originators in the log (*PPIC1* (F_o)) since labor cost is usually one of the major cost factors.

Table 23 provides a PPI in the cost perspective. *PPIC1* considers the number of originators in a log and can be considered a more accurate proxy of actual costs.

Definition 19 (PPIC1: The total number of originators in the log) *Let F_o be the total number of originators in the log.*

- $F_o = \sum_{q=1}^{m} \begin{cases} 1 & \text{if } O_q \in \{\sum_{0<k<|c|} \sum_{0<i<n} \pi_o(e_{k,i})\} \\ 0 & \text{otherwise} \end{cases}$

Table 23. Process Performance Indicators in Cost Perspective

PPI#	Explanation	Measure	Aggregation Function
PPIC1	The total number of originators in a log	Count of elements	–

This indicator is defined based on the assumption that all resources are full-time equivalents. Assuming that wages are similar among full-time employees, we can evaluate the costs of resources by comparing the number of resources before and after BPR. If the event log contains detailed cost information, then the cost analysis can be more accurate. For example, Tu and Song [171] suggest how to analyze manufacturing costs by applying existing process mining techniques on cost-enhanced event logs.

6.4.3 Quality

A typical approach to evaluating the quality of a process is to check the satisfaction of customers [19]. This external quality is primarily measured through customer surveys, and it is unlikely that this information is available in event logs. For this reason, here, we define PPI metrics which evaluate the extent of standardization on process flows or time-related values. In other words, our analysis focuses on internal process quality, assuming that improved internal quality, e.g., less variable process operating times, is likely to lead to improved customer satisfaction. Four process performance indicators are defined in this perspective (see Table 24).

Table 24. Process Performance Indicators in Quality Perspective

PPI#	Explanation	Measure	Aggregation Function
PPIQ1	Matching rate compared to a reference model	Matching rate	–
PPIQ2	Variation of time for cases in a log	Cycle/Operation/Waiting time	STDEV
PPIQ3	Variation of time of an activity (a_1)	Cycle/Operation/Waiting time	STDEV
PPIQ4	Variation of time for events performed by an originator (o_1)	Cycle/Operation/Waiting time	STDEV

With regard to *PPIQ1*, it can be referred in Definition 4 in Chapter 4.2. It provides a

detailed explanation for calculating the matching rate. Indicators *PPIQ2*, *PPIQ3*, and *PPIQ4* are similar to *PPIT2*, *PPIT3*, and *PPIT4*, but using the standard deviation as aggregation function. These indicators are used to evaluate how diverse are the variations of the time values in the process, per activity, and per resource. Lower standard deviation values entail more stable, streamlined, or standardized processes. As remarked before, more streamlined processes are likely to lead to higher customer satisfaction [172]. Different quality-related indicators may be adopted, such as success rate or failure rate of an activity or a case, cancellation rate, yield rate (for manufacturing processes), or repurchase rate. Information to calculate these indicators, however, is not commonly available in standard event logs that can be handled by process mining tools.

6.4.4 Flexibility

Flexibility evaluates the ability of a process of reacting to changes and handling unexpected situations. To assess flexibility, we introduce three indicators presented in Table 25. Note that an aggregation function is not applicable for these indicators, since they all consider global counters across cases in an event log.

Table 25. Process Performance Indicators in Flexibility Perspective

PPI#	Explanation	Measure	Aggregation Function
PPIF1	The total number of variants in the log	Count of elements	–
PPIF2	The total number of relations in the process model	Count of elements	–
PPIF3	The total number of relations in the social network	Count of elements	–

PPIF1 (F_v), the total number of variants in logs, is defined in Definition 20.

Definition 20 (PPIF1: The total number of variants in a log) *Let F_v be the total number of variants in the log.*

- $F_v = \sum_{r=1}^{o} \begin{cases} 1 & \text{if } V_r \in \{\sum_{0<k<|c|} \pi_{var}(c_k)\} \\ 0 & \text{otherwise} \end{cases}$

In the formula, a variant is a finite set of traces; thus, a high number of variants indicates that logs have diverse case patterns. In other words, a business process with many variants has the ability to handle different types cases. *PPIF2* (F_{ar}) and *PPIF3* (F_{or}) are defined in Definition 21 and 22, respectively.

Definition 21 (PPIF2: The total number of relations in a process model) *Let F_{ar} be the total number of relations in the process model.*

$$\text{- } F_{ar} = \sum_{0<k\leq|c|}\sum_{0<i<j\leq n}\begin{cases}1 & \text{if } c_k \in L \wedge e_{k,i}, e_{k,j} \in c_k \wedge a_l, a_m \in A \wedge e_{k,i} > e_{k,j} \\ & \quad \wedge \pi_a(e_{k,i}) = a_l \wedge \pi_a(e_{k,j}) = a_m \\ 0 & \text{otherwise}\end{cases}$$

Definition 22 (PPIF3: The total number of relations in a social network) *Let F_{or} be the total number of relations in the social network.*

$$\text{- } F_{or} = \sum_{0<k\leq|c|}\sum_{0<i<j\leq n}\begin{cases}1 & \text{if } c_k \in L \wedge e_{k,i}, e_{k,j} \in c_k \wedge a_l, a_m \in A \wedge e_{k,i} > e_{k,j} \\ & \quad \wedge \pi_o(e_{k,i}) = o_l \wedge \pi_o(e_{k,j}) = o_m \\ 0 & \text{otherwise}\end{cases}$$

PPIF2 (F_{ar}) and *PPIF3* (F_{or}) assess the flexibility of a process through measures characterizing process models and social networks discovered from event logs. In particular, they focus on the complexity of the models discovered, intended as number of relations. For example, a higher value of *PPIF2* signifies that the process model is more complex and able to handle a higher variety of cases with different control flow. Similarly, higher values of *PPIF3* signify that more people are cooperating in the execution of a process.

6.5 Evaluation

To validate the proposed approach, we conducted case studies in a hospital organization, where some of the best practices were implemented in their BPR projects. In the case, we have collected real-life event logs from their information systems before and after the projects and computed the indicator values proposed in this chapter. The values are investigated and compared the effects of the best practices in [36]. The case studies show that the proposed method is used to check whether best practices are correctly implemented or not. Furthermore, the indicator values (i.e. PPIs) present the effects of best practices in a quantitative manner.

6.5.1 Context

The case study has been conducted at a tertiary hospital in Korea hosting about 1400 beds and 40 operation rooms. The *extra resources* best practice, that suggests increasing the number of resources if the capacity is not enough, was applied to improve outpatient processes in the clinical neuroscience center and payment processes in the hospital.

BP1: In April 2013, the hospital constructed the new building where the renovated clinical neuroscience center, the cancer center, and the intensive care unit were moved. The hospital increased the number of resources, i.e., clinical doctors, in the centers and the unit. The objective was to increase the ability to provide care to patients by increasing the number of resources and

installing additional clinical equipment. To evaluate the effects of the change of the processes, outpatient logs collected in the clinical neuroscience center were analyzed.

BP2: One of the problems in the hospital was the long delay in the payment process. To avoid a long queue in a receiving teller, the hospital introduced payment devices (KIOSKs). In late 2013, the hospital relocated the KIOSKs on the basis of their usage rate. More KIOSKs were removed from the areas where accessibility by patients was low and installed near the payment counters. By relocating the payment devices, the hospital tried to increase the number of payments using KIOSKs and decrease payments handled in payment counters.

In order to understand the effects of best practice implementation, we extracted EHR (Electronic Health Record) outpatient logs for a month before and after the changes. With regard to *BP1*, we collected one month of data at the clinical neuroscience center in July of 2012 and in July of 2013. For *BP2*, we used event logs about patients' payments for medical expenses through KIOSKs in July and December of 2013. The lag between the BPR implementations and to-be data collection was sufficiently large to avoid the transition period between the as-is and to-be configurations. A summary of the event logs of *BP1* and *BP2* is shown in Table 26.

Table 26. Summary of event logs

	BP1			BP2		
Indicator	Before	After	Variation (%)	Before	After	Variation (%)
Number of cases	1337	2243	67.8	9360	11504	22.9
Number of events	6901	11444	65.8	66582	81084	21.8
Number of activity types	17	17	0.0	17	17	0.0
Number of originators	359	475	32.3	1252	1231	-1.7

6.5.2 Assessing implementation of best practices

In this case, the extra resources best practice was applied and the related measure is the total number of resources in Table 21. Table 26 shows the number of resources before and after the best practice was applied. In the first log for BP1, originators were increased from 359 to 475 (32.3% increase), whereas there was no significant difference in the log for BP2. Considering the resources directly related to BPR, we calculated the discrepancy in the number of clinicians involved in the neuroscience center and the number of KIOSKs located next to the payment counter as given in Table 27. In BP1, the number of doctors who provided clinical services was increased from 25 to 33 (32% increase). The indicator for BP2 was also increased from 4 to 5. Therefore, we concluded that extra resources were implemented well in both cases.

With regard to BP1, the hospital sought to improve the ability to provide care to more patients and to provide more services by employing additional resources. Thus, we investigated

Table 27. The changes of implementation measures

Case	Indicator	Before	After	Variation(%)
BP1	The number of doctors in CNSC	25	33	32.0
BP2	The number of KIOSKs	4	5	25.0

the number of patients and events before and after the BPR. Table 26 shows that the neuroscience center managed 1,337 patients and 6,901 events were logged in July 2012. Meanwhile, after BPR, in July 2013, 2,243 patients visited the center and 11,444 were logged. In BP2, the hospital increased the capacity to handle payment activities by adding more self-payment devices. We analyzed the number of events involving each KIOSK (see Table 28). The utilization of existing KIOSKs generally decreased, but overall the total number of events involving KIOSKs in the event log increased by 24.4%. Also, the usage of KIOSKs was more uniform after BPR, as demonstrated by the standard deviation decreasing from 731.33 to 442.34 (about 40%).

Table 28. The changes of additional implementation measures in BP2

Elements (Frequency)	Before	After	Variation(%)
KIOSK A	3479	2607	-25.1
KIOSK B	2654	540	-4.3
KIOSK C	2327	2145	-7.8
KIOSK D	1437	1494	4.0
KIOSK E (Added)	–	3524	–
Total	9897	12310	24.4
Average	2474.25	3077.5	24.4
Standard deviation	731.33	442.34	-39.5

6.5.3 PPIs application

To quantitatively investigate the effect of the best practice implementations, we calculated PPIs as proposed in Chapter 6.4. Table 29 shows the PPIs for BP1. For the time perspective, all PPIs decreased after BPR. The average case cycle time decreased by 5%. Waiting times of key activities, such as test and consultation, which directly affects satisfaction of patients [158] decreased by about 13%. For the cost perspective, the number of clinicians increased by about 32%, which should have resulted in an increase of the expenses for the hospital. Regarding the quality perspective, we calculated the matching rate between a reference model provided by the hospital and the process model discovered from the event log using the frequency mining

plugin [40]. The matching rate slightly declined after BPR, from 87% to 85%. Also, we analyzed the discrepancy of standard deviations of cycle time for cases in the log and key activities in the process. The standard deviations decreased except for the value of consultation. A lower standard deviation means that the hospital was able to provide the same level of services and it increases the satisfaction of patients, i.e., perceived quality. In the flexibility perspective, we compared the number of variants in the process. The number of process variants increased by 27.5%. However, the discovered process models were very similar and the number of relations among activities in the model remained almost the same before and after BPR (162 to 163). Thus, while the process remained almost the same, the care pathways of outpatients became more diverse and varied. In the social network, the number of relations increased by 38.6%, since the network became more complex as the number of resources involved in the process increased.

Table 29. The changes of PPIs in BP1

PPM	PPI	Before	After	Variation(%)
Time	Average of cycle time for cases in the log (min.)	79.53	79.51	-4.6
	Average of cycle time of consultation (min.)	35.09	33.81	-3.6
	Average of cycle time of test (min.)	11.90	10.60	-10.9
	Average of waiting time of consultation (min.)	27.08	23.72	-12.4
	Average of waiting time of test (min.)	7.71	6.61	-14.3
Cost	The number of doctors in the log	25	33	32.0
Quality	The matching rate compared to the reference model	0.87	0.85	-2.3
	Standard deviation of cycle time for cases in the log (min.)	99.88	84.11	-15.8
	Standard deviation of cycle time of consultation (min.)	27.91	30.16	8.1
	Standard deviation of cycle time of consultation registration (min.)	73.58	65.48	-11.0
	Standard deviation of cycle time of test (min.)	17.42	16.68	-4.2
	Standard deviation of cycle time of test registration (min.)	63.89	45.72	-28.4
Flexibility	The total number of variants in the log	494	630	27.5
	The total number of relations in the process model	162	163	0.6
	The total number of relations in the social network	2840	3936	38.6

Table 30 shows the PPIs for BP2. The average of cycle time for cases and that of the payment activities decreased by about 6%. Regarding the cost perspective, the number of KIOSKs increased, which should have resulted in an increase of the costs for the hospital. For the quality

perspective, the standard deviation for cases in the log decreased slightly from 90.76 to 88.68. However, the standard deviation of the cycle time of the payment slightly increased; thus, we were not able to identify stabilization of payment cycle time according to the growth of KIOSKs. In the flexibility perspective, the number of variants in the log and the number of relations in the social network increased after BPR. However, there was no noticeable difference in the number of relations in the process model, since the new KIOSK did not change the control flow of the process.

Table 30. The changes of PPIs in BP2

PPM	PPI	Before	After	Variation(%)
Time	Average of cycle time for cases in the log (min.)	85.86	80.78	-5.9
	Average of cycle time of variant 1* (min.)	39.5	32	-19.0
	Average of cycle time of variant 2* (min.)	35.4	36.2	2.3
	Average of cycle time of variant 3* (min.)	37.5	35.9	-4.3
	Average of cycle time of variant 4* (min.)	37	36.8	-0.5
	Average of cycle time of variant 5* (min.)	45.1	33.5	-25.7
	Average of waiting time of payment (min.)	9.07	8.42	-7.2
	Average of cycle time of KIOSK A (min.)	10.86	9.47	-12.8
	Average of cycle time of KIOSK B (min.)	5.8	5.16	-11.0
	Average of cycle time of KIOSK C (min.)	10.28	8.46	-17.7
	Average of cycle time of KIOSK D (min.)	8.81	10.25	16.3
	Average of cycle time of KIOSK E (min.)	–	9.18	–
Cost	The number of KIOSKs in the log	4	5	25.0
Quality	Standard deviation of cycle time for cases in the log (min.)	90.76	88.68	-2.3
	Standard deviation of cycle time of payment (min.)	23.44	25	6.7
	Standard deviation of cycle time of KIOSK A (min.)	23.74	26.83	13.0
	Standard deviation of cycle time of KIOSK B (min.)	12.68	8.99	-29.1
	Standard deviation of cycle time of KIOSK C (min.)	27.24	24.19	-11.2
	Standard deviation of cycle time of KIOSK D (min.)	19.47	23.78	22.1
	Standard deviation of cycle time of KIOSK E (min.)	–	21.44	–
Flexibility	The total number of variants in the log	2913	3224	0.7
	The total number of relations in the process model	218	228	4.6
	The total number of relations in the social network	9377	10575	12.8

*Variant 1: Registration → Consultation → Scheduling → Payment → Prescription printing → Treatment

*Variant 2: Registration → Consultation → Scheduling → Payment → Prescription printing

*Variant 3: Registration → Consultation → Payment → Prescription printing

*Variant 4: Registration → Consultation → Payment

*Variant 5: Registration → Consultation → Scheduling → Payment → Treatment

6.5.4 Organizational relevance

The best practice implementation yielded positive effects on the time perspective PPIs in both BP1 and BP2, particularly concerning the average cycle time of the main activities in both cases, i.e., test and consultation in BP1 and payment in BP2. Concerning the cost perspective, both cases showed that adding more resources implied a noticeable increase of costs. Note that the analysis did not cover other costs that were incurred for the implementation of the best practice and for which there was no trace in the event log, e.g., the cost of constructing a new building in BP1 and the costs of relocating the payment devices in BP2. Overall, we concluded that BPR led to negative effects in the cost perspective in both cases. In the quality perspective, PPIs showed both positive and negative effects resulting from the application of the best practice. In BP1, standard deviations of most of the time-related values remained roughly unchanged, except for the matching rate and the time-related values of the consultation activity, which decreased. Similar to BP1, only some of the time-related values in BP2 decreased, and others indicated the opposite effect. Thus, we could not conclude whether the implementation of the best practice had a positive or negative effect on the process. Regarding flexibility, we found that BPR led to an increase of process flexibility in both BP1 and BP2.

To summarize, the application of the increase resource best practice in BP1 and BP2 lead to the following effects on the process: Time – positive, Cost – negative, Quality – neutral, and Flexibility – positive. This evaluation coincides with the suggestions made by Reijers and Mansar [17] for the same best practice.

6.6 Summary and Discussion

This chapter proposed a structured approach to assessing the implementation and benefits of business process redesign best practices based on established process analysis techniques, i.e. process mining. The proposed framework has been validated using case studies in a hospital, focusing on the best practices of *extra resources* (human and physical). The result obtained substantially agree with the conclusions drawn in the literature about the effect of best practices in the time, cost, quality and flexibility perspectives on process performance. The proposed framework, while contributing to the body of literature concerned with the validation of BPR best practices, also represents a ready to use tool for practitioners to conduct advanced BPR process analysis.

Our work has important implications for both research and practice. From an academic research standpoint, the proposed framework provides a sound and verified method to assess the implementation of BPR best practices univocally. As such, it shifts the paradigm of BPR best practice evaluation towards evidence-based decision making. BPR best practices have been assessed in previous work often based on second-hand data, such as process participants and executive interviews [173,174]. Our framework enables the assessment of BPR best practices based

on evidence, i.e., data collected from process executions. Moreover, the proposed framework can be applied by other researchers to improve the knowledge base about BPR best practice effectiveness. This enables building a large-scale knowledge repository based on case studies that have performed BPR assessments. Such a repository may collect information such as service sectors, relevant business processes, goals of redesigns, applied redesign heuristics, utilized BPIs and PPIs, and application results of case studies. This information allows to improve continuously our knowledge about the effectiveness of different process redesign best practices and possibly to define new evidence-based process redesign best practices. A further contribution of this chapter is to link the realms of business process redesign and process mining. While process mining has been used extensively to discover business processes and analyze their conformance to business requirements [37], it has not been used so far for assessing business process redesign in a structured and reusable manner.

As far as implications for practice are concerned, the proposed methodology gives practitioners a ready to use tool to assess process redesign improvements. Process mining is becoming an increasingly mainstream technique for process analysis commonly accepted by practitioners. Forrester, for instance, reports in [175] that 75% of interviewed business decision-makers are aware of process mining and are using it in their daily routine or planning to use it in the next year. Also, while conducting our case studies, we noted an increasing sensibility of executives to understand the evidence provided by process mining tools, which facilitated the communication of our results.

Our work also has several limitations. From a methodological standpoint, the defined indicators need to be validated in the design phase. The suggested indicators were based on the literature review and the experience of the authors. As such, their robustness can be improved by implementing a validation phase involving other experts in the indicators design phase. Also, our framework can be extended by defining additional BPI indicators for other BPR best practice not considered by Reijers and Mansar [17] and by including a mechanism for generating domain specific PPIs. Furthermore, additional PPIs can be developed by employing enhanced logs. In this chapter, for example, we included only one PPI for the cost dimension since it is generally unachievable to obtain cost-enhanced logs. However, if event logs including cost information are available, we can define more direct cost-related PPIs. Therefore, future research should extend our framework to cover more effective and practical indicators.

VII Conclusion

This chapter concludes this dissertation. Chapter 7.1 presents a summary and the implications of this research. Finally, Chapter 7.2 provides directions for future research.

7.1 Summary and Implications

This dissertation aimed to devise a data analysis methodology using process mining for process diagnosis and redesign in healthcare. As such, we built a generic data analysis methodology consisting of the four steps presented in Chapter III: data preparation, preprocessing, analysis, and post-hoc analysis. Of most significance here was the specification for extracting event logs for process mining, starting with the common data model, i.e., a standardized clinical data configuration, and the presenting of two categories with which to effectively conduct process mining analysis: clinical process types and process mining types. We also presented additional steps for interpreting the results obtained from the data analysis with the help of domain experts and performed a post-hoc analysis to improve clinical processes with simulations or to evaluate these processes using previous data analysis results.

We defined three research frameworks that needed to be constructed: 1) a framework for diagnosing clinical processes for outpatients, inpatients, and clinical pathways, 2) a framework for redesigning clinical processes with a simulation-based approach, and 3) a framework for evaluating the effects of process redesign. These are briefly described below.

First, we developed a comprehensive data analysis framework for three clinical process types: outpatients, inpatients, and clinical pathways as described in Chapter IV. For each category, we provided a specific goal and suitable fine-grained techniques that were based on existing approaches or which were developed for the first time in this research. Furthermore, we summarized four real-life case studies to validate the effectiveness of our approach.

Second, we developed a decision support framework for simulation-based redesign analysis, as described in Chapter V. The proposed framework employed process mining and discrete event simulation and was composed of four steps: data preparation, process mining analysis, simulation modeling and evaluation, and experiment and decision support. We suggested a mechanism for obtaining simulation parameters from process mining analysis from control-flow and performance perspectives and automatically constructed a reliable and robust simulation model based on these parameters. The proposed framework was constructed around a specific goal (e.g., a decrease in waiting times) and the applicability of the framework was validated with a case study.

Finally, we developed a framework for evaluating the effects of process redesign (Chapter VI). We defined two types of indicators: best practice implementation indicators to assess whether a specific best practice has been correctly implemented and process performance indicators to understand the impact of applied best practices. These indicators were explicitly connected to

process mining functionalities, and we demonstrated how to obtain these indicators from clinical event logs. The usefulness of the methodology was demonstrated with real-life data before and after a specific process improvement.

This research has implications for both research and practice. Academically, this research links process mining and process redesign in healthcare. While process mining has been used extensively to discover best business processes, it has not yet been used to build and assess process redesign in a structured and reusable manner. Thus, it is believed that this research can act as a motivation for others to extend the use of process mining in healthcare. Also, our framework proposes a new paradigm that encourages an evidence-based approach using data collected from various processes rather than relying on second-hand data for process redesign as seen in previous research.

In terms of practical implications, this research serves as a fully applicable guideline for analyzing and improving clinical processes because the proposed methodology has been developed with the user in mind. The fundamental contributions of this research explained above (e.g., creating event logs with CDM, providing process mining functionalities, and developing systematic frameworks) are evidence of this. Furthermore, it is believed that this study will be a useful tool for clinical process management because its feasibility and usefulness have been demonstrated through various case studies.

This research also contributes to improving the current practice in healthcare organizations. As mentioned, clinical expert systems tend only to serve as tools to help doctors make decisions such as diagnosis and prescriptions in a way that improves patient outcomes, or only focus on describing current situations, measuring performance rather than analyzing the clinical process as a whole. Beyond these limitations, the proposed methodology will produce the following effects. Regarding decision support for care providers, the data analysis framework for clinical pathways will be of significant benefit in analyzing the current situation and generating new forms of clinical pathways (CP) based on data. In particular, given that CP is currently developed manually by doctors, it is believed that this framework will reduce the burden on doctors. This methodology will also allow the diagnosis and understanding of clinical processes from a holistic perspective, something which has not been attempted before. Most administrative teams in hospitals manually investigate improvements in clinical processes, which is not an ideal course of action in practice. However, the proposed methodology and comprehensive frameworks presented in this research enable the diagnosis of problematic issues present in a clinical process, improvements to be made to the process, and an evaluation of the improvements to be conducted with an automatic, evidence-based approach. Therefore, it is believed that this research will enhance the financial situation of healthcare organizations by decreasing the length of stay and increasing the turnover of patients, as well as enhancing customer satisfaction by minimizing waiting times.

Furthermore, a subset of this research, building a redesign model and assessing its effects, can

be utilized in other fields, such as the service or manufacturing industries. In terms of redesign assessment (addressed in Chapter VI), we looked specifically at a service process within a tour agency [20]. We identified the effects of *numerical involvement and split responsibilities best practices* by analyzing logs before and after a redesign. As such, this approach has a flexibility that means it can be employed in other fields.

7.2 Future Research

This dissertation has several limitations which can act as an opportunity for future research. First, the data analysis framework for diagnosis can be extended to emergency department processes. This research included outpatients, inpatients, and clinical pathways as targets, but emergency room care is also considered a clinical process (addressed in the CDM [4]). Emergency room care also requires the analysis and improvement of its clinical processes because it also faces several issues that need to be resolved (e.g., limited resources and the need for rapid treatment). Also, regarding the data-driven methodology for redesign, we demonstrated a connection with the simulation approach, which is utilized to predict the effects of best practices. In addition to the simulation approach, we can extend the optimization approach using data analysis. Realistic optimization parameters can be derived from data analysis using process mining, and then optimized redesign methods can be developed using optimization analysis. In doing so, identifying the optimum combination of factors in a redesign in the face of limited resources is possible, leading to better results. Furthermore, it is necessary to perform more case studies using other clinical logs using the methodology presented in this research. Continuously modifying and improving the proposed framework based on this additional validation can lead to more effective clinical practice.

In addition, process mining in healthcare can be extended using different levels from the BPM lifecycle (e.g., multiple and individual process instances); this research only focused on the process model level. In hospitals, there are numerous healthcare processes including clinical workflow processes for outpatients, inpatients, and emergency rooms as well as administrative processes for human resources or financial management. Building a clinical process repository or process querying framework for manage these would therefore be useful. In addition, BPM at the individual process instance level can be combined with the healthcare environment. For example, it is necessary to develop a method for continuously monitoring patients based on real-time data, i.e., streaming data, and analyzing them in real-time from a performance and conformance perspective to improve the clinical outcome of patients.

Process mining in healthcare also needs to be developed in a way that reflects future hospitals. It is believed that hospitals will change dramatically with the further development of technology. This could include patient-centered hospitals that involve collaborating with patients [176], home healthcare to cope with the increase in chronic disease cases [177], and evidence-based hospitals for better outcomes [178]. Furthermore, ICT development [7], including biometric scanners,

interactive hospital beds, and medical wristwatches for monitoring vital signs, will lead to the changes in clinical processes as well as increasing the volume of collected data. For example, [97] applied process mining from the data collected from wireless tracking to effectively manage and utilize resources by identifying patient's behavior patterns. As such, we need to develop and expand process mining in healthcare in keeping with these changes.

Finally, support that covers the holistic data analysis framework needs to be provided. Process mining tools have been developed by researchers and are continually evolving. However, the focus has generally been on the applications of general business processes and not on approaches that take into account the medical environment. In addition, there is a lack of user-friendly tools that enable both the diagnosis and redesign of clinical processes, including process analysis, management, and improvement in a healthcare environment. Therefore, future research should provide support for medical practitioners.

Bibliography

[1] R. S. Mans, W. M. P. van der Aalst, and R. J. Vanwersch, *Process mining in healthcare: evaluating and exploiting operational healthcare processes.* Springer International Publishing, 2015.

[2] J. Mendling, B. Baesens, A. Bernstein, and M. Fellmann, "Challenges of smart business process management: an introduction to the special issue," *Decision Support Systems*, vol. 100, pp. 1–5, 2017.

[3] W. M. P. van der Aalst, *Process mining: data science in action.* Springer-Verlag Berlin Heidelberg, 2016.

[4] J. M. Overhage, P. B. Ryan, C. G. Reich, A. G. Hartzema, and P. E. Stang, "Validation of a common data model for active safety surveillance research," *Journal of the American Medical Informatics Association*, vol. 19, no. 1, pp. 54–60, 2011.

[5] OECD, *Health at a Glance 2017: OECD Indicators.* OECD Publishing, 2017.

[6] R. Thakur, S. H. Hsu, and G. Fontenot, "Innovation in healthcare: Issues and future trends," *Journal of Business Research*, vol. 65, no. 4, pp. 562–569, 2012.

[7] B. W. Pickering, J. M. Litell, V. Herasevich, and O. Gajic, "Clinical review: the hospital of the future-building intelligent environments to facilitate safe and effective acute care delivery," *Critical Care*, vol. 16, no. 2, p. 220, 2012.

[8] M. Dumas, M. La Rosa, J. Mendling, and H. A. Reijers, *Fundamentals of business process management.* Springer-Verlag Berlin Heidelberg, 2013.

[9] Á. Rebuge and D. R. Ferreira, "Business process analysis in healthcare environments: A methodology based on process mining," *Information systems*, vol. 37, no. 2, pp. 99–116, 2012.

[10] J. A. M. Gray, *Evidence-based healthcare and public health: how to make decisions about health services and public health.* Elsevier Health Sciences, 2009.

[11] M. A. Musen, B. Middleton, and R. A. Greenes, *Clinical Decision-Support Systems.* Springer London, 2014, pp. 643–674.

[12] D. Blum, S. X. Raj, R. Oberholzer, I. I. Riphagen, F. Strasser, and S. Kaasa, "Computer-based clinical decision support systems and patient-reported outcomes: A systematic review," *Patient*, vol. 8, no. 5, pp. 397–409, 2015.

[13] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success," *BMJ*, vol. 330, no. 7494, p. 765, 2005.

[14] M. A. Mach and M. S. Abdel-Badeeh, "Intelligent techniques for business intelligence in healthcare," in *2010 10th International Conference on Intelligent Systems Design and Applications*, 2010, p. 545–550.

[15] W. Bonney, "Applicability of business intelligence in electronic health record," *Procedia - Social and Behavioral Sciences*, vol. 73, pp. 257–262, 2013.

[16] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, "Process mining in healthcare: A literature review," *Journal of biomedical informatics*, vol. 61, pp. 224–236, 2016.

[17] H. A. Reijers and S. L. Mansar, "Best practices in business process redesign: an overview and qualitative evaluation of successful redesign heuristics," *Omega*, vol. 33, no. 4, pp. 283–306, 2005.

[18] J. Mendling, *Metrics for process models: empirical foundations of verification, error prediction, and guidelines for correctness*. Springer Science & Business Media, 2008.

[19] M. Hammer and J. Champy, "Reengineering the corporation: A manifesto for business revolution," *Business Horizons*, vol. 36, no. 5, pp. 90–91, 1993.

[20] M. Cho, M. Song, M. Comuzzi, and S. Yoo, "Evaluating the effect of best practices for business process redesign: An evidence-based approach based on process mining techniques," *Decision Support Systems*, vol. 104, pp. 92–103, 2017.

[21] T. H. Davenport and J. E. Short, *The new industrial engineering: information technology and business process redesign*. Center for Information Systems Research, Sloan School of Management, Massachusetts Institute of Technology, 1990.

[22] T. H. Davenport, *Process innovation: reengineering work through information technology*. Harvard Business Press, 1993.

[23] K. Jensen, *Coloured Petri nets: basic concepts, analysis methods and practical use*. Springer Science & Business Media, 2013.

[24] W. Reisig and G. Rozenberg, *Lectures on petri nets i: basic models*. Springer Science & Business Media, 1998.

[25] A. H. ter Hofstede, W. M. P. van der Aalst, M. Adams, and N. Russell, *Modern Business Process Automation: YAWL and its support environment.* Springer Science & Business Media, 2009.

[26] R. Dijkman, J. Hofstetter, and J. Koehler, *Business Process Model and Notation.* Springer-Verlag Berlin Heidelberg, 2011.

[27] A. W. Scheer, *Business process engineering: reference models for industrial enterprises.* Springer Science & Business Media, 2012.

[28] F. Leymann, D. Karastoyanova, and M. P. Papazoglou, *Business Process Management Standards.* Springer Berlin Heidelberg, 2010, pp. 513–542.

[29] J. B. Hill, J. Sinur, D. Flint, and M. J. Melenovsky, *Gartner's position on business process management.* Gartner, 2006.

[30] R. K. L. Ko, "A computer scientist's introductory guide to business process management (bpm)," *XRDS: Crossroads*, vol. 15, no. 4, pp. 11–18, 2009.

[31] M. Weske, *Business Process Management: Concepts, Languages, Architectures*, 2012.

[32] H. Leopold, *Natural language in business process models.* Springer International Publishing, 2013.

[33] H. van der Aa, H. Leopold, and H. A. Reijers, "Comparing textual descriptions to process models–the automatic detection of inconsistencies," *Information Systems*, vol. 64, pp. 447–460, 2017.

[34] V. Andrikopoulos, S. Benbernou, M. Bitsaki, O. Danylevych, M. Hacid, W. van den Heuvel, D. Karastoyanova, B. Kratz, F. Leymann, M. Mancioppi, and K. Mokhtari, *Survey on business process management.* S-Cube Consortium, 2008.

[35] S. J. van Zelst, B. F. van Dongen, and W. M. P. van der Aalst, "Event stream-based process discovery using abstract representations," *Knowledge and Information Systems*, vol. 54, no. 2, pp. 407–435, 2018.

[36] W. M. P. van der Aalst, T. Weijters, and L. Maruster, "Workflow mining: Discovering process models from event logs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 9, pp. 1128–1142, 2004.

[37] W. M. P. van der Aalst, H. A. Reijers, A. J. M. M. Weijters, B. F. van Dongen, A. K. A. de Medeiros, M. Song, and H. M. W. Verbeek, "Business process mining: An industrial application," *Information Systems*, vol. 32, no. 5, pp. 713–732, 2007.

[38] W. M. P. van der Aalst and E. Damiani, "Processes meet big data: Connecting data science with process science," *IEEE Transactions on Services Computing*, vol. 8, no. 6, pp. 810–819, 2015.

[39] F. Mannhardt, "Multi-perspective process mining," Ph.D. dissertation, Technische Universiteit Eindhoven.

[40] M. Cho, M. Song, and S. Yoo, "A systematic methodology for outpatient process analysis based on process mining," *International Journal of Industrial Engineering: Theory, Applications and Practice*, vol. 22, no. 4, pp. 480–493, 2015.

[41] A. K. A. de Medeiros, A. J. Weijters, and W. M. P. van der Aalst, "Genetic process mining: an experimental evaluation," *Data Mining and Knowledge Discovery*, vol. 14, no. 2, pp. 245–304, 2007.

[42] C. W. Günther and W. M. P. van der Aalst, "Fuzzy mining – adaptive process simplification based on multi-perspective metrics," in *Business Process Management*, G. Alonso, P. Dadam, and M. Rosemann, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 328–343.

[43] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst, "Discovering block-structured process models from event logs - a constructive approach," in *Application and Theory of Petri Nets and Concurrency*, J.-M. Colom and J. Desel, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 311–329.

[44] A. Rozinat, R. S. Mans, M. Song, and W. M. P. van der Aalst, "Discovering simulation models," *Information systems*, vol. 34, no. 3, pp. 305–327, 2009.

[45] A. Weijters, W. M. P. van der Aalst, and A. A. De Medeiros, "Process mining with the heuristics miner-algorithm," *Technische Universiteit Eindhoven, Tech. Rep. WP*, vol. 166, pp. 1–34, 2006.

[46] J. Munoz-Gama, *Conformance Checking and Diagnosis in Process Mining: Comparing Observed and Modeled Processes*. Springer International Publishing, 2016.

[47] A. Weijters and W. M. P. van der Aalst, "Rediscovering workflow models from event-based data using little thumb," *Integrated Computer-Aided Engineering*, vol. 10, no. 2, pp. 151–162, 2003.

[48] W. M. P. van der Aalst, H. T. de Beer, and B. F. van Dongen, "Process mining and verification of properties: An approach based on temporal logic," in *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*, R. Meersman and Z. Tari, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 130–147.

[49] M. Song and W. M. P. van der Aalst, "Towards comprehensive support for organizational mining," *Decision Support Systems*, vol. 46, no. 1, pp. 300–317, 2008.

[50] W. M. P. van der Aalst, H. A. Reijers, and M. Song, "Discovering social networks from event logs," *Computer Supported Cooperative Work (CSCW)*, vol. 14, no. 6, pp. 549–593, 2005.

[51] M. Song and W. M. P. van der Aalst, "Supporting proces mining by showing events at a glance," in *Seventeenth Annual Workshop on Information Technologies and Systems (WITS'07), Montreal, Canada, December 8-9, 2007*, K. Chari and A. Kumar, Eds., 2007, pp. 139–145.

[52] A. Winter, R. Haux, E. Ammenwerth, B. Brigl, N. Hellrung, and F. Jahn, *Health Information Systems*. London: Springer London, 2011, pp. 33–42.

[53] R. Haux, "Health information systems – past, present, future," *International Journal of Medical Informatics*, vol. 75, no. 3, pp. 268–281, 2006.

[54] F. Abbo, "Medical practice management system," Patent US Patent Application No. 10/447,512, 2003.

[55] A. Hoerbst and E. Ammenwerth, "Electronic health records," *Methods of information in medicine*, vol. 49, no. 4, pp. 320–336, 2010.

[56] J. L. Sepulveda and D. S. Young, "The ideal laboratory information system," *Archives of Pathology and Laboratory Medicine*, vol. 137, no. 8, pp. 1129–1140, 2013.

[57] R. Kaushal, K. G. Shojania, and D. W. Bates, "Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review," *Archives of internal medicine*, vol. 163, no. 12, pp. 1409–1416, 2003.

[58] O. Ratib, M. Swiernik, and J. M. McCoy, "From pacs to integrated emr," *Computerized Medical Imaging and Graphics*, vol. 27, no. 2-3, pp. 207–215, 2003.

[59] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek *et al.*, "Observational health data sciences and informatics (ohdsi): opportunities for observational researchers," *Studies in health technology and informatics*, vol. 216, p. 574, 2015.

[60] C. Reich, P. Ryan, R. Belenkaya, K. Natariajan, and C. Blacketer, *OMOP Common Data Model Specifications*, 2018. [Online]. Available: https://github.com/OHDSI/CommonDataModel/wiki

[61] H. Kim, J. Choi, I. Jang, J. Quach, and L. Ohno-Machado, "Feasibility of representing data from published nursing research using the omop common data model," *AMIA Annual Symposium proceedings. AMIA Symposium*, vol. 2016, pp. 715–723, 2016.

[62] Y. Xu, X. Zhou, B. Suehs, A. Hartzema, M. Kahn, Y. Moride, B. Sauer, Q. Liu, K. Moll, M. Pasquale, V. Nair, and A. Bate, "A comparative assessment of observational medical outcomes partnership and mini-sentinel common data models and analytics: Implications for active drug safety surveillance," *Drug Safety*, vol. 38, no. 8, pp. 749–765, 2015.

[63] E. Voss, R. Makadia, A. Matcho, Q. Ma, C. Knoll, M. Schuemie, F. DeFalco, A. Londhe, V. Zhu, and P. Ryan, "Feasibility and utility of applications of the common data model to multiple, disparate observational health databases," *Journal of the American Medical Informatics Association*, vol. 22, no. 3, pp. 553–564, 2015.

[64] J. Marc Overhage, P. Ryan, C. Reich, A. Hartzema, and P. Stang, "Validation of a common data model for active safety surveillance research," *Journal of the American Medical Informatics Association*, vol. 19, no. 1, pp. 54–60, 2012.

[65] S. You, S. Lee, S.-Y. Cho, H. Park, S. Jung, J. Cho, D. Yoon, and R. Park, "Conversion of national health insurance service-national sample cohort (nhis-nsc) database into observational medical outcomes partnership-common data model (omop-cdm)," *Studies in Health Technology and Informatics*, vol. 245, pp. 467–470, 2017.

[66] A. Matcho, P. Ryan, D. Fife, and C. Reich, "Fidelity assessment of a clinical practice research datalink conversion to the omop common data model," *Drug Safety*, vol. 37, no. 11, pp. 945–959, 2014.

[67] P. Rijnbeek, "Converting to a common data model: What is lost in translation?: Commentary on "fidelity assessment of a clinical practice research datalink conversion to the omop common data model"," *Drug Safety*, vol. 37, no. 11, pp. 893–896, 2014.

[68] M. J. Rho, S. R. Kim, S. H. Park, K. S. Jang, B. J. Park, and I. Y. Choi, "Development common data model for adverse drug signal detection based on multi-center emr systems," in *2013 International Conference on Information Science and Applications (ICISA)*, 2013, pp. 1–7.

[69] T. Ong, M. Kahn, B. Kwan, T. Yamashita, E. Brandt, P. Hosokawa, C. Uhrich, and L. Schilling, "Dynamic-etl: A hybrid approach for health data extraction, transformation and loading," *BMC Medical Informatics and Decision Making*, vol. 17, 2017.

[70] S. Liu, Y. Wang, N. Hong, F. Shen, S. Wu, W. Hersh, and H. Liu, "On mapping textual queries to a common data model," in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 2017, pp. 21–25.

[71] W. Yang and Q. Su, "Process mining for clinical pathway: Literature review and future directions," in *2014 11th International Conference on Service Systems and Service Management (ICSSSM)*, 2014, pp. 1–5.

[72] R. Mans, W. M. P. van der Aalst, and R. Vanwersch, "Process mining in healthcare: opportunities beyond the ordinary," *BPM reports*, vol. 1326, 2013.

[73] M. Ghasemi and D. Amyot, "Process mining in healthcare: a systematised literature review," *International Journal of Electronic Healthcare*, vol. 9, no. 1, pp. 60–88, 2016.

[74] A. P. Kurniati, O. Johnson, D. Hogg, and G. Hall, "Process mining in oncology: A literature review," in *2016 6th International Conference on Information Communication and Management (ICICM)*, 2016, pp. 291–297.

[75] E. Rojas, M. Arias, and M. Sepúlveda, "Clinical processes and its data, what can we do with them," in *International Conference on Health Informatics (HEALTHINF)*, 2015, pp. 642–647.

[76] T. Erdoğan and A. Tarhan, "Process mining for healthcare process analytics," in *Software Measurement and the International Conference on Software Process and Product Measurement (IWSM-MENSURA), 2016 Joint Conference of the International Workshop on.* IEEE, 2016, pp. 125–130.

[77] T. G. Erdogan and A. Tarhan, "Systematic mapping of process mining studies in healthcare," *IEEE Access*, vol. 6, pp. 24 543–24 567, 2018.

[78] E. Kim, S. Kim, M. Song, S. Kim, D. Yoo, H. Hwang, and S. Yoo, "Discovery of outpatient care process of a tertiary university hospital using process mining," *Healthcare informatics research*, vol. 19, no. 1, pp. 42–49, 2013.

[79] F. Caron, J. Vanthienen, K. Vanhaecht, E. Van Limbergen, J. De Weerdt, and B. Baesens, "Monitoring care processes in the gynecologic oncology department," *Computers in biology and medicine*, vol. 44, pp. 88–96, 2014.

[80] H. Baek, M. Cho, S. Kim, H. Hwang, M. Song, and S. Yoo, "Analysis of length of hospital stay using electronic health records: A statistical and data mining approach," *PLOS ONE*, vol. 13, no. 4, pp. 1–16, 2018.

[81] M. Rovani, F. M. Maggi, M. de Leoni, and W. M. P. van der Aalst, "Declarative process mining in healthcare," *Expert Systems with Applications*, vol. 42, no. 23, pp. 9236–9251, 2015.

[82] F. Caron, J. Vanthienen, and B. Baesens, "Healthcare analytics: Examining the diagnosis–treatment cycle," *Procedia Technology*, vol. 9, pp. 996–1004, 2013.

[83] F. Caron, J. Vanthienen, K. Vanhaecht, E. Van Limbergen, J. Deweerdt, and B. Baesens, "A process mining-based investigation of adverse events in care processes," *Health Information Management Journal*, vol. 43, no. 1, pp. 16–25, 2014.

[84] X. Xu, T. Jin, Z. Wei, and J. Wang, "Incorporating topic assignment constraint and topic correlation limitation into clinical goal discovering for clinical pathway mining," *Journal of healthcare engineering*, vol. 2017, pp. 261–264, 2017.

[85] Z. Huang, X. Lu, H. Duan, and W. Fan, "Summarizing clinical pathways from event logs," *Journal of Biomedical Informatics*, vol. 46, no. 1, pp. 111–127, 2013.

[86] A. Partington, M. Wynn, S. Suriadi, C. Ouyang, and J. Karnon, "Process mining for clinical processes: a comparative analysis of four australian hospitals," *ACM Transactions on Management Information Systems*, vol. 5, no. 4, pp. 19:1–19:18, 2015.

[87] S. Yoo, M. Cho, E. Kim, S. Kim, Y. Sim, D. Yoo, H. Hwang, and M. Song, "Assessment of hospital processes using a process mining technique: Outpatient process analysis at a tertiary hospital," *International Journal of Medical Informatics*, vol. 88, pp. 34–43, 2016.

[88] P. Delias, M. Doumpos, E. Grigoroudis, P. Manolitzas, and N. Matsatsinis, "Supporting healthcare management decisions via robust clustering of event logs," *Knowledge-Based Systems*, vol. 84, pp. 203–213, 2015.

[89] R. Mans, H. Reijers, D. Wismeijer, and M. van Genuchten, "A process-oriented methodology for evaluating the impact of it: A proposal and an application in healthcare," *Information Systems*, vol. 38, no. 8, pp. 1097–1115, 2013.

[90] M. van Genuchten, R. Mans, H. Reijers, and D. Wismeijer, "Is your upgrade worth it? process mining can tell," *IEEE software*, vol. 31, no. 5, pp. 94–100, 2014.

[91] F. Pegoraro, E. Santos, E. de Freitas Rocha Loures, G. da Silva Dias, L. dos Santos, and R. Coelho, "Short-term simulation in healthcare management with support of the process mining," *Advances in Intelligent Systems and Computing*, vol. 746, pp. 724–735, 2018.

[92] F. Rismanchian and Y. Lee, "Process mining–based method of designing and optimizing the layouts of emergency departments in hospitals," *Health Environments Research and Design Journal*, vol. 10, no. 4, pp. 105–120, 2017.

[93] D. Forsberg, B. Rosipko, and J. L. Sunshine, "Analyzing pacs usage patterns by means of process mining: Steps toward a more detailed workflow analysis in radiology," *Journal of digital imaging*, vol. 29, no. 1, pp. 47–58, 2016.

[94] C. Fernandez-Llatas, A. Lizondo, E. Monton, J.-M. Benedi, and V. Traver, "Process mining methodology for health process tracking using real-time indoor location systems," *Sensors*, vol. 15, no. 12, pp. 29 821–29 840, 2015.

[95] C. Alvarez, E. Rojas, M. Arias, J. Munoz-Gama, M. Sepúlveda, V. Herskovic, and D. Capurro, "Discovering role interaction models in the emergency room using process mining," *Journal of Biomedical Informatics*, vol. 78, pp. 60–77, 2018.

[96] R. J. C. Bose and W. M. P. van der Aalst, "Process diagnostics using trace alignment: opportunities, issues, and challenges," *Information Systems*, vol. 37, no. 2, pp. 117–141, 2012.

[97] C. Fernández-Llatas, J.-M. Benedi, J. M. García-Gómez, and V. Traver, "Process mining for individualized behavior modeling using wireless tracking in nursing homes," *Sensors*, vol. 13, no. 11, pp. 15 434–15 451, 2013.

[98] D. C. Kelleher, R. J. C. Bose, L. J. Waterhouse, E. A. Carter, and R. S. Burd, "Effect of a checklist on advanced trauma life support workflow deviations during trauma resuscitations without pre-arrival notification," *Journal of the American College of Surgeons*, vol. 218, no. 3, pp. 459–466, 2014.

[99] N. Saelim, P. Porouhan, and W. Premchaiswadi, "Improving organizational process of a hospital through petri-net based repair models," in *2016 14th International Conference on ICT and Knowledge Engineering (ICT KE)*, 2016, pp. 109–115.

[100] 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*, 2009, pp. 22–27.

[101] M. L. van Eck, X. Lu, S. J. J. Leemans, and W. M. P. van der Aalst, "Pm2: A process mining project methodology," in *Advanced Information Systems Engineering*, J. Zdravkovic, M. Kirikova, and P. Johannesson, Eds. Springer International Publishing, 2015, pp. 297–313.

[102] J. De Weerdt, F. Caron, J. Vanthienen, and B. Baesens, "Getting a grasp on clinical pathway data: An approach based on process mining," in *Emerging Trends in Knowledge Discovery and Data Mining*, T. Washio and J. Luo, Eds. Springer Berlin Heidelberg, 2013, pp. 22–35.

[103] J. Lismont, A.-S. Janssens, I. Odnoletkova, S. vanden Broucke, F. Caron, and J. Vanthienen, "A guide for the application of analytics on healthcare processes: A dynamic view on patient pathways," *Computers in Biology and Medicine*, vol. 77, pp. 125–134, 2016.

[104] G. T. Lakshmanan, S. Rozsnyai, and F. Wang, "Investigating clinical care pathways correlated with outcomes," in *Business Process Management*, F. Daniel, J. Wang, and B. Weber, Eds. Springer Berlin Heidelberg, 2013, pp. 323–338.

[105] P. Rattanavayakorn and W. Premchaiswadi, "Analysis of the social network miner (working together) of physicians," in *2015 13th International Conference on ICT and Knowledge Engineering (ICT Knowledge Engineering 2015)*, 2015, pp. 121–124.

[106] E. Rojas, M. Sepúlveda, J. Munoz-Gama, D. Capurro, V. Traver, and C. Fernandez-Llatas, "Question-driven methodology for analyzing emergency room processes using process mining," *Applied Sciences*, vol. 7, no. 3, 2017.

[107] P. Ahmad, S. Qamar, and S. Q. A. Rizvi, "Techniques of data mining in healthcare: a review," *International Journal of Computer Applications*, vol. 120, no. 15, pp. 38–50, 2015.

[108] M. U. Khan, J. P. Choi, H. Shin, and M. Kim, "Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008, pp. 5148–5151.

[109] C. Chien and G. J. Pottie, "A universal hybrid decision tree classifier design for human activity classification," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 1065–1068.

[110] S. S. Moon, S. Y. Kang, W. Jitpitaklert, and S. B. Kim, "Decision tree models for characterizing smoking patterns of older adults," *Expert Systems with Applications*, vol. 39, no. 1, pp. 445–451, 2012.

[111] C.-L. Chang and C.-H. Chen, "Applying decision tree and neural network to increase quality of dermatologic diagnosis," *Expert Systems with Applications*, vol. 36, no. 2, pp. 4035–4041, 2009.

[112] Y. Xie, G. Schreier, D. C. Chang, S. Neubauer, Y. Liu, S. J. Redmond, and N. H. Lovell, "Predicting days in hospital using health insurance claims," *IEEE journal of biomedical and health informatics*, vol. 19, no. 4, pp. 1224–1233, 2015.

[113] H. Baek, M. Cho, S. Kim, H. Hwang, M. Song, and S. Yoo, "Analysis of length of hospital stay using electronic health records: A statistical and data mining approach," *PLoS one*, vol. 13, no. 4, p. e0195901, 2018.

[114] Divya and S. Agarwal, "Weighted support vector regression approach for remote healthcare monitoring," in *2011 International Conference on Recent Trends in Information Technology (ICRTIT)*, 2011, pp. 969–974.

[115] S. Belciug, "Patients length of stay grouping using the hierarchical clustering algorithm," *Annals of the University of Craiova-Mathematics and Computer Science Series*, vol. 36, no. 2, pp. 79–84, 2009.

[116] T.-S. Chen, T.-H. Tsai, Y.-T. Chen, C.-C. Lin, R.-C. Chen, S.-Y. Li, and H.-Y. Chen, "A combined k-means and hierarchical clustering method for improving the clustering efficiency of microarray," in *2005 International Symposium on Intelligent Signal Processing and Communication Systems*, 2005, pp. 405–408.

[117] D. Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G. Wang, "Algorithmic prediction of health-care costs," *Operations Research*, vol. 56, no. 6, pp. 1382–1392, 2008.

[118] T. H. A. Soliman, A. A. Sewissy, and H. AbdelLatif, "A gene selection approach for classifying diseases based on microarray datasets," in *2010 2nd International Conference on Computer Technology and Development*, 2010, pp. 626–631.

[119] R. K. Jha, B. S. Sahay, and P. Charan, "Healthcare operations management: a structured literature review," *DECISION*, vol. 43, no. 3, pp. 259–279, 2016.

[120] B. T. Denton, *Handbook of healthcare operations management.* Springer-Verlag New York, 2013.

[121] D. Gupta and W.-Y. Wang, *Patient Appointments in Ambulatory Care.* Springer US, 2012, pp. 65–104.

[122] L. V. Green and S. Savin, "Reducing delays for medical appointments: A queueing approach," *Operations Research*, vol. 56, no. 6, pp. 1526–1538, 2008.

[123] S. Deo and I. Gurvich, "Centralized vs. decentralized ambulance diversion: A network perspective," *Management Science*, vol. 57, no. 7, pp. 1300–1319, 2011.

[124] N. Yankovic and L. V. Green, "Identifying good nursing levels: A queuing approach," *Operations research*, vol. 59, no. 4, pp. 942–955, 2011.

[125] D. Gupta and B. Denton, "Appointment scheduling in health care: Challenges and opportunities," *IIE transactions*, vol. 40, no. 9, pp. 800–819, 2008.

[126] B. Pearce, N. Hosseini, K. Taaffe, N. Huynh, and S. Harris, "Modeling interruptions and patient flow in a preoperative hospital environment," in *Proceedings of the 2010 Winter Simulation Conference*, 2010, pp. 2261–2270.

[127] B. Liang, A. Turkcan, M. E. Ceyhan, and K. Stuart, "Improvement of chemotherapy patient flow and scheduling in an outpatient oncology clinic," *International Journal of Production Research*, vol. 53, no. 24, pp. 7177–7190, 2015.

[128] L. Wang, "An agent-based simulation for workflow in emergency department," in *2009 Systems and Information Engineering Design Symposium*, 2009, pp. 19–23.

[129] M. Laskowski and S. Mukhi, "Agent-based simulation of emergency departments with patient diversion," in *Electronic Healthcare*, D. Weerasinghe, Ed. Springer Berlin Heidelberg, 2009, pp. 25–37.

[130] A. K. Kanagarajah, P. Lindsay, A. Miller, and D. Parker, "An exploration into the uses of agent-based modeling to improve quality of healthcare," in *Unifying Themes in Complex Systems*, A. Minai, D. Braha, and Y. Bar-Yam, Eds. Springer Berlin Heidelberg, 2008, pp. 471–478.

[131] S. A. Erdogan and B. Denton, "Dynamic appointment scheduling of a stochastic server with uncertain demand," *INFORMS Journal on Computing*, vol. 25, no. 1, pp. 116–132, 2013.

[132] P. P. Wang, "Static and dynamic scheduling of customer arrivals to a single-server system," *Naval Research Logistics (NRL)*, vol. 40, no. 3, pp. 345–360, 1993.

[133] B. Denton and D. Gupta, "A sequential bounding approach for optimal appointment scheduling," *IIE transactions*, vol. 35, no. 11, pp. 1003–1016, 2003.

[134] M. A. Begen, R. Levi, and M. Queyranne, "A sampling-based approach to appointment scheduling," *Operations Research*, vol. 60, no. 3, pp. 675–681, 2012.

[135] M. A. Begen and M. Queyranne, "Appointment scheduling with discrete random durations," *Mathematics of Operations Research*, vol. 36, no. 2, pp. 240–257, 2011.

[136] R. Velásquez and M. T. Melo, "A set packing approach for scheduling elective surgical procedures," in *Operations Research Proceedings 2005*, H.-D. Haasis, H. Kopfer, and J. Schönberger, Eds. Springer Berlin Heidelberg, 2006, pp. 425–430.

[137] A. Jebali, A. B. H. Alouane, and P. Ladet, "Operating rooms scheduling," *International Journal of Production Economics*, vol. 99, no. 1-2, pp. 52–62, 2006.

[138] B. T. Denton, A. J. Miller, H. J. Balasubramanian, and T. R. Huschka, "Optimal allocation of surgery blocks to operating rooms under uncertainty," *Operations research*, vol. 58, no. 4-part-1, pp. 802–816, 2010.

[139] S. Batun, B. T. Denton, T. R. Huschka, and A. J. Schaefer, "Operating room pooling and parallel surgery processing under uncertainty," *INFORMS journal on Computing*, vol. 23, no. 2, pp. 220–237, 2011.

[140] D. Yoon, E. K. Ahn, M. Y. Park, S. Y. Cho, P. Ryan, M. J. Schuemie, D. Shin, H. Park, and R. W. Park, "Conversion and data quality assessment of electronic health record data at a korean tertiary teaching hospital to a common data model for distributed network research," *Healthcare informatics research*, vol. 22, no. 1, pp. 54–58, 2016.

[141] R. P. J. C. Bose, R. S. Mans, and W. M. P. van der Aalst, "Wanna improve process mining results?" in *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2013, pp. 127–134.

[142] S. Suriadi, R. Andrews, A. H. ter Hofstede, and M. T. Wynn, "Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs," *Information Systems*, vol. 64, pp. 132–150, 2017.

[143] H. Bueno, J. S. Ross, Y. Wang, J. Chen, M. T. Vidan, S.-L. T. Normand, J. P. Curtis, E. E. Drye, J. H. Lichtman, P. S. Keenan *et al.*, "Trends in length of stay and short-term outcomes among medicare patients hospitalized for heart failure, 1993-2006," *JAMA*, vol. 303, no. 21, pp. 2141–2147, 2010.

[144] T. Rotter, L. Kinsman, E. L. James, A. Machotta, H. Gothe, J. Willis, P. Snow, and J. Kugler, "Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs," *Cochrane Database of Systematic Reviews*, 2010.

[145] Z. Huang, X. Lu, and H. Duan, "On mining clinical pathway patterns from medical behaviors," *Artificial Intelligence in Medicine*, vol. 56, no. 1, pp. 35–50, 2012.

[146] T. Romeyke and H. Stummer, "Clinical pathways as instruments for risk and cost management in hospitals-a discussion paper," *Global journal of health science*, vol. 4, no. 2, pp. 50–59, 2012.

[147] R. Mans, H. Schonenberg, G. Leonardi, S. Panzarasa, A. Cavallini, S. Quaglini, and W. M. P. van der Aalst, "Process mining techniques: an application to stroke care," *Studies in health technology and informatics*, vol. 136, pp. 573–578, 2008.

[148] D. Parmenter, *Key performance indicators: developing, implementing, and using winning KPIs*. John Wiley & Sons, 2015.

[149] B. Li, H. Yang, Y. Wei, R. Su, C. Wang, W. Meng, Y. Wang, L. Shang, Z. Cai, L. Ji *et al.*, "Is it time to change our reference curve for femur length? using the z-score to select the best chart in a chinese population," *PLoS ONE*, vol. 11, no. 7, p. e0159733, 2016.

[150] L. Deng, Y. Hu, J. P. Y. Cheung, and K. D. K. Luk, "A data-driven decision support system for scoliosis prognosis," *IEEE Access*, vol. 5, pp. 7874–7884, 2017.

[151] W. van der Aalst, A. Adriansyah, A. K. A. de Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. van den Brand, R. Brandtjen, J. Buijs, A. Burattin, J. Carmona, M. Castellanos, J. Claes, J. Cook, N. Costantini, F. Curbera, E. Damiani, M. de Leoni, P. Delias, B. F. van Dongen, M. Dumas, S. Dustdar, D. Fahland, D. R. Ferreira, W. Gaaloul, F. van Geffen, S. Goel, C. Günther, A. Guzzo, P. Harmon, A. ter Hofstede, J. Hoogland, J. E.

Ingvaldsen, K. Kato, R. Kuhn, A. Kumar, M. La Rosa, F. Maggi, D. Malerba, R. S. Mans, A. Manuel, M. McCreesh, P. Mello, J. Mendling, M. Montali, H. R. Motahari-Nezhad, M. zur Muehlen, J. Munoz-Gama, L. Pontieri, J. Ribeiro, A. Rozinat, H. Seguel Pérez, R. Seguel Pérez, M. Sepúlveda, J. Sinur, P. Soffer, M. Song, A. Sperduti, G. Stilo, C. Stoel, K. Swenson, M. Talamo, W. Tan, C. Turner, J. Vanthienen, G. Varvaressos, E. Verbeek, M. Verdonk, R. Vigo, J. Wang, B. Weber, M. Weidlich, T. Weijters, L. Wen, M. Westergaard, and M. Wynn, "Process mining manifesto," in *Business Process Management Workshops*, F. Daniel, K. Barkaoui, and S. Dustdar, Eds. Springer Berlin Heidelberg, 2012, pp. 169–194.

[152] B. F. Van Dongen, A. K. A. de Medeiros, H. Verbeek, A. Weijters, and W. M. P. van der Aalst, "The prom framework: A new era in process mining tool support," in *International Conference on Application and Theory of Petri Nets*. Springer, 2005, pp. 444–454.

[153] T. J. Carney, G. P. Morgan, J. Jones, A. M. McDaniel, M. Weaver, B. Weiner, and D. A. Haggstrom, "Using computational modeling to assess the impact of clinical decision support on cancer screening improvement strategies within the community health centers," *Journal of Biomedical Informatics*, vol. 51, pp. 200–209, 2014.

[154] S. Jacobson, S. Hall, and J. Swisher, "Discrete-Event Simulation of Health care Systems," in *Patient Flow SE - 12*, ser. International Series in Operations Research & Management Science, R. Hall, Ed. Springer US, 2013, vol. 206, pp. 273–309.

[155] P. K. Sahoo, S. K. Mohapatra, and S.-L. Wu, "Analyzing healthcare big data with prediction for future health condition," *IEEE Access*, vol. 4, pp. 9786–9799, 2016.

[156] M. H. Rutberg, S. Wenczel, J. Devaney, E. J. Goldlust, and T. E. Day, "Incorporating discrete event simulation into quality improvement efforts in health care systems," *American Journal of Medical Quality*, vol. 30, no. 1, pp. 31–35, 2013.

[157] N. R. Hoot, L. J. LeBlanc, I. Jones, S. R. Levin, C. Zhou, C. S. Gadd, and D. Aronsky, "Forecasting Emergency Department Crowding: A Discrete Event Simulation," *Annals of Emergency Medicine*, vol. 52, no. 2, pp. 116–125, 2008.

[158] A. Berhane and F. Enquselassie, "Patients' preferences for attributes related to health care services at hospitals in amhara region, northern ethiopia: a discrete choice experiment," *Patient preference and adherence*, vol. 9, pp. 1293–1301, 2015.

[159] W. Cao, Y. Wan, H. Tu, F. Shang, D. Liu, Z. Tan, C. Sun, Q. Ye, and Y. Xu, "A web-based appointment system to reduce waiting for outpatients: A retrospective study," *BMC Health Services Research*, vol. 11, no. 1, p. 318, 2011.

[160] C. Nessim, J. Winocour, D. P. M. Holloway, R. Saskin, and C. M. B. Holloway, "Wait times for breast cancer surgery: effect of magnetic resonance imaging and preoperative investigations on the diagnostic pathway." *Journal of oncology practice / American Society of Clinical Oncology*, vol. 11, no. 2, pp. e131–e138, 2015.

[161] T. Cayirli and E. Veral, "Outpatient Scheduling in Health Care: a Review of Literature," *Production and Operations Management*, vol. 12, no. 4, pp. 519–549, 2003.

[162] Process Analyzer, "Process Analyzer Website," http://pa.postech.ac.kr.

[163] D. Muller, "AutoMod®: modeling complex manufacturing, distribution, and logisitics systems for over 30 years," in *Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World*, 2013, pp. 4037–4051.

[164] B. Wetzstein, Z. Ma, and F. Leymann, "Towards measuring key performance indicators of semantic business processes," in *Business Information Systems*, W. Abramowicz and D. Fensel, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 227–238.

[165] A. Del-RíO-Ortega, M. Resinas, C. Cabanillas, and A. Ruiz-Cortés, "On the definition and design-time analysis of process performance indicators," *Information Systems*, vol. 38, no. 4, pp. 470–490, 2013.

[166] R. S. Kaplan and D. P. Norton, *The balanced scorecard: translating strategy into action.* Harvard Business Press, 1996.

[167] V. Popova and A. Sharpanskykh, "Modeling organizational performance indicators," *Information Systems*, vol. 35, no. 4, pp. 505–527, 2010.

[168] M. Al-Mashari, Z. Irani, and M. Zairi, "Business process reengineering: a survey of international experience," *Business Process Management Journal*, vol. 7, no. 5, pp. 437–455, 2001.

[169] M. H. Jansen-Vullers, P. Kleingeld, M. Loosschilder, M. Netjes, and H. A. Reijers, "Tradeoffs in the performance of workflows–quantifying the impact of best practices," in *International Conference on Business Process Management.* Springer-Verlag Berlin Heidelberg, 2007, pp. 108–119.

[170] W. M. P. van der Aalst, H. T. de Beer, and B. F. van Dongen, "Process mining and verification of properties: An approach based on temporal logic," in *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*, R. Meersman and Z. Tari, Eds. Springer Berlin Heidelberg, 2005, pp. 130–147.

[171] T. B. H. Tu and M. Song, "Analysis and prediction cost of manufacturing process based on process mining," in *2016 International Conference on Industrial Engineering, Management Science and Application (ICIMSA)*, 2016, pp. 1–5.

[172] P. Lillrank, "The quality of standard, routine and nonroutine processes," *Organization Studies*, vol. 24, no. 2, pp. 215–233, 2003.

[173] M. Al-Mashari, Z. Irani, and M. Zairi, "Business process reengineering: a survey of international experience," *Business Process Management Journal*, vol. 7, no. 5, pp. 437–455, 2001.

[174] S. M. Siha and G. H. Saad, "Business process improvement: empirical assessment and extensions," *Business Process Management Journal*, vol. 14, no. 6, pp. 778–802, 2008.

[175] C. Richardson, C. Mines, R. Heffner, N. Fenwick, J. Rymer, C. L. Clair, and C. Tajima, "The new discipline of digital business automation," *Forrester*, 2016.

[176] C. V. Fiorio, M. Gorli, and S. Verzillo, "Evaluating organizational change in health care: the patient-centered hospital model," *BMC health services research*, vol. 18, no. 1, p. 95, 2018.

[177] S. Shen, "Hospital to home: Design to prevent social loneliness among people with chronic heart failure," 2015.

[178] D. R. Bardach, *Evidence-based hospitals*. University of Kentucky, 2015.

Acknowledgements

본 박사학위 논문을 작성하기까지 많은 분들의 응원과 격려, 도움이 있었습니다. 지금의 제가 있기까지 도움을 주신 많은 분들께 이 지면을 빌어 감사의 말을 올리고자 합니다.

우선, 울산과 포항에서의 대학원 기간 동안 한결같이 애정 어린 충고와 격려, 가르침을 주신 존경하는 송민석 교수님께 감사의 말씀 드립니다. 항상 부족했던 저를 교수님께서는 관심과 애정으로 지도해 주셨습니다. 저에게는 버팀목이었던 교수님 덕분에, 오늘의 제가 있는 것 같습니다. 교수님께서 몸소 보여주신 삶에 대한 태도와 마음가짐을 항상 가슴에 새기고 살아가도록 하겠습니다. 교수님의 첫 박사 제자로서 부족한 점이 많고, 아직도 채워야 할 부분이 많습니다. 앞으로도 많은 가르침 부탁 드립니다.

부족한 저의 지도를 맡아 주시고, 많은 도움을 주신 Marco Comuzzi 교수님께 감사의 뜻을 전합니다. (I would like deeply to express my gratitude to my technical advisor, Professor Marco Comuzzi. Thank you for your guidance and support.) 논문 및 연구 수행과 관련하여 많은 조언을 해주신 분당서울대병원 유수영 교수님께도 감사 드립니다. 또한, 논문 심사를 흔쾌히 수락해 주시고, 부족한 논문에 대해 날카로운 지적과 조언을 주신 권대일 교수님, 임치현 교수님께도 감사의 뜻을 전합니다. 이 외에, 저에게 많은 가르침을 주신 유니스트 경영학과, 경영공학과 여러 교수님들께도 감사의 말씀을 드립니다.

울산과 포항에서 함께 동고동락했던 AIM/BPI 연구실 선후배 및 동기 여러분들에게도 감사의 말을 전합니다. 먼저, 포항공대 AIM 연구실 식구인 호정, 도현, 민규, 정은, 규남, 덕상, 현아, 정우, 종원, 성희, 규동이형에게 고맙다는 말을 하고 싶습니다. 까칠한 연구실 선배를 만나 고생했음에도 따뜻한 마음으로 늘 응원해 주어 고맙습니다. 여러분의 앞날에 충만한 성취가 있기를 기원합니다. 또한, 유니스트에서 함께 생활했던 Bernardo Nugroho Yahya 박사님, 홍성철 박사님, 한나누나, 영준이형, 용혁, 숙영, 광운, 예림, 민우, 태현, 민주에게도 고마움을 표합니다. 여러분과 함께한 추억은 가슴 속에 소중히 간직하도록 하겠습니다. 서로의 박사학위 과정 생활을 응원하며 함께 고생해 온 수진, 성환, 범철, 학부 때부터 학업 외적으로 정신적 지주가 되어준 장배, 정석에게도 감사함을 표하며, 나의 유일한 대학원 동기 재선이와 멀리서 응원해 준 경성이에게도 고마움을 전합니다. 그리고 저의 학위 과정 동안 항상 힘이 되어 주었던 민정이에게 고맙다는 말을 전합니다.

마지막으로, 묵묵히 뒤에서 걱정하시고 응원해주신 부모님께 감사의 말씀을 올립니다. 부모님의 은혜 덕분에 학교 생활과 학위 과정을 잘 마무리 할 수 있었습니다. 항상 감사한 마음을 가지고 보답하면서 살겠습니다. 그리고, 누나와 매형, 곧 태어날 조카에게도 감사의 말을 전합니다.

본 논문이 저 한 명의 성과가 아닌 많은 분들의 도움으로 완성되었다는 것을 잊지 않겠습니다. 앞으로 더 노력하고 정진하여 많은 분들의 기대에 부응하는 사람이 되겠습니다. 이 지면에 다 표현하지는 못했지만, 그동안 저를 응원해주신 모든 분들께 다시 한 번 감사의 말씀을 전합니다.

2018년 7월, 조민수 올림

Curriculum Vitae

Minsu Cho

Ulsan National Institute of Science and Technology (UNIST)
Department of Management Engineering
E-mail: mcho@unist.ac.kr / minsu.cho.29@gmail.com
WWW: http://mscho90.wordpress.com

Education

2013.3 – 2018.8	School of Management Engineering Ulsan National Institute of Science and Technology (Doctor of Philosophy)
2009.3 – 2013.2	School of Technology Management Ulsan National Institute of Science and Technology (Bachelor of Science)

Academic Experiences

2016.3 – 2018.8	Research Assistant Pohang University of Science and Technoogy
2016.9 – 2017.3	Visiting Researcher Eindhoven University of Technology
2013.3 – 2016.2	Teaching & Research Assistant Ulsan National Institute of Science and Technology

Publications: Journals (International)

[1] H. Baek†, **M. Cho**†, S. Kim, H. Hwang, M. Song*, and S. Yoo*, "Analysis of length of hospital stay using electronic health records: A statistical and data mining approach." *PLoS ONE*, 13(4): e0195901, 2018 (†: equal contribution).

[2] **M. Cho**, M. Song*, M. Comuzzi, and S. Yoo, "Evaluating the effect of best practices for business process redesign: An evidence-based approach based on process mining techniques" *Decision Support Systems*, 104:92-103, 2017.

[3] S. Yoo, **M. Cho**, E. Kim, S. Kim, Y. Sim, D. Yoo, H. Hwang, and M. Song*, "Assessment of Hospital Processes Using a Process Mining Technique: Outpatient Process Analysis at a Tertiary Hospital." *International Journal of Medical Informatics*, 88:34-43, 2016.

[4] **M. Cho**, M. Song*, and S. Yoo, "A Systematic Methodology for Outpatient Process Analysis based on Process Mining." *International Journal of Industrial Engineering: Theory, Applications, and Practice*, 22(4):480-493, 2015.

[5] S. Yoo, **M. Cho**, S. Kim, E. Kim, S. Park, K. Kim, H. Hwang, and M. Song*, "Conformance Analysis of Clinical Pathway Using Electronic Health Record Data." *Healthcare Informatics Research*, 21(3):161-166, 2015.

Publications: Journals (Domestic)

[1] **M. Cho**, D. Kim, M. Song*, G. Kim, C. Jung, and K. Kim, "A Development on a Predictive Model for Buying Unemployment Insurance Program Based on Public Data." *The Korean Journal of BigData*, 2(2):17-31, 2017.

Publications: Conference Proceedings (International)

[1] **M. Cho**, M. Song, C. Müller, P. Fernandez, A. del-Río-Ortega, M. Resinas, and A. Ruiz-Cortés, "A New Framework for Defining Realistic SLAs: An Evidence-Based Approach" *In Business Process Management Forum, BPM 2017, Vol. 297 of Lecture Notes in Business Information Processing*, Springer International Publishing, Barcelona, Spain, September 10-15, 2017.

[2] C. Müller, **M. Cho**, P. Fernandez, M. Song, M. Resinas, and A. Ruiz-Cortés, "Devising an SLA-Aware Methodology to Improve Process Performance" *In Jornadas de Ciencia e Ingeniería de Servicios*, Spain, July 19-21, 2017.

[3] H. Yang, M. Park, **M. Cho**, M. Song, and S. Kim, "A System Architecture for Manufacturing Process Analysis based on Big Data and Process Mining Techniques." *In IEEE International Conference on Big Data*, pp.1024-1029, Washington D.C, USA, October 27-30, 2014.

[4] **M. Cho**, M. Song, and S. Yoo, "A Systematic Methodology for Outpatient Process Analysis Based on Process Mining." *In Asia Pacific Business Process Management, Vol. 181 of Lecture Notes in Business Information Processing*, pp.31-42, Springer International Publishing, Brisbane, Australia, July 3-4, 2014.

[5] **M. Cho**, S. Pyo, K. Lee, and M. Jung, "Forecasting Spot Price of Crude Oil using Three-Agent Model", *the 13rd Asia Pacific Industrial Engineering and Management Systems Conference*, Patong Beach, Phuket, Thailand, December 2-5, 2012.

[6] **M. Cho**, S. Pyo, and M. Jung, "Risk Analysis of Uncertainty in Pricing of Crude Oil", *Proceedings of 1st International Youth Conference*, pp.41-49, Vladivostok, Russia, May 29-30, 2012.

Publications: Conference Proceedings (Domestic)

[1] J. Lim, M. Song, **M. Cho**, S. Yoo, K. Kim, H. Baek, and S. Kim, "A methodology for deriving CP variation from patient types using clinical performance analysis" *In Industrial Engineering and Management Science*, Gyeongju Hotel Hyundai, Gyeongju, Korea, April 4-7, 2018.

[2] G. Park, **M. Cho**, M. Song, and J. Lee, "A methodology for discovering a yield-based optimal resource path in semiconductor manufacturing" *In Industrial Engineering and Management Science*, Gyeongju Hotel Hyundai, Gyeongju, Korea, April 4-7, 2018.

[3] M. Lee, J. Kwahk, S. H. Han, M. Song, **M. Cho**, Y. Koh, D. Kim, S. Oh, H. Kim, G. Chae, and J. Lee, "Defining and Analyzing IoT Intelligence" *In 2017 Fall Conference of the Ergonomics Society of Korea*, Wellihillipark, Gangwon, Korea, November 29 - December 2, 2017.

[4] **M. Cho**, M. Song, D. Kim, K. Kim, C. Jung, and K. Kim, "A Predictive Model for Buying Unemployment Insurance Program" *In Korea Bigdata Society*, Yonsei University, Seoul, Korea, May 12, 2017.

[5] J. Lim, **M. Cho**, M. Song, D. Kim, M. Choi, S. Yoo, K. Kim, H. Baek, and S. Kim, "Assessment of clinical pathways based on patients' records: A case study" *In Industrial Engineering and Management Science*, Yeosu Expo Convention Center, Yeosu, Korea, April 26-29, 2017.

[6] **M. Cho**, M. Song, and S. Yoo, "Personal clinician scheduling using Discrete Event Simulation based on Process Mining." *In KORMS/KIIE/ESK/KSIE/KSS Joint Spring Conference*, Jeju ICC, Jeju, Korea, April 13-16, 2016.

[7] H. Yi, M. Song, **M. Cho**, and D. Kim, "Impact of Business Process Re-engineering on Reorganization of Process." *In KORMS/KIIE/ESK/KSIE/KSS Joint Spring Conference*, Jeju ICC, Jeju, Korea, April 13-16, 2016.

[8] **M. Cho**, M. Song, and S. Yoo, "A Method for Developing Clinical Pathway using Event Logs." *In KORMS/KIIE/ESK/KSIE/KSS Joint Spring Conference*, Ramada Hotel, Jeju, Korea, April 8-11, 2015.

[9] **M. Cho**, M. Song, and S. Yoo, "Developing Clinical Pathway using Process Mining.", *In the 2014 Korea Bigdata Society Fall Conference*, Kintex, Goyang, Korea, September 18, 2014.

[10] Y. Shim, **M. Cho**, M. Song, and S. Yoo, "Delta analysis for organizational environment using process mining: A case study." *In Industrial Engineering and Management Science*, pp.425-430, Bexco, Busan, Korea, May 16-17, 2014.

[11] E. Kim, S. Kim, **M. Cho**, M. Song, and S. Yoo, "Simulation of Optimal Number of Outpatient Payment KIOSK Estimation Using Process Mining", *Proceedings of the 2013 Korean Society of Medical Informatics Spring Conference*, Asan Hospital, South Korea, June 13-14, 2013.

[12] **M. Cho**, M. Song, and S. Yoo. "Payment Process Analysis Based on Process Mining", *the 2013 Korea Intelligent Information System Society Spring Conference*, Sogang University, Korea, June 1, 2013.

[13] **M. Cho**, M. Song, and S. Yoo, "Healthcare Process Simulation Based on Process Mining", *the 2013 Korean Institute of Industrial Engineers/ Korean Operations Research and Management Science Society Fall Co-conference*, The Ocean Resort, Yeosu, Korea, May 24-25, 2013.

[14] S. Pyo, **M. Cho**, K. Lee, and M. Jung, "Interrelationship Analysis between Crude Oil Price and the World Economic Indices", *the 2012 Korean Institute of Industrial Engineers*, Ansan, Korea, November 2, 2012.

[15] K. Lee, **M. Cho**, S. Pyo, and M. Jung, "Relationship Analysis between Oil Price and Equity Returns of Domestic Industry using a Regressing Model", *Proceedings of The 2012 IE/MS Joint Spring Conference*, E1.4, pp.1-6, Gyeongju, Korea, May 10-11, 2012.

[16] K. Lee, **M. Cho**, and M. Jung, "Risk analysis of oil pricing in crude oil trading", *Proceedings of The 2011 KIIE Fall Conference*, Seoul, Korea, pp.1-5, November 5, 2011.

Patents

[1] M. Song and **M. Cho**, "Process Simulation Model Discovery System of Outpatient Consultation and Process Simulation Model Discovery Method", registered in Korea. (Patent No. *10-1601916* (2016))

Awards

[1] The *Best Student Paper* at the 13rd Asia Pacific Industrial Engineering and Management Systems Conference.

[2] The *Best Paper* at 2013 Korean Society of Medical Informatics Spring Conference.

[3] The *Achievement Award* at Ulsan National Institute of Science and Technology.

[4] The *Third-prize Award in the process mining case competition* at the 2015 Asia-Pacific Conference on Business Process Management.