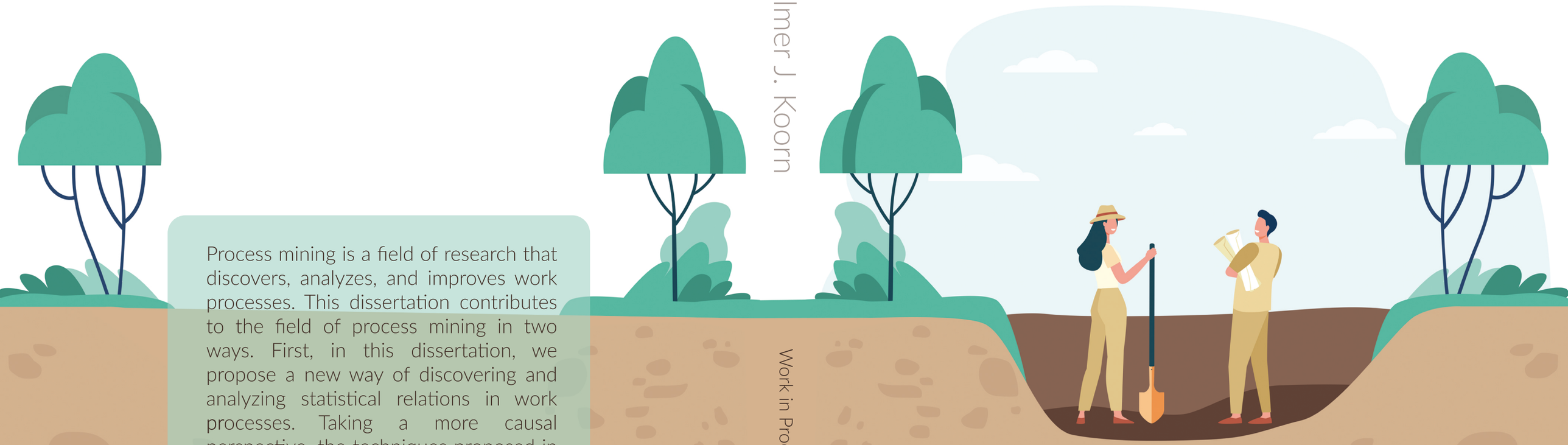
Jelmer J. Koorn

Process mining is a field of research that discovers, analyzes, and improves work processes. This dissertation contributes to the field of process mining in two ways. First, in this dissertation, we propose a new way of discovering and analyzing statistical relations in work processes. Taking a more causal perspective, the techniques proposed in this dissertation help understand the impact of causes and effects in work processes. Second, this dissertation adds value by providing support for qualitative evaluations in process mining projects. In this dissertation, we propose a set of guidelines that help improve the validity of a qualitative evaluation when domain experts are involved in the process.

Jelmer J. Koorn

Work in Process: Unearthing Meaning using Process Mining



Work in Process:
Unearthing Meaning using Process Mining

Work in Process: Unearthing Meaning using Process Mining

Jelmer J. Koorn

Assessment committee:
prof. dr. ir. Remco Dijkman, Technical University Eindhoven
dr. Jorge Munoz-Gama, Pontificia Universidad Catolica de Chile
prof. dr. Arno P. J. M. Siebes, Utrecht University
prof. dr. Marco Spruit, Leiden University Medical Center
prof. dr. Barbara Weber, University of St. Gallen

SIKS Dissertation Series No. 2022-15
The research reported in this thesis has been carried out under auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Work in Process: Unearthing Meaning using Process Mining
ISBN: 978-90-393-7471-9

Druk: ProefschriftMaken
Cover image: Freepik.com, pch.vector

Work in Process
Unearthing Meaning using Process Mining

Betekenis Ontdekken met Process Mining

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de rector magnificus,
prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 30 mei 2022 des ochtends te 10.15 uur

door

Jelmer Jan Koorn

geboren op 5 november 1994,
te Groningen, Nederland

Promotor:	Prof. dr. ir. H.A. Reijers
Copromotor:	Prof. dr. H. Leopold
	dr. ir. X. Lu

Dit proefschrift werd (mede) mogelijk gemaakt door financiële steun van NWO in het TACTICS project [project code 628.011.004].

The Post-Season Interview

Usually, in a dissertation, this is where you would find the Acknowledgments section. It is a place where the author, I, disclose one of the untold stories about how this book came into existence. I draw the attention of the reader, you, and put the spotlight on those who have supported me along the way. However, the story of gratitude is but one of the untold stories of a journey we call a Ph.D. trajectory. It is difficult for me to only talk about this journey while ignoring other stories. Writing only about thankfulness sketches a utopia in what, to me, is a place of reflection on my journey. Sometimes, it makes me feel like a sports player who scores a last-minute goal and excessively celebrates it in a game where the score was already long decided.

Thinking about it, a Ph.D. trajectory is almost like a sports season. Some matches are won, some matches are lost, tournaments are played in which prizes can be won, and at the end of the season, there is a winner. In my trajectory, some papers got published, some got rejected, and I was fortunate enough to lift a junior-league trophy from behind my screen for one of my publications at a conference.

What I am writing here is the post-season interview. Realizing that my emotions are still running high, I might look back on this season differently later on. The season is not even over when I am sitting down to write this. The last match, my defense, is still coming up. So what should I say, or more accurately write, in such an interview? Media training will tell me to keep it superficial, focus on the achievements, and look ahead. Fortunately, or unfortunately, I had a different kind of training. Trained by the principles of scientific communication. Transparency is an important norm I want to adhere to. It is such a story that I want to tell in this interview.

What do I see when I look back at this season? I first see the team I am the captain of and playing with. My co-authors and supervisory team, who helped me think about the tactics and contributed to winning matches. Those who thought and wrote with me to develop and ultimately publish work, which you will find a lot of in this book. I also see the club, my colleagues at the university, they are the players that did not join for the matches but were there for the training sessions. These sessions helped my team prepare for the matches, improving us before we stepped onto the pitch. I see the sponsors who supported us with proper training facilities. NWO and Lunet Zorg provided us with equipment and circumstances to perform. And I see the fans, those who were on the sidelines, my friends and family, who cheered us on, who celebrated with us when we won, and who believed in us even when we might have lost that faith.

Then I see a neutral party, the referees in the game, the reviewers. They decide what passes during a match and what is brought to a stop. Their decisions can be helpful and their arguments useful. Sometimes, we feel they make the wrong decisions, disagree with their arguments, or feel they take the wrong perspective. Either way, they are part of the game and we need to learn how to play with them.

I also see matches we could have played better. Matches that we should have lost, or were lucky to win. This too is part of the season we have played. It is in these matches that I need to look at myself and see how I could have played better. I look at the team to see if I should have changed tactics, players, or positions. Next, I look at the club, sponsors, and fans to see their role in this. In the end, each and everyone one of them has contributed to the team its success. That does not mean there are no lessons to be learned, things to improve upon to be better next season.

In this competition, only the winners of the season get the chance to write up their experience in a post-season interview. Consider the players and teams who did not end up winning this season and hope to get back in the next. I feel they deserve to have post-season interviews in which they can share their lessons as well. There are also players who will be starting their first season shortly. They may want to use this post-season interview to plan ahead and optimize their performance and tactics. We, of course, cannot give up our competitive advantage by sharing our specific strengths and weaknesses with them. Besides, the specific lessons we learned in our season are not relevant to these other players and teams. These lessons only helped my team against our opponents. Nonetheless, reflecting on our performance is helpful. Success and failure are both part of the game. Looking more closely at the season, what I see is that single matches may be won or lost by individual actions but the season result is always a reflection of the team and the larger environment in which it plays.

In the end, we won this season. The dissertation is in your hands. The season is concluded in a final match in which I, as the captain of the team, have to defend the performances of the team over the last season. If that is successful, we get ready for a beautiful ceremony to pick up a wonderful trophy. So perhaps, this is not really a post-season interview, but rather a post-season analysis report. One in which I analyze the season to truly understand what made it successful. On the one hand, looking at the things that went well, which I should be grateful for and cherish to make sure it is around for next season. On the other hand, looking at the things that could be done differently, things that should change next time around.

That time has come for me, the time for a post-season analysis. Sitting down, thinking, writing, and speaking about how to make next season a success. Let me end this story with a quote from a sportsman that captures what I learned during my season as a Ph.D. candidate. A quote that abstracts and helps me make next season a success. May that next season be in sports, academia, or elsewhere:

Ability is what you are capable of doing.
Motivation determines what you do.
Attitude determines how well you do it.
Lou Holtz

Jelmer Jan Koorn,
April 12th 2022

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

"*Skepticism is unbelief in cause and effect*"
Ralph Waldo Emerson (1860), The Conduct of Life: "Worship", *p. 192*

These words were written in the context of worship by Ralph Waldo Emerson (1860), an American essayist. Worship not only in terms of religious beliefs but also worship as a means to merge faith with science and arts. The concepts of cause and effect are seen as vital to connect these diverse facets of mankind. The concepts of cause and effect provide a perspective through which we can study how the world progresses. We can use the concepts of cause and effect to direct efforts in a way that serves us best. In other words, cause and effect help us understand what we can do to achieve the goals we have.

Naturally, not only individuals aim to achieve goals. Collectively, people also aim to achieve goals (McKenna, 2020). Organizations are social arrangements through which people can collectively work towards such goals (Buchanan & Huczynski, 2019). Organizational goals can be: producing a product, selling a service, or providing aid to those in need. To achieve organizational goals, actors in organizations perform activities. As individuals collaborate to work towards an organizational goal, there is a need for the coordination of the activities that are performed (Okhuysen & Bechky, 2009; Rollinson, 2008). To coordinate the activities we can think of a collection of activities as a *work process*[1]. We define a work process as a set of related activities that is performed by various actors within an organization that result in some form of output.

To make the concept of work process more tangible, let us consider an example. Consider the healthcare domain, specifically a healthcare organization that provides direct care for clients (e.g. a hospital or a care facility for people with a disability). The main goal of such an organization is to ensure the well-being of its clients to the best possible degree. Within this context, care staff treats clients with the goal to cure them or care for them. Several activities are performed in such a setting. When a patient enters the hospital, doctors try to diagnose the patient. Specific experts occasionally perform extra tests. Then, the doctors decide on a treatment for the patient. A nurse provides the treatment and further care for the client. Finally, the doctors and nurses monitor the patient to see if the treatment leads to the desired result.

We know that organizations vary in many ways; their domain, the product or service they offer, whether they are for-profit or non-profit, and so on. Regardless of these differences, all organizations have in common that they want to achieve goals. To do so, the management of an organization wants to understand and optimize how it can achieve its organizational goals. In light of the healthcare example above, it is interesting for both the care staff and the hospital management to be able to understand the impact of each activity on the status of the patient.

Insights into the impact of activities are not only relevant in a healthcare context. For example, think of a financial organization that assesses loan applications. Its management may ask themselves: How can we assess a loan application in the most accurate way possible? Alternatively, consider a customer services department of an organization. The management may wonder: How can we deal with complaints

[1]The term *work process* is explained in more detail in 1.1

such that customers are most satisfied? A railway operator that temporarily closed a train station has to consider: How can we best organize alternative transport to compensate for the closed train station?

What all these organizational examples have in common is that there is value for organizations in understanding how causes and effects are related in their work process. With respect to these organizational work process examples:

- *Causes* relate to activities. Causes from the previous examples can be a request for extra information for the loan application and the choice of specific treatment by the medical staff.
- *Effects* relate to successive activities and/or the goal of the work process. Effects from the previous examples can be the status of a patient (e.g. healthy, sick, or deceased) and the travel time of a customer.

Studying cause and effect proves to be challenging (Pearl & Mackenzie, 2018). The concepts of cause and effect are often complex, as the previous organizational examples show. Capturing cause and effect in data is necessary to allow us to study the concepts (Pearl, 2009). The concepts have been studied from as early as the ancient Greeks (Barnes, 1991). The onset of new data, or a new perspective on existing data, allows us to better capture the complexity surrounding cause and effect (Pearl & Mackenzie, 2018). New data and perspectives on existing data provide an opportunity to better study the relationship between cause and effect.

In an organizational setting, the generation of new data and reuse of existing data are increasingly recognized as potential sources of organizational value (Constantiou & Kallinikos, 2015; Mikalef et al., 2018). More and more, organizations collect data and try to recycle data initially gathered for other purposes (Hashem et al., 2015; Sagiroglu & Sinanc, 2013; Villars et al., 2011). Consider data that an organization generates during the performance of activities within the organization. IT systems routinely and automatically log this data, and managers may actively collect new data by tracking the progress of activities to assess employee performance (Paulk et al., 1993; Van der Aalst, 2016). Business process management is a field that uses this specific data to help organizations gain insights into their work processes (Dumas et al., 2013). Process mining is a family of techniques within the business process management field that supports organizations in the steps of discovering, analyzing, and improving work processes (De Weerdt et al., 2013; Rojas et al., 2016). These fields mostly consider the order in which activities are performed, but do not focus on how cause and effect may play a role in a work process [2], in Section 1.3 we develop this notion further.

Ultimately, the management in an organization wants to understand how each activity contributes to the organizational goal. A causal perspective helps organizations discover the impact of their activities by framing it in the concepts of cause and effect. The data on work processes, also referred to as event data, presents an opportunity to study cause and effect in an organizational context. Using work processes as a lens we can study the impact of specific activities on the organizational goal. The work process lens helps structure and understand *what* and *when* activities are performed to reach an organizational goal. Once we create such an understanding, the causal

[2] Where the idea of using this specific data from IT systems is new, the notion of looking at processes to discover causality was already put forward by Aristotle (Barnes, 1991) in his definition of the *four causes*.

perspective helps us study the *how*. This is meant in the sense of: how does each activity have an impact on the organizational goal?

In what follows, we first elaborate on business process management and process mining by providing *background* literature on both topics. Then, we elaborate on the *main challenges* that this dissertation addresses. The next section translates these challenges and focuses on the *contributions* of this research, after which the *outline* of the dissertation structure is explained. This chapter concludes with a *list of publications*.

1.1 Background

The goal of *business process management* is to help organizations understand, analyze, and ultimately take advantage of improvement opportunities in work processes (Dumas et al., 2013). The combination of these two goals, understanding and improving work processes, fits the desire of an organization that seeks to understand the impact of its activities. Fundamental for this discussion is a reflection on the concept of a work process, Section 1.1.1 will elaborate on this. *Process mining* can be used to discover and analyze these work processes. Process mining is an evidence-based approach that uses existing data from information systems in an organization. The basis for process mining techniques and their analyses are data extracted from IT systems. This data needs to be extracted in a certain form; this form is referred to as an *event log*. These event logs are extracted from various information systems that are used within organizations. Therefore, these event logs provide valuable insights into how work processes are executed in real-life (Van der Aalst, 2016).

Two of the most prominent steps in process mining are: *discovery* and *evaluation*. One of the first steps in process mining is called process *discovery*. In this step, the event log is used to learn, and often visualize, how a work process is executed in real-life. The goal here is to learn about the work process and create an understanding of how the work process is executed in real-life. In the *evaluation* step of a process mining project, we turn our attention to the validation of the results. People that discover and analyze the work process are often not the same people that perform the activities in a work process. As such, sharing and evaluating the insights generated about the work process is an important step in a process mining project.

At present, traditional process mining discovery techniques on the one hand, and limited support for evaluations on the other hand form obstacles for organizations who aim to understand the impact of their activities. Below, we discuss the related work for these two challenges. The first challenge describes how process mining *discovery* techniques focus mainly on the order of activities. To elaborate on that, the following section discusses the literature in which we specifically look at the three most common techniques in process mining discovery. The second challenge relates to the lack of methodological support for evaluations with domain experts in process mining evaluations. Therefore, in a follow-up section, we reflect on the current process mining *evaluations*.

1.1.1 Work Process

Before delving into the literature regarding process mining, let us return to the notion of a work process. This term is a more inclusive version of the more familiar term business process (e.g. (Dumas et al., 2013)). In this dissertation, we use the term work process for two reasons: purpose and formality. *Purpose* refers to the fundamental mission of an organization. In terms of purpose, the term business often has a commercial connotation (Richards & Schmidt, 2013). *Formality* describes the type of organization of the group of people working towards a shared goal. In terms of formality, the term business often refers to a group of people that is formally established, for example through registration as a legal entity. Reflecting on the concepts of purpose and familiarity, we can see how the term 'business' scopes us to only consider a narrow part of the concepts. In this dissertation, we adopt a more inclusive scope of activities and processes. Let us consider a scenario to exemplify this:

A group of neighbors from an apartment building spontaneously come together and decide they want to organize monthly barbeques for the whole neighborhood. To do this, they need to make sure several tasks are performed, e.g. rent a barbeque, buy drinks and food, get a party tent, etc. The goal of these barbeques is to build a stronger sense of community in the neighborhood.

In this scenario, it would be difficult to argue that these neighbors are collaborating in a *business* process. The monthly barbeques are not aimed at generating a profit and the group of people is not formally registered. However, we can still discover, analyze, and improve how these neighbors work. To do this justice, in this dissertation, we use the term *work* process. A work process is a set of related activities that is performed by a group of actors with a shared purpose that result in some form of output. We refer to an organization as a, more straightforward, synonym for the group of actors working towards a shared purpose (Richards & Schmidt, 2013).

1.1.2 Process Discovery

Over time, many process discovery techniques have been proposed (Augusto et al., 2018). On an aggregate level, the vast majority of the discovery techniques can be categorized into three types: procedural, declarative, and hybrid. The first group of techniques produces *procedural* models. These techniques are the most commonly used in process mining discovery. These techniques analyze the sequence of activities performed in a work process. Examples include Petri nets (Song et al., 2015; Verbeek et al., 2017), causal nets (Nguyen et al., 2017; Yahya et al., 2016), BPMN models (Augusto et al., 2017; Broucke & De Weerdt, 2017) or process trees (Buijs et al., 2012; Leemans et al., 2013). The second type of technique is are techniques that produce *declarative* models. These techniques focus on constraints that determine the order of activities in a work process. Examples of work that use such a technique are (Bernardi et al., 2014; Schönig et al., 2016). The final group of techniques produces *hybrid* models. These techniques combine elements from both procedural and declarative models. Examples of works that use such a technique are (De Smedt et al., 2015; Maggi et al., 2014). What all these techniques have in common is that they aim to

discover the control flow of a work process, that is, the execution constraints among the process' activities. A shortcoming of the control-flow approach adopted by most discovery techniques is that certain insights cannot be obtained. Most importantly, we cannot use these techniques to assess the impact of activities on the process. The importance of this research gap is best explained using an example.

The Research Gap Illustrated

Let us consider the healthcare domain as an example. Specifically, a mental health setting with a residential care facility. Care staff support clients with mental and/or physical disabilities around the clock in their daily lives. In such a context, the main goal of the care organization is to optimize the well-being of its clients. As such, the organization is interested in knowing the impact of the activities performed by both care staff and clients that relate to the well-being of the clients.

In terms of understanding activities with an impact, consider the aggression incidents that take place in the residential care facilities. Let us introduce client Taylor and caretaker Jordan. An aggressive incident would involve client Taylor 'lashing out' at caretaker Jordan. Taylor can shout, hit, or throw objects to Jordan. Jordan has to deal with the situation and make sure Taylor calms down. After the incident, caretaker Jordan has to file a report describing the incident. Using traditional process mining discovery techniques, we use these reports to show what activities caretaker Jordan has historically taken to respond to the aggression of client Taylor. This already provides insights into the interplay between the client and caretaker. However, what traditional process mining discovery techniques cannot tell us is how the response of caretaker Jordan impacts the aggression of client Taylor. In other words, we cannot tell what Jordan could do to reduce the chances of Taylor becoming aggressive in the future.

These aggressive incidents have a large negative impact on the well-being of both clients and care staff as they cause a large amount of stress and possible bodily harm. To improve the well-being of client Taylor, the organization needs to be able to discover the impact of one activity on another. In this dissertation, we address this shortcoming and propose a way to discover work processes from a novel perspective.

1.1.3 Process Mining Evaluations

At the end of every process mining project there comes a moment where the results and insights of the project are *evaluated*. The evaluation at the end of a process mining project can be done in both a quantitative and qualitative manner. Here, a quantitative evaluation refers to those evaluations in which numerical data is generated to assess the findings (Creswell, 2013). Think here of for example benchmarking the performance of a proposed algorithm. These types of evaluations are often performed using a computer to generate the data. A qualitative evaluation refers to those evaluations in which other types of data are generated (e.g. textual data) (Creswell & Poth, 2016). Think here of, for example, interviews or focus groups. In contrast to quantitative evaluations, these qualitative evaluations often involve humans as participants to generate the data.

As process mining projects are often performed in real-world scenarios there are usually external experts involved, for example from an organization in which the

data is collected. These experts are referred to as domain experts. Process mining methodologies recognize this and generally advocate for the involvement of domain experts (Emamjome et al., 2019). Domain experts are essential to achieve the goal of producing organizational value and help explain the why and how questions behind the findings of the project. Although the importance of the involvement of domain experts is acknowledged, what all process mining methodologies have in common is limited support for *how* to perform evaluations when domain experts are involved. In this dissertation, we address this challenge and propose a set of guidelines to perform exactly these types of evaluations.

1.2 Key Challenges

In this dissertation, two key challenges are addressed that relate to the process discovery (main challenge 1) and the evaluation (main challenge 2) step in process mining projects.

- **Main challenge 1**: Identify, analyze, and visualize statistical relations in work processes.

Generally, process mining projects focus on the *discovery* of a process in terms of a sequential ordering of activities. Sequential order here refers to the central question: how are the activities in a process ordered from start to end? This approach does not always provide the insights required to understand, analyze, and improve a work process. Recall the healthcare example where we analyze an aggressive incident with client Taylor and caretaker Jordan from earlier. In this context, taking a sequential order approach would provide the healthcare organization with insights such as *after an aggressive incident of a client, the caretaker responded by taking measure X*. These insights cannot support an organization in gaining insights into how the activities impact each other and the process as a whole. To exemplify, the healthcare organization may wonder how a caretaker can best respond to an incident. To this end, a causal point of view in process mining discovery is required to provide such insights. Here, statistical, and in particular dependency relations can be identified. As such, this challenge describes how additional digging is required to unearth the meaning behind the order of activities.

- **Main challenge 2**: Provide actionable support to improve validity of qualitative evaluations with domain experts in process mining projects.

At the end of a process mining project, an evaluation step ensures that findings from the project are assessed. The evaluation is often performed in collaboration with domain experts. The involvement of domain experts is widely advocated and recognized in the process mining methodologies. Currently, there is limited support in terms of guidelines, frameworks, or methodologies for both researchers and practitioners to perform qualitative evaluations. As a result, informal ways of working have evolved due to the lack of central guidelines. To improve the validity of the outcomes

of process mining evaluations, the challenge is to bring transparency and coherence in the approaches to qualitative evaluations in which domain experts are involved.

We have now introduced an overview of the two main challenges this dissertation aims to tackle. In the next section, we describe how this dissertation contributes to addressing these challenges.

1.3 Research contributions

In this dissertation, we present contributions to two types of literature. First, we describe contributions to the body of literature covering process mining *discovery* techniques. Second, we present the contributions in terms of the *evaluations* performed in process mining projects. Below, we outline each of these types of literature and the specific contributions of this dissertation.

We contribute the following techniques to the process mining **discovery** literature:

- **Contribution 1**: A process mining technique that uses well-established statistical mechanisms to detect, analyze, and visualize statistical relations within work processes.
- **Contribution 2**: A novel technique to detect, analyze, and visualize statistical relations between actions and states in a work process.
- **Contribution 3**: An extension to existing process mining discovery techniques to help identify confounding variables in processes.

We contribute the following to the process mining **evaluation** literature:

- **Contribution 4**: An overview of current practices in process mining case study evaluations in which domain experts are involved.
- **Contribution 5**: A set of guidelines to support qualitative evaluations in process mining projects in which domain experts are involved.

The main focus of this dissertation is encompassed in these five contributions. In Section 1.4 the outline of this dissertation is discussed connected to the contributions that are made in each part of the dissertation. Next to these five main contributions, other insights are generated. Most notably, the proposed techniques in process discovery and analysis are applied in two healthcare case studies. The application and evaluation of the techniques also result in various healthcare domain-specific insights. Two insights can be defined in this respect: (1) the use of existing process mining techniques to study behavior as a work process, and (2) the generation of insights into domain-specific healthcare questions, specifically in the field of aggressive behavior and sepsis. These insights will be reflected upon in Chapter 8 where the *practical relevance* of this dissertation is discussed.

1.4 Dissertation Outline

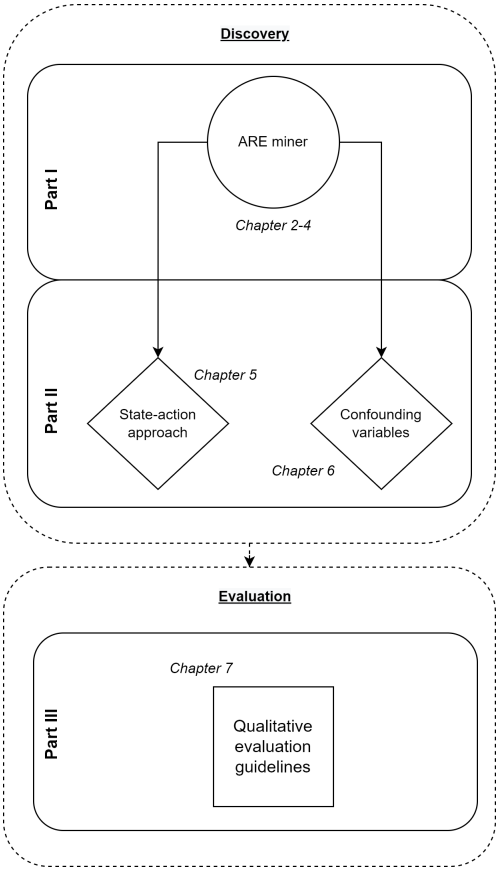


Figure 1.1: Outline of the dissertation

The dissertation is divided into three parts, which are visualized in Figure 1.1. The first part describes how the ARE miner came into existence. The second part describes two separate extensions of the ARE miner. The third part takes a methodological stance and looks at qualitative evaluations in process mining projects. Below, we provide a summary of the works of each part and show how they contribute to this dissertation.

1.4.1 Part I

Chapter 2 describes the application of process mining techniques to discover patterns of aggressive behavior in clients with intellectual disabilities (Koorn et al., 2019b). The novelty in this work lies in the use of process mining to discover behavioral patterns. Previous work mostly focuses on what is referred to as *static* analysis. The term static refers to the focus on single aggressive incidents. The introduction of process

mining techniques results in a more *dynamic* analysis. The term dynamic is used as we study the development of incidents in clients over time. Thus, this enables a study of the behavior, rather incidents, of clients. As such, we can see that the behavior of clients can be seen as a process: it is the development of incidents over time. This chapter introduces a case study that is revisited in several following chapters. Hence, information regarding the raw data generation and its conversion to event logs can be found here. This work opens a new way of using process mining to study behavior but is limited in the sense that it only considers the incidents. When the discovered patterns are presented and discussed, inevitably the questions of cause and effect arise. In other words, how can we extend this process into a work process in which multiple activities are captured? For example, were there any measures taken by a caretaker that influence the client? Specifically, we expect that there are dependency relations in these activities that can help explain the patterns of aggressive behavior.

Chapter 3 addresses these questions (Koorn et al., 2020). It formally introduces the Action-Response-Effect (ARE) miner to capture not only the incident but also study its causes and effects. To study the causes and effects within work processes the concept of process activity is too generic. Therefore, the more fine-grained concepts of action, response, and effect are formalized in this paper. Next to the formalization, the contribution of this work comes from the algorithm introduced. The algorithm takes specific ARE event logs as input, it then uses process mining and statistics to uncover potential dependency patterns, and returns these potential dependency patterns in graphical form as output. The novelty of this work is the combination of statistics and process mining with a specific application to study behavior. This combination of statistics and process mining helps produce graphical representations of potential dependency patterns in actions, responses, and effects. In addition, the combination helps to effectively filter relations in the process such that the potential dependency patterns are easy to understand. The ARE miner is evaluated in a case study of aggressive behavior in clients with intellectual disabilities. However, it is important to understand the behavior of the miner better outside of the case study and to compare it with existing process mining techniques.

Chapter 4 extends the work of the ARE miner (Koorn et al., 2022a). The ARE miner is extended in a variety of ways, most prominently the technique is quantitatively evaluated on an artificial data set. This evaluation provides insights into the behavior of the miner in a large variety of data settings. In addition, the quantitative evaluation is used to benchmark the performance of the ARE miner against that of a naive and filtered directly follows graph (DFG) technique. The evaluation shows that the ARE miner generally is better at providing insights into dependency patterns compared to both DFG techniques. Next to that, the case study is revisited and new insights are generated. The algorithm is improved by automating the determination of a key parameter value that was previously required to be manually inputted. In addition, the graphical representations it produces are enhanced by including the strength of the dependency relation. All these extensions result in a well-rounded process mining technique that can detect, analyze, and visualize dependency patterns.

1.4.2 Part II

The second part includes two extensions of the ARE miner. The *first extension* in Chapter 5 is a work that proposes a more generalized version of the ARE miner (Koorn et al., 2022c). This work proposes a novel technique to study the interaction and statistical relations between actions and states. As such, we move away from the more strict approach adopted in the ARE miner where we define a specific set of activities (i.e. an action is followed by a response is followed by an effect) is required. This work adopts a more relaxed approach to the work process. Relaxed as we do not strictly define states and actions. As a result, we can study less tangible input than a work activity, such as the status of a patient as a state or a decision as an action. In addition, the relaxed definitions leave more freedom in the order of states and activities. This allows us to choose either a state or an action as starting point of the work process. We apply the technique in a sepsis work process in a hospital setting. The technique describes the creation of a state-action log that is subsequently analyzed using a set of statistical tests after which a graphical representation is created.

The *second extension* in Chapter 6 is based on a work that proposes a way to detect, display, and deal with confounding variables in potential dependency patterns (Koorn et al., 2022b). Confounding variables are variables, or activities in process terms, which provide an alternative explanation to the one captured in a potential dependency pattern. This extension is an important addition to the ARE miner as it detects potential dependency relations. To test if a potential dependency relation is real, an important check is to explore alternative explanations, e.g. confounding variables. Thus, this extension is a step moving from potential dependency relations to causal relations. Previous work in process mining, to a certain extent, recognizes the potential effect of confounding variables. The novelty of this extension is twofold. First, the combination of detecting, visualizing, and dealing with confounding variables is new. Second, the technique is broadly applicable to other process mining discovery techniques. The technique consists of a three-step approach where we first enrich event logs, then we use statistics to detect and analyze potential confounding variables, and finally, we produce graphical representations of the results.

1.4.3 Part III

Chapter 7 includes the final publication of this dissertation. In this work, a more meta-level lens is used to approach the field of process mining from a methodological perspective (Koorn et al., 2021). This part is inspired by the focus on healthcare in the case studies presented in the first and second parts of this dissertation. In these two parts, qualitative evaluations play a key role. In the final publication, we use a literature study to reflect on and enhance existing evaluation methodologies in process mining projects. Specifically, the study exclusively focuses on the evaluations of process mining projects in which domain experts are involved. The contribution of the work is twofold. First, it analyzes and describes current evaluation practices. Second, it proposes a set of guidelines to support future qualitative evaluations. The literature study captures process mining case studies in which domain experts are involved. It analyzes how the evaluation in these studies is conducted in terms of goals and methods. This analysis shows that evaluations often have a qualitative

nature, but that a systematic approach is lacking. In this work, we build on this by providing a set of six validation guidelines. These guidelines aim to enhance the accuracy and meaning of findings in qualitative evaluations.

1.5 List of Publications

This dissertation is based on a number of publications by the author. In total, one journal article, three conference papers, and one workshop paper are included. Additionally, one chapter of the dissertation is based on work-in-progress.

Journal article:

Koorn, J. J., X. Lu, H. Leopold & H. A. Reijers (2022a), "From action to response to effect: mining statistical relations in work processes", *Information Systems*, DOI: `https://doi.org/10.1016/j.is.2022.102035`.

Conference articles:

Koorn, J. J., X. Lu, H. Leopold & H. A. Reijers (2020), "Looking for meaning: discovering action-response-effect patterns in business processes", in: *International Conference on Business Process Management*, Springer, pp. 167–183.

Koorn, J. J., I. M. Beerepoot, V. S. Dani, X. Lu, I. Van de Weerd, H. Leopold & H. A. Reijers (2021), "Bringing rigor to the qualitative evaluation of process mining findings: an analysis and a proposal", in: *2021 3rd International Conference on Process Mining (ICPM)*, IEEE, pp. 120–127.

Koorn, J. J., X. Lu, F. Mannhardt, H. Leopold & H. A. Reijers (2022c), "Uncovering complex relations in patient pathways based on statistics: the impact of clinical actions", in: *Proceedings of the 55th Hawaii International Conference on System Sciences*.

Workshop article:

Koorn, J. J., X. Lu, H. Leopold & H. A. Reijers (2019b), "Towards understanding aggressive behavior in residential care facilities using process mining", in: *International Conference on Conceptual Modeling*, Springer, pp. 135–145.

Work in progress:

Koorn, J. J., X. Lu, H. Leopold & H. A. Reijers (2022b), "Mining statistical relations for better decision making in healthcare processes".

1.5.1 Not Included Work

Next to that, the author has been involved in a number of other research papers that are *not* included in this dissertation.

Journal article (under review):

Beerepoot, I. M., N Martin & J. J. Koorn (Under review), "Evaluating process mining insights with healthcare professionals: the FEI funnel", *Business Process Management Journal*.

Conference articles:

Koorn, J. J., H. Leopold & H. A. Reijers (2018b), "A task framework for predicting the effects of automation", in: *Proceedings of the 26th European Conference on Information Systems (ECIS)*, p. 141.

Beerepoot, I. M., J. J. Koorn, I. Van de Weerd, B. Van den Hooff, H. Leopold & H. A. Reijers (2019), "Working around health information systems: the role of power", in: *Proceedings of the 40th International Conference on Information Systems (ICIS)*, p. 2.

Workshop articles:

Dani, V. S., H. Leopold, J. M. E. M. Van der Werf, X. Lu, I. M. Beerepoot, J. J. Koorn & H. A. Reijers (2021a), "Towards understanding the role of the human in event log extraction", in: *19th conference on Business Process Management*, *17th International Workshop on Business Process Intelligence (BPI'21)*.

Dani, V. S., E. Mahendrawathi, J. J. Koorn, J. M. E. M. Van der Werf, H. Leopold & H. A. Reijers (2021b), "Pair modeling: does one plus one add up?", in: *19th conference on Business Process Management*, *14th International Workshop on Social and Human Aspects of Business Process Management (BPMS2'21)*.

Koorn, J. J., H. Leopold & H. A. Reijers (2018a), "A task framework for predicting the effects of automation", in: *39th International Conference on Information Systems, ICIS 2018*, *6th International Workshop on the Changing Nature of Work (CNoW): Bridging the Workplace of People, Data and Things*.

Koorn, J. J., S. Zhang & J. V. Nickerson (2019a), "A fresh look at the impact of technology on occupations through software", in: *40th International Conference on Information Systems, ICIS 2019*, *8th International Workshop on the Changing Nature of Work (CNoW): Digital Ecosystems and the Changing Nature of Work - Reconfiguring technology, practices, and organizations*.

Part I: From Pattern to Dependency Relation

CHAPTER 2

Towards Understanding Aggressive Behavior in Residential Care Facilities Using Process Mining

Reading Guide. In this chapter we look into the patterns related to aggressive incidents. We approach these incidents from a process mining perspective to study not only single incidents but also the sequence of incidents. This sequence of incidents we refer to as aggressive behavior. In this work, we make some interesting first observations regarding aggressive behavior. More importantly, this work forms the foundation from which we expand into building a technique. This technique we will address in the following chapters of the first part.

This chapter is based on the following publication:

Koorn, J. J., X. Lu, H. Leopold & H. A. Reijers (2019b), "Towards understanding aggressive behavior in residential care facilities using process mining", in: *International Conference on Conceptual Modeling*, Springer, pp. 135–145.

2.1 Introduction

Many people with intellectual disabilities live in residential care facilities, where they can get support for their daily needs. A central goal of these facilities is to make sure that their clients have an appropriate quality of life. While this primarily means the care facilities attend to the needs of their clients, it also requires them to deal with undesired behavior of clients, such as aggression.

Aggression is a complex phenomenon and has many facets. It involves physical aggression towards other people, verbal aggression, physical aggression towards objects, self-injurious behavior, and sexually inappropriate behavior (Crocker et al., 2006). In general, aggressive behavior is considered a threat to both staff and clients. Studies have shown that staff members who deal with aggressive behavior are more likely to experience stress and even have burnouts (Hensel et al., 2014; Hensel et al., 2012; Mills & Rose, 2011). What is more, aggressors can be met with severe restrictive measures (e.g., seclusion) as well as suffer physical injuries as a consequence of their aggressive behavior (Nieuwenhuis et al., 2017; Tenneij & Koot, 2008; Van den Bogaard et al., 2018). Hence, there is a desire to minimize the impact of aggression on staff and clients and, ultimately, prevent aggression incidents from happening altogether.

To do so, many researchers have investigated which factors may contribute to aggressive behavior. By mostly focusing on client characteristics, they found links to factors such as: age (Cooper et al., 2009; Tyrer et al., 2006), severity of intellectual disability (Cooper et al., 2009; Crocker et al., 2006), and gender (Cooper et al., 2009; Crocker et al., 2006). However, some researchers argue that focusing on client characteristics is too limited (McClintock et al., 2003). More recent research has followed up on this and started investigating how characteristics from the aggression incident itself, such as time, location, and trigger, can help to understand aggressive behavior (Nijman & Palmstierna, 2002). In that light, observational data describing aggression incidents are increasingly digitally recorded in Information Systems.

Nevertheless, what all these studies have in common is that they take a static perspective. Due to the increasing digitization of behavioral data, new opportunities emerge to analyze this phenomenon from new dynamic perspectives. That is, to consider the changes in the behavior of clients over time rather than for a single incident or client. Such a perspective has the potential to uncover, for instance, how aggression evolves over time and whether different aggression incidents are related. Process mining is such an emerging field that provides techniques to support the analyses of data from a causal perspective (Van der Aalst, 2016).

Therefore, we use this chapter to study aggression incidents using the technology of process mining. We analyze data from 1,115 clients from a Dutch residential care facility over a period of three years. We find that, on a high level, we can distinguish between cases exclusively showing the same type of aggressive behavior and cases showing a variety of types of aggressive behavior. Moreover, we find that although the division into both groups is useful, the repetition of the same type of aggressive behavior is the most frequently observed behavior. Lastly, physical aggression towards other people plays a key role as it occurs most often and usually follows after any other type of aggressive behavior.

The rest of the chapter is organized as follows. First, the current status of process mining in healthcare is discussed in Section 2.2. Then, Section 2.3 provides a detailed description of the methodology of this chapter. This is followed by Section 2.4 describing the results of the analyses. In Section 2.5 we discuss these results and the limitations of this study. Finally, in Section 2.6 the conclusions are presented.

2.2 Process Mining in Healthcare

Process mining is a family of data analysis techniques that aims to discover, monitor, and improve organizational processes by analyzing data from so-called event logs (Van der Aalst, 2016). These event logs are generated by various information systems that are used in organizations and, therefore, capture how organizational processes are executed in real life. Process mining has been applied in various healthcare settings (Rojas et al., 2016). Among others, process mining has been used to analyze patient care processes (Fei, Meskens, et al., 2010; Kim et al., 2013), dentistry processes (Bakhshandeh et al., 2017; Mans et al., 2012a), and cancer treatment processes (Binder et al., 2012).

Despite the general potential of process mining in healthcare, the application of process mining in this domain is often associated with particular challenges (Mans et al., 2012b). One of the most common issues in this context is the absence of accurate timestamps. Healthcare information systems often only capture the day of an activity and not the exact point in time. As a result, the exact order of certain activities remains unclear. Another issue is the absence of a clear start and endpoint of a process. Often, the data entries for a single patient span several years. This, however, does not mean that all this data relates to the same treatment process. It could be that the patient received several treatments at the same time or had a recurrence after a couple of years. We have encountered both issues, and the steps taken to handle these issues are discussed in the following section.

2.3 Methodology

Below, we will first present an overview of the data and describe how the data was extracted. Then, we explain how the extracted data is converted to event logs. Once the event logs are created, we can discuss the process of data cleaning and filtering. Finally, we explain how the cleaned and filtered event logs are used to discover patterns of aggressive behavior.

2.3.1 Data Extraction and Overview

For our research, we acquired a data set from a Dutch healthcare organization that operates 54 residential care facilities in the Netherlands. It specializes in providing care for people with mild intellectual disabilities (IQ between 50 and 70), borderline intellectual functioning (IQ between 70 and 85), co-occurring psychiatric disorders, and physical disabilities. The total number of clients this organization cares for is about 3,000 (the exact numbers vary over the period we consider). Our data set covers all aggression incidents reported on all wards starting from the 1st of January

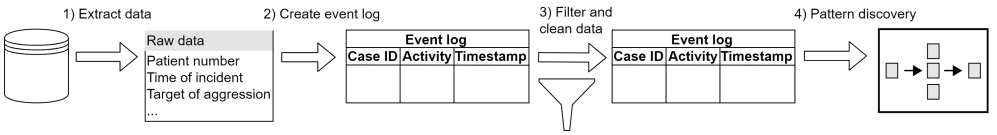


Figure 2.1: Methodology

Client ID	Timestamp	Time slot	Target	Means	Consequences
L002eR	04/09/2015	16:00-17:00	Themselves	Teeth	Visible injuries
L002eR	11/09/2015	17:00-18:00	Themselves	Hands	Visible injuries
LHZ02	03/05/2016	11:00-12:00	Staff member	Hands	Pain 5 min.
LHZ02	22/05/2016	10:00-11:00	Objects	Hands	Damage to property
LH88E3	26/12/2016	19:00-20:00	Staff member	Verbal	Felt threatened
H030E	02/02/2017	14:00-15:00	Objects	Feet	Damage to property

Table 2.1: Snippet of the raw data as an example

2015 until the 31st of December 2017. The total number of incidents in our data set is 21,706.

Table 2.1 shows an extract of the raw data we obtained. We can see that each entry about an aggression incident includes: a reference to the aggressor (client ID), the day (timestamp), an approximate point of time (time slot), information about the target (e.g. objects), the means (e.g. hands), and the consequences (e.g. pain 5 min.). The caretakers self-report on these incidents through a standardized digital reporting form with pre-defined categories.

To apply process mining to our data set, several requirements need to be met. The starting point of every process mining analysis is a so-called event log. These event logs must contain at least three specific attributes: (1) a unique case identifier (case ID), (2) an activity description, and (3) an appropriate timestamp. As illustrated by the extract shown in Table 2.1, the raw data set does not fulfill these criteria. Among others, a case ID is missing (note that a client ID is not a case ID since a single client might be associated with several cases) and there is no clear notion of an activity. Against this background, our methodology includes several preparatory steps represented in Figure 2.1. The first step is described above in the section *Data Extraction and Overview*.

2.3.2 Event Log Creation

The second step was the creation of an event log that is suitable for a process mining analysis. To obtain such an event log, we had to introduce 1) a suitable activity attribute, 2) an appropriate case ID, and 3) a proper timestamp. Table 2.2 shows an extract of the final event log.

Activity Design. In order to introduce a suitable activity attribute, we designed a simple algorithm that categorized each incident into one out of four aggression categories: physical aggression towards other people (PP), self-injurious behavior (SIB), physical aggression towards objects (PO), and verbal aggression (VA). Due

Case ID	Activity	Timestamp
L002eR-2015-09	Auto mutilation (SIB)	04-09-2015 16:00
L002eR-2015-09	Auto mutilation (SIB)	11-09-2015 17:00
LHZ02-2016-05	Physical aggression towards others (PP)	03-05-2016 11:00
LHZ02-2016-05	Physical aggression towards objects (PO)	22-05-2016 10:00
LH88E3-2016-12	Verbal aggression (VA)	26-12-2016 19:00
H030E-2017-02	Physical aggression towards objects (PO)	02-02-2017 14:00

Table 2.2: Snippet of the event log

to privacy reasons, we did not include data on sexually inappropriate behavior. We designed the algorithm using a rule book approach based on (Nijman et al., 1999). The rules were adjusted based on the input from experts of the healthcare organization we collaborated with. It is important to note that a single incident can be associated with multiple types of aggressive behavior. In such cases, only the most severe type of aggressive behavior is considered. This is in accordance with previous literature (Nijman et al., 1999), which proposed a hierarchy of the severity of aggressive behavior. The four categories we consider are mentioned in their order of severity, that is, PP is the most severe and VA is the least.

Case ID design. Choosing the client ID as the case ID may result in a process that spans over three years. To illustrate that this may lead to a distorted view of the data, consider the example of two clients. Client 1 has an SIB incident in early January 2015 and another in late December 2017. Client two has an SIB incident in early May 2016 and another in late May 2016. If the client ID is chosen as the case ID, both clients would be considered to show the same pattern of aggressive behavior, although in the former case it is hard to argue that the two incidents are related.

Based on validation with experts, we decided to slice the data for each client in three ways: traces of 24 hours, seven days, and one month. To this end, we created a case ID by combining the client ID with the year, month, week, and hour values from the timestamp. Thus, for example, one specific case for the client LHZ02 would be the case LHZ02-2015-01 including all incidents from January 2015. We found similar patterns using all three forms of data slicing. Thus, for consistency purposes, we present our results on the basis of the month level.

Timestamp Design. In our raw data set, the date (dd-mm-yyyy) of an incident was provided. To make the timestamp more precise we added available data about the time slot and combined both into a single timestamp.

2.3.3 Data Cleaning and Filtering

The event log obtained after the preparatory steps outlined above contains a total of 21,706 aggression incidents related to 1,115 clients or 8,557 cases (client months). To be able to detect relevant patterns using process mining, we applied three specific filters: (1) an event filter, (2) a time filter, and (3) a recurrence filter. After applying all

three filters our final sample size used for process mining contained 16,794 incidents (78.2% of the total sample) spread over 822 clients included in 4,149 cases (48% of the total sample).

The *event filter* removed all incidents for which no aggression incident type could be determined (N=322). The *time filter* was applied to exclude cases for which erroneous data was recorded for the considered time frame from 2015 to 2017 (N=20). Finally, we applied a *recurrence filter* to exclude all cases that only contained one single activity (N=4570). While this had a considerable impact, our analysis aimed to discover behavioral patterns. In order to detect a pattern at least two activities need to be included in a case.

2.3.4 Pattern Discovery

To discover relevant patterns, we used the commercial process mining tool Disco[1]. After we completed the exploratory analysis, it became clear that the data could be split into two large groups: cases that exclusively contain the same type of aggressive behavior (e.g., case L002eR-2015-09 in Table 2.2), and cases that contain a mix of aggressive behavior types (e.g., case LHZ02-2016-05 in Table 2.2). Based on this insight, we analyzed the patterns for each of these groups separately.

More specifically, we analyzed the data of each group from three angles: frequency, time, and variation. *Frequency* relates to the visual patterns and captures information about the number of incidents and transitions. *Time* captures details regarding the duration between incidents. *Variation* looks at the various traces (i.e., specific orders of incidents) of which the patterns consist. When performing the analyses, we looked at both the grander schemes of all types of aggression combined per group of behavior as well as at each type of aggression within each group of behavior.

Finally, to make sure our understanding of the discovered patterns is correct, we validated our results with a behavioral expert from the care organization.

2.4 Results

We first present the patterns that represent mixed aggressive behavior. Then, we go into detail about the patterns discovered capturing homogeneous aggressive behavior.

2.4.1 Mixed behavior

The patterns resulting from all cases containing at least two types of aggressive behavior are summarized in Figure 2.2. It shows the frequency of occurrence of each type of aggressive behavior as well as the transition frequencies (see respective arc label). In total, Figure 2.2 captures 2,170 cases and 10,713 incidents.

Figure 2.2 illustrates that there are two types of aggressive behavior that are relatively frequent: physical aggression against other people (PP) and verbal aggression (VA). The other two types of aggression (physical aggression towards objects (PO) and self-injurious behavior (SIB)) account for less than 25% of all incidents. Further-

[1] https://fluxicon.com/disco/

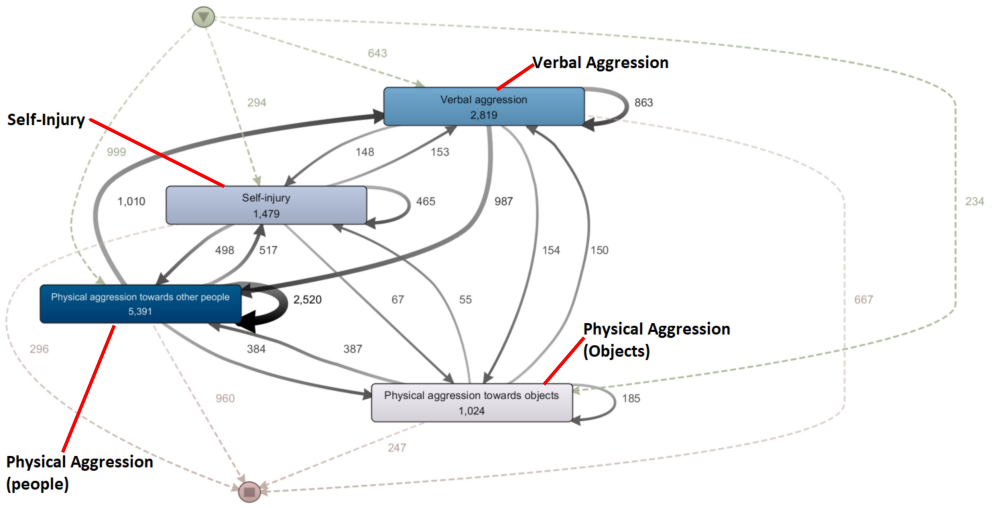


Figure 2.2: Directly-follows graph for the mixed behavior group, for a detailed account regarding the way this graph should be interpreted please refer to (Van der Aalst, 2016, p. 223)

more, we can see that around half the cases start or end with an incident of PP. This indicates that PP serves as both enabler of and amplifier to other types of aggression.

		To			
		VA	*PP*	*PO*	*SIB*
From	*VA*	**62.0%**	29.7%	4.3%	4.0%
	PP	12.0%	**76.4%**	4.8%	6.8%
	PO	15.9%	**44.1%**	35.2%	4.9%
	SIB	7.2%	26.3%	3.0%	**63.5%**

Table 2.3: Relative transition frequencies

Table 2.3 provides further insights into the underlying patterns by showing the relative transition frequencies (i.e. in how many percent of the cases we observe a transition from one type of behavior to another). The bold figures indicate the most frequent transitions and the underlined figures are the second-most frequent transitions for each type. The data allows for four interesting observations. First, the repetition of the same type of aggressive behavior is the most frequent pattern (see VA-VA, PP-PP, and SIB-SIB). Second, focusing on patterns without repetition, the most frequent transition is to PP (see VA-PP and SIB-PP). Third, PO shows the reverse of both previous points, its most frequent transition is to PP and its second most frequent transition is to itself (PO-PO). Fourth, PO and SIB are rarely followed by other types of aggressive behavior besides a repetition of the same behavior or a transition to PP.

2.4.2 Homogeneous Behavior

The results for the group with homogeneous behavior are summarized in Figure 2.3. It shows the frequency of occurrence of each type of aggressive behavior as well as the

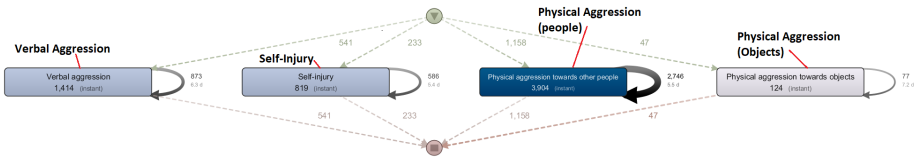


Figure 2.3: Directly-follows graph for the homogeneous behavior group

transition frequencies (see respective arc label). In total, Figure 2.3 captures 1,979 cases and 5,388 incidents.

The numbers in Figure 2.3 show that PP is the most frequent type of repetitive behavior (60%) followed by VA (25%). Only about 10% of the incidents relate to SIB and 4% to PO incidents. The same distribution holds when we are looking at the number of cases rather than the number of incidents.

Looking into the average number of incidents per month allows for further insights. Interestingly, this number varies per type of aggression: SIB (3.5 incidents) and PO (3.4 incidents) have a higher average than PP and VA (both 2.6 incidents). In addition, when we consider the aspect of time, we observe that the median duration between two incidents is considerably longer for PO. Here the median between two PO incidents is the longest with 6 days, whereas the median for PP and VA is 4 days and SIB is 3 days. This shows that there is a negative relation between the average number of incidents per month and the median duration between two incidents. In other words, the higher the average number of incidents per month, the shorter the median duration between two incidents. However, PO is an exception to this since the average number of incidents per month is on the high end, but the median number of days between two incidents is high as well.

2.5 Discussion

Taking a holistic view of the results from our analysis, we identify three findings: (1) repetition of the same type of behavior occurs most frequently, (2) PP plays a central role, and (3) PO exhibits deviating patterns.

First, we see that *repeating* the same type of aggressive behavior is the most frequently observed transition between two incidents. Although the number of cases included in the mixed and within the homogeneous group is roughly equal, we can see that the most observed behavior within the mixed group is still a repetition of the same type of behavior. Through the validation with a behavioral expert, we found that this can be explained by the fact that the behavior of clients is usually determined by their habits. As such, it is unexpected for clients to change their behavior. However, most clients also follow behavioral adaptation treatments aimed at changing this behavior. In this light, it is unexpected to see such a high percentage of repeating the behavior. The behavioral expert hypothesized that the high percentage of repeating behavior could indicate that the treatments are often not successful.

Second, in both groups of aggressive behavior, PP is the most frequently observed behavior accounting for around 60% of the total number of incidents. Besides being the most frequently observed type of aggressive behavior, in Figure 2.2, we see that PP in a mixed behavior environment is the epicenter of the system. There is relatively little interaction among the other forms of aggressive behavior except when accommodated through PP. Through the expert validation we found that one possible explanation is that when an aggressive incident occurs, a caretaker usually intervenes, thereby moving the target of aggression towards him/herself.

Third, PO represents an exception to the general patterns observed in both mixed and homogeneous behavior groups. We observed that cases showing this type of aggression follow a different evolution of behavior compared to the other types of aggression. We see for example that the pattern PO-PO is not as frequently observed as with the other types of aggression. In addition, if it is observed, the time between two incidents is relatively long. This is interesting as it could indicate that PO is a fundamentally different type of aggression compared to the other types of aggression.

It is important to note that our findings are subject to a number of limitations. More specifically, we identify three main limitations to this study. First, our data set is not representative in a statistical sense. While our data set contains a considerable number of patients and incidents, we cannot extrapolate our findings to other care facilities. Second, our data set may contain different kinds of biases due to manual reporting. For example, caretakers may report incidents in bulk, meaning multiple incidents are included in a single report at the end of the day or week in order to reduce administrative load. Although we know that this was not used very frequently, the existing bulk reports are counted as a single incident as there is no indication about how many incidents are reported in one bulk report. Another reporting bias relates to VA, SIB, and PO. From the discussion with experts, we learned that these incidents are sometimes reported less frequently since they are perceived as less severe by staff. Last, in terms of data slicing we use a blunt-force approach to slice the cases into months, ideally, we would introduce a time box (such as introduced in Rinner et al., 2018) to avoid this challenge in future work.

2.6 Conclusion

Research looking into the aggressive behavior in clients with intellectual disabilities has exclusively investigated this phenomenon from a static perspective. In this chapter, we advanced on this by using a process mining approach to look into the evolution of aggressive behavior in clients with intellectual disabilities. This enabled us to obtain insights into the relations among different types of aggression and to infer patterns of aggressive behavior. We found that there are mainly two groups of clients: those with homogeneous and those with mixed aggressive behavior. Among others, we found that repetitive behavior is the most frequently observed behavior. In addition, results show that physical aggression towards other people plays a central role in a majority of the behavioral patterns

With these insights, this research contributed to a better understanding of aggressive behavior aiding further development in this field. From a practitioner's point of view, the discovered patterns can aid the development of prevention and treatment

techniques. Despite the interesting findings, we are aware of the limitations of our study. Therefore, we plan to follow up on our work by further developing insights into the discovered patterns. We aim to identify causal relations between the static characteristics of incidents and the behavioral patterns, thereby uncovering more refined patterns of aggression. Furthermore, we plan to include additional care facilities to increase the external validity of our results.

CHAPTER 3

Looking for Meaning: Discovering Action-Response-Effect Patterns in Business Processes

Reading Guide. In the previous chapter we discovered patterns in aggressive behavior by applying process mining techniques. This provides insights into what the behavior looks like. However, questions remain regarding the how question? In other terms, what are the causes and effects of the aggressive incidents? In this chapter, we further address these types of questions as we set out to look into the potential dependency relations between the activities within an aggressive incident. We do this by proposing a novel technique to detect, analyze, and visualize potential dependency patterns in work processes.

This chapter is based on the following publication:

Koorn, J. J., X. Lu, H. Leopold & H. A. Reijers (2020), "Looking for meaning: discovering action-response-effect patterns in business processes", in: *International Conference on Business Process Management*, Springer, pp. 167–183.

3.1 Introduction

The desire to improve organizational processes has led to the adoption of process mining in many industries (De Weerdt et al., 2013; Rojas et al., 2016). One of the key advantages of process mining is that it enables organizations to understand, analyze, and improve their processes based on process execution data, so-called event logs. Such event logs capture how organizational processes are executed and can be extracted from various information systems that are used in organizations (Van der Aalst, 2016).

While the advantages of process mining have been demonstrated in many domains, the application of process mining is still associated with different challenges. One particularly important challenge is to provide the user with a process representation that a) is easy to understand and b) allows the user to obtain the required insights about the process execution. To this end, various process discovery algorithms have been proposed, including the heuristic miner (Weijters & Van der Aalst, 2003), the fuzzy miner (Günther & Van der Aalst, 2007), and the inductive miner (Leemans et al., 2013). What all of these algorithms have in common is that they focus on discovering the control flow of a process, i.e., the order constraints among events.

In many scenarios, however, the control flow perspective is not sufficient for understanding and improving the process. As an example, consider the care process of a residential care facility supporting clients with their daily needs. The main goal of this process is to ensure the well-being of clients. One of the main factors negatively affecting the well-being of clients is incidents of aggressive behavior, e.g. when clients verbally or physically attack other clients or staff. Staff responds to aggressive incidents with one or multiple measures ranging from verbal warnings to seclusion. A key question in the context of process improvement is which of these measures leads to desired (i.e., de-escalation of aggressive behavior) or undesired (i.e., escalation of aggressive behavior) outcomes.

From a process perspective, this requires understanding the *action-response-effect* patterns. In the healthcare process, we consider the aggressive incidents as *actions*, the countermeasures taken to the incident as *responses*, and the follow-up incidents as *effects*. Action-response-effect patterns are not accounted for in existing discovery algorithms. As a result, their application to such event logs leads to a process representation that is either hard to read (because it contains too many connections) or it does not allow the user to obtain actual insights about the process (because it does not show the effect of behavior).

Recognizing the limitation of existing algorithms with respect to showing meaningful insights into action-response-effect patterns, we use this chapter to propose a novel discovery technique. We leverage well-established statistical tests to analyze event logs in order to discover simplified graphical representations of business processes. We simplify the resulting models by highlighting the statistically significant dependency relations according to statistical tests, while insignificant relations are hidden. We conduct an evaluation with an event log from a Dutch residential care facility containing a total of 21,706 aggression incidents related to 1,115 clients. We show that our technique allows us to obtain important insights that existing discovery algorithms cannot reveal.

EID	CID	Timestamp	action	Response(s)
1	1	12-05 09:53	VA	Warning
2	1	13-05 13:35	PO	Distract Client, Seclusion
3	1	26-05 09:32	VA	Warning
4	1	26-05 11:02	PP	Distract Client
5	2	21-06 14:51	VA	Distract Client
6	1	23-06 21:23	VA	Distract Client
7	2	24-06 17:02	VA	-
8	3	29-08 11:22	VA	Warning
9	3	31-08 08:13	PO	Warning, Seclusion
10	3	31-08 10:48	PP	Distract Client

Legend: EID = Event identifier, CID = Client identifier, VA = Verbal Aggression, PP = Physical Aggression (People), PO = Physical Aggression (Objects),

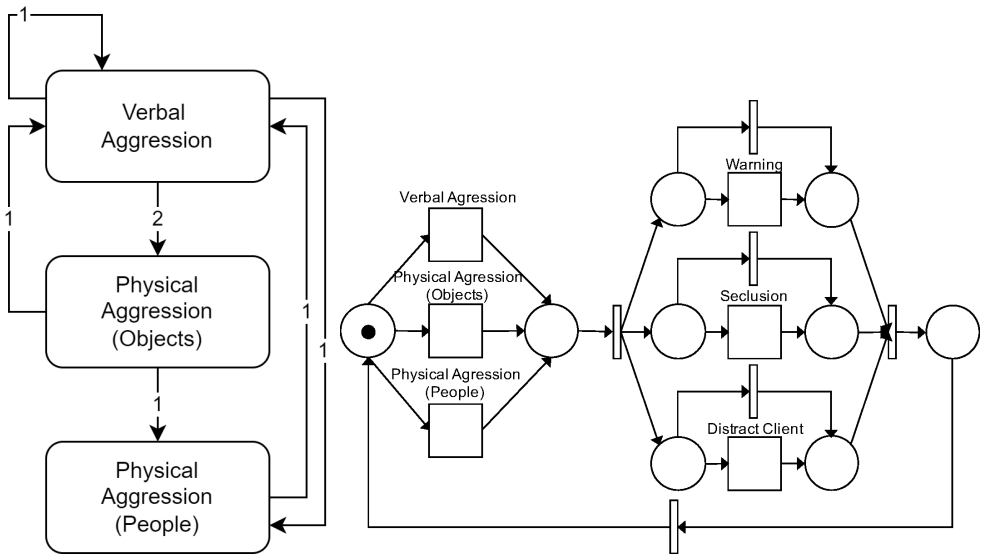
Table 3.1: Excerpt from an action-response log of a care process

The rest of the chapter is organized as follows. Section 3.2 describes and exemplifies the problem of discovering *action-response-effect* patterns. Section 3.3 introduces the formal preliminaries for our work. Section 3.4 describes our proposed technique for discovering *action-response-effect* patterns. Section 3.5 evaluates our technique by applying it to a real-world data set. Section 3.6 discusses related work and Section 3.7 concludes the chapter.

3.2 Problem statement

Many processes contain action-response-effect patterns. As examples consider healthcare processes where doctors respond to medical conditions with a number of treatments, service processes where service desk employees respond to issues with technical solutions, and marketing processes where customers may respond to certain stimuli such as ad e-mails with increased demand. Let us reconsider the example of the healthcare process in a residential care facility in order to illustrate the challenge of discovering an understandable and informing process representation from an event log containing action-response relations. The particular aspect of interest are incidents of aggressive behavior from the clients and how these are handled by staff. Table 3.1 shows an excerpt from a respective event log. Each entry consists of an event identifier EID (which, in this case, is equal to the incident number), a case identifier CID (which, in this case, is equal to the client identifier), a timestamp, an aggressive incident (action), and one or more responses to this event.

Figure 3.1 a) shows the directly–follows-graph that can be derived from the events of this log. It does not suggest any clear structure in the process. Although this graph is only based on twelve events belonging to three different event classes, it seems that almost any behavior is possible. In addition, this representation does not provide any insights into certain hidden patterns (Van der Aalst, 2019). However, if we take a closer look, we can see that there are effects to a certain response. For instance, we



(a) Directly-follows graph (b) Petri net

Figure 3.1: Representations resulting from the action-response log from Table 3.1

can see that over time the aggressive incidents related to client 1 escalate from verbal aggression to physical aggression against objects and people. The verbal aggression event in June (EID = 6) is probably unrelated to the previous pattern since it occurs several weeks after. To gain an even deeper understanding, we need to take both the response and its effect into account. Both client 1 and 2 escalate from verbal aggression to physical aggression after the verbal aggression was only countered with a warning.

These examples illustrate that explicitly including the responses and effects in the discovery process is important for answering the question of how to possibly respond to an action when a certain effect (e.g. de-escalating aggressive behavior) is desired. Therefore, our objective is to discover a model that: (1) shows the action-response-effect process, and (2) reveals the dependency patterns of which responses lead to desired or undesired outcomes (effect). There are two main challenges associated with accomplishing this:

1. *Graphical representation*: From a control-flow perspective, action-response relations are a loop consisting of a choice between all actions and a subsequent and-split that allows to execute or skip each responses. Figure 3.1 b) illustrates this by showing the Petri net representing the behavior from the log in Table 3.1. Obviously, this representation does not allow understanding which responses lead to a desired or undesired effect.
2. *Effective filtering mechanism*: The possible number of responses calls for a filtering mechanism that allows inferring meaningful insights from the model. In the example above, we only have three event classes and three event response classes (plus the “no response"). This results in eight possible responses. In the

case of 5 response event classes, we already face 32 (=2^5 possible responses. Including all these response arcs in a model will likely lead to an unreadable model that does not allow inferring the desired insights.

In the next sections, we propose a technique that creates graphical representations of dependency patterns in action-response effect logs.

3.3 Preliminaries

In this section, we formalize the concept of action-response-effect event logs.

Definition 3.3.1 (action-response-effect Log) *Let $\mathcal{E}$ be the universe of event identifiers. Let $\mathcal{C}$ be the universe of case identifiers. Let $d_1, ..., d_n$ be the set of attribute names (e.g., timestamp, resource, location). Let A be the set of actions and R a finite set of responses. An action-response log L is defined as $L = (E, \pi_c, \pi_l, \pi_r, \pi_{d_1}, ..., \pi_{d_n}, <)$, where*

- *$E \subseteq \mathcal{E}$ is the set of events,*
- *$\pi_c : E \to \mathcal{C}$ is a surjective function linking events to cases,*
- *$\pi_l : E \to A$ is a surjective function linking events to actions,*
- *$\pi_r : E \to 2^R$ is a surjective function linking events to a set of responses,*
- *$\pi_{next} : E \to C$ is a surjective function linking events to the effects,*
- *$\pi_{d_i} : E \to \mathcal{U}$ is a surjective function linking the attribute d_i of each event to its value,*
- *$< \subseteq E \times E$ is a strict total ordering over the events.*

Given an action-response log L according to Definition 3.3.1, we shall use the shorthand notation $\sigma = \langle e_1, \ldots, e_n \rangle$ in the remainder of this chapter to refer to an event trace that consists of n events with an identical case identifier. Furthermore, for any pair of events e_i and e_j with $i < j$, it holds that $e_i < e_j$ according to the strict total ordering of the events in log L.

The set of response events $\{r_1^e, \ldots, r_n^e\}$ of an event e is given by the function π_r, we write $\pi_r(e) = \{r_1^e, \ldots, r_n^e\}$. For each trace $\sigma = \langle e_1, \ldots, e_n \rangle$, the sequence of responses is $\langle \pi_r(e_1), \ldots, \pi_r(e_n) \rangle$. For example, in the action-response log listed in Table 3.1, for event e_1: $\pi_c(e_1) = 1$ is the case of event e_1, $\pi_l(e_1) =$ "Verbal Aggression" is the action of e_1, and $\pi_r(e_1) = \{$"Warning"$\}$ is the set of responses of e_1.

Effects of Responses

As we discussed, we aim to investigate whether a certain response to an action has an effect on the follow-up event. As such, we measure the effectiveness of a response to an action by studying the effect. For this aim, we first define the effects of events by using the function π_{next} and introducing parameter ϵ for elapsed time. For each trace $\sigma = \langle e_1, ..., e_n \rangle$, we define the effect for each e_i, where $1 \leq i < n$ as follows: if the elapsed time to the next event e_{i+1} is less than ϵ, the effect $\pi_{next}(e_i)$ of e_i is the action of e_{i+1}, else we say that the effect is a silent action τ. Formally, if $\pi_{time}(e_{i+1}) - \pi_{time}(e_i) \leq \epsilon$, then $\pi_{next}(e_i) := \pi_{action}(e_{i+1})$, else $\pi_{next}(e_i) := \tau$.

To test the hypothesis of whether an effect is independent of the response to an action, the number of observed events is compared to the number of expected events of different responses and effects. To calculate the number of observed events, we create a matrix (table) where each cell is filled with the number of observed events of a response and an effect. Let $a \in A$ be an action, $R = \{r_1, \cdots, r_m\}$ be a set of responses,

ID	Timestamp	Action	Response(s)	Effect
1	12-05 09:53	VA	Warning	PO
1	13-05 13:35	PO	Distract Client, Seclusion	τ
1	26-05 09:32	VA	Warning	PP
1	26-05 11:02	PP	Distract Client	τ
2	21-06 14:51	VA	Distract Client	VA
1	23-06 21:23	VA	Distract Client	τ
2	24-06 17:02	VA	-	τ
3	29-07 11:22	VA	Warning	PO
3	31-07 08:13	PO	Warning, Seclusion	PP
3	31-07 10:48	PP	Distract Client	τ

Legend: VA = Verbal Aggression, PP = Physical Aggression (People), PO = Physical Aggression (Objects)

Table 3.2: Excerpt of the event log action-response-effect

and $C = \{c_1, \cdots, c_n\}$ a set of effects. We define a $|R| \times |C|$ matrix, where each row represents a response r_i, each column represents an effect c_j, and each cell counts the number of observed events that have response r_i and effect c_j. We have

$$freq_{a,R,C} = \begin{pmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,n} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{m,1} & f_{m,2} & \cdots & f_{m,n} \end{pmatrix}$$

where

$$f_{i,j} = freq_L(a, r_i, c_j) = \quad |\{e \in L \mid \pi_l(e) = a \wedge r_i \in \pi_r(e) \wedge \pi_{next}(e) = c_j\}| \tag{3.1}$$

For instance, given a log L as listed in Table 3.2, $freq_L$("VA","Warning","PO") $=|\{e_1, e_8\}|= 2$. Considering the observed section in Table 3.3 and omitting the column totals and row totals, it exemplifies a matrix $freq_{a,R,C}$. If the effects are independent of responses, then we should observe that the distribution of effects of a response is similar to the *total distribution*.

Each row r_i presents the distribution of effects $c_1, ..., c_k$ to the response r_i. To test whether each individual response r_i has an influence on the effects, we define $freq_{a,r,C}$ as a 2× |C| matrix:

$$freq_{a,r,C} = \begin{pmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,n} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,n} \end{pmatrix}$$

where $f_{1,j} = freq_L(a, r, c_j)$ and

$$f_{2,j} = \quad |\{e \in L \mid \pi_l(e) = a \wedge r \notin \pi_r(e) \wedge \pi_{next}(e) = c_j\}| \tag{3.2}$$

An example of $freq_{a,r,C}$ where r is "Terminate contact" is listed in Table 3.4. In the following section, our approach first performs a chi-squared test which allows us to calculate the expected values and test the dependency between responses and effects. The chi-square test compares the observed frequencies to the expected frequencies. If

Observed	**PO**	**PP**	**VA**	τ	*Total*
Warning	250	400	200	50	*900*
Held with force	20	50	50	10	*130*
Seclusion	30	50	20	10	*110*
Terminate contact	100	100	90	10	*300*
Distract client	100	150	40	10	*310*
Total	*500*	*750*	*400*	*100*	**1750**
Expected	**PO**	**PP**	**VA**	τ	*Total*
Warning	257.1	385.7	205.7	51.4	*900*
Held with force	37.1	55.7	29.7	7.4	*130*
Seclusion	31.4	47.1	25.1	6.3	*110*
Terminate contact	85.7	128.6	68.6	17.1	*300*
Distract client	88.6	132.9	70.9	17.7	*310*
Total	*500*	*750*	*400*	*100*	**1750**

Legend: VA = Verbal Aggression, PP = Physical Aggression (People), PO = Physical Aggression (Objects)

Table 3.3: Excerpt of the tables used to perform high-level statistical tests; horizontal categories: effect, vertical categories: response

Observed	**PO**	**PP**	**VA**	τ	*Total*
Terminate contact = 0	300	500	210	90	*1100*
Terminate contact = 1	100	100	90	10	*300*
Total	*400*	*600*	*300*	*100*	*1400*
Expected	**PO**	**PP**	**VA**	τ	*Total*
Terminate contact = 0	314.3	471.4	235.7	78.6	*1100*
Terminate contact = 1	85.7	128.6	64.3	21.4	*300*
Total	*400*	*600*	*300*	*100*	*1400*

Legend: VA = Verbal Aggression, PP = Physical Aggression (People), PO = Physical Aggression (Objects)

Table 3.4: Excerpt of the tables for an individual response used to perform statistical tests; horizontal categories: effect

they differ significantly, then the null hypothesis is rejected, which means we cannot rule out that there is a dependency relation between the response and the effect.

The complete event logs containing action-response-effect are used in the technique proposed in this chapter. The next section elaborates on this.

3.4 Discovery technique for action-response logs

In this section, we propose an algorithm to implement a discovery technique, see Algorithm 1. This technique builds on the formalization introduced previously. The goal is to create understandable process models that provide the user with the required insights into the execution of the process. First, we describe the pre-processing that needs to take place (Input for Algorithm 1). Then, we elaborate on the technique which consists of three main stages: (1) high-level statistics (line 1-7 in Algorithm 1), (2) detailed statistics (line 8-11), and (3) identifying influential points (line 12-19).

Algorithm 1 Compute graph

Input: Event log L
Output: Graph $G = (V, \prec)$
1: {***STAGE 1: High-Level Statistics***}
2: **for** $a \in A$ **do**
3: Initiate matrix $O[a] \leftarrow freq_{a,R,C}$ {*see Equation 3.2, calculate the observed values*}
4: Compute matrix $E[a]$ {*calculate the expected values by following the chi-square test, see (Cochran, 1952)*}
5: Compute $\chi^2_a = \frac{(O[a]-E[a])^2}{E[a]}$ {*To test the dependence between responses R and effects* $A \cup \{\tau\}$}
6: **if** χ^2_a is significant **then**
7: {$O[a]$ *differs from* $E[a]$*, thus responses R have a statistically significant influence on the effects C*}
8: {***STAGE 2: Detailed Statistics***}
9: **for** response $r \in R$ **do**
10: Compute matrix $O[a]_r$, $E[a]_r$, and $\chi^2_{a,r}$
11: **if** $\chi^2_{a,r}$ is significant **then**
12: {***STAGE 3: Influential Points***}
13: Compute adjusted standardized residuals ASR_c {*see Section 3.4.4*}
14: **for** effect $c \in A \cup \{\tau\}$ **do**
15: **if** ASR_c is significant **then**
16: {*draw the arc from r to c*}
17: $V \leftarrow V \cup \{a_s\} \cup \{r\}$, $\prec \leftarrow \prec \cup \{(a_s, r)\} \cup \{(r, c)\}$
18: **end if**
19: **end for** {*effect*}
20: **else**
21: {$\chi^2_{a,r}$ *is insignificant, i.e., r has no significant influence on C. We do not draw node r or any arc from r to C*}
22: **end if**
23: **end for** {*response*}
24: **else**
25: {*Observed* $O[a]$ *follows the expected values* $E[a]$*, thus response R has no statistically significant influence on the effects C; thus, no arcs are drawn*}
26: **end if**
27: **end for** {*action*}
28: **return** G

3.4.1 Pre-processing the Event Log

We first pre-process the log to obtain the effects of responses. As we are studying the effect of a response to an action, the duration between a response and its effect influences the likelihood of a dependency relation between the two. Let us return to our example: if there is an aggressive incident, there is a given response to this incident. However, if the next incident takes place after a long time (e.g. a year) we doubt that this new incident is still dependent on the response to the initial action. Thus, we defined the parameter epsilon (ϵ), see Section 3.3. ϵ represents the maximum duration between two events in which the first event is still considered to have an effect on the second event. For our specific example, we define ϵ equaling seven days in line with the input of an expert. Based on the ϵ, we introduce state τ. It represents the state we reach if there is no next incident within the defined duration of ϵ. In Table 3.2 we can see, for example, that distracting the client seems to be related to τ.

3.4.2 Computing High-Level Statistics

After pre-processing the event log, we investigate for each action the significant relation between the responses and the effects. In our example, the client shows a certain type of aggressive behavior (the action). Given this, we are interested in how the re-

sponse of a caretaker to that incident has an effect on the follow-up incident. Hence, we will explain the technique with a fixed initial action.

In Table 3.3, an example of the observed and calculated expected frequencies can be found given the action is physical aggression against objects (see lines 2 & 3 in Algorithm 1). This allows us to perform a Pearson Chi-square test (Cochran, 1952) (see line 4). Based on a confidence level α (usually 95%), the calculated χ^2 is compared to the Chi-square distribution to see if there is at least one pair of response-effect significantly different. If the chi-square score is insignificant, the action is excluded from the graphical representation (see line 22). If the Chi-square is significant (see line 5), this indicates that the effects may depend on the response. We then move to the second stage, see Section 3.4.3.

We demonstrate this first stage by applying it to a designed example based on our case study presented in Table 3.3. Based on the observed values, we can calculate the expected values in the table, for example, the expected value for the first cell: response *Terminate Contact* and effect $VA = \frac{\mathrm{N_r} \times \mathrm{N_c}}{E[\mathrm{a}][\mathrm{r}][\mathrm{c}]} = \frac{300 \times 400}{1750} = 68.6$. We know from Table 3.3 that there are five response classes and four effect classes, so the degrees of freedom: $c = (5-1) \times (4-1) = 12$. Given all this, we can calculate the Chi-square score for the overall table:

$$\chi^2_c = \sum_{i=1}^{5} \frac{(O_{\text{Warning,PO}} - E_{\text{Warning,PO}})^2}{E_{\text{Warning,PO}}} + ... + \frac{(O_{\text{Distract client},\tau} - E_{\text{Distract client},\tau})^2}{E_{\text{Distract client},\tau}}$$

$$\chi^2_{12} = \frac{(250-257.1)^2}{257.1} + ... + \frac{(10-17.7)^2}{17.7} = 63.47$$

Now we need to determine if this score is significantly different from the mean of the Chi-distribution (Fisher & Yates, 1938). The formula for calculating the p-value is complex and will thus not be discussed in detail in this chapter. For more details, we refer to (Fisher & Yates, 1938). In our case the p-value (< 0.001) corresponding to our Chi-square score is significant. This shows that for at least one pair of response-effect given action PO there is a significant difference from the expected frequency. Thus, we perform a Chi-square test for each individual response.

3.4.3 Computing Detailed Statistics

In the second stage of the algorithm, we perform the Chi-square test again on each response class to determine for which response we need to perform post-hoc statistical tests (see lines 6 - 8 in Algorithm 1). For this purpose, we create dummy variables. A dummy variable is made for each individual response, which takes the value of 0 or 1. The new table we create is a 2 x 4 table where the rows represent the response either taking a 0 or 1 value, see Table 3.4. Note that the degrees of freedom change to three now. The same formulas are used to calculate the individual response Chi-square score and the corresponding p-value. A Bonferroni correction (Haynes, 2013) is made to correct the critical value for the fact that on the same table multiple sets of analyses are performed. The Chi-square test identifies for which responses there is at least one effect that is significantly different from the expected frequency. If the Chi-square score is significant, we create a node for the response and perform post-hoc tests to identify the exact pairs of response-effect that are significant (see line 9).

We will demonstrate this stage on our designed example. We test five times (one for each response). Thus, we apply the Bonferroni correction (Haynes, 2013) on confidence level of 95% (meaning $\alpha = 0.05$): $\frac{0.05}{5} = 0.01$. If we take Table 3.4, we can use the same formulas as presented in the previous section to calculate the expected values. Note that we assume independence of responses. Thus, if there are two responses, the action is counted twice: once for response 1 and once for response 2. Therefore, the observed frequencies in Table 3.3 are not necessarily equal to those in Table 3.4. If we perform the Chi-square test for the response *terminate contact* we get a Chi-square score of 31.96 with a p-value < 0.001. Thus, for the response *terminate contract* there is at least one effect that is significantly different from the expected frequency. A post-hoc test will identify the exact pairs for which this is true.

3.4.4 Identifying Influential Points

In the last stage, the post-hoc tests are performed to test which exact pairs of response-effect have a significant contribution to the Chi-square test score. For this, the adjusted standardized residuals (ASR) (Agresti, 2003) are calculated (see line 10 in Algorithm 1). They represent a normalization of the residuals (observed - expected frequency). As the residuals can take either a positive or negative value we use two-sided testing. In order to improve the interpretability, we transform the α level into a critical value. We refer to (Fisher & Yates, 1938) for details on this. If $|ASR| > criticalscore$ the difference between observed and expected frequency is significant. A significant score means that a specific pair of response-effect has a significant impact on the overall test score. We will refer to these as *influential points*. If the score is insignificant, no arc is drawn for that pair of response-effect

For each influential point, arcs are drawn in the graphical representation (see lines 11-14). We first draw an arc from the action to the responses. On this arc, we indicate the observed frequency of the behavior. Then, we draw an arc from the response to the effect(s) for which we found a significant relation. On the arc, we display the observed frequency followed by the expected frequency in brackets. If the observed frequency is larger than the expected frequency, i.e. the response leads to an increase in the frequency of effect, we draw a thick arc. Correspondingly, if the observed frequency is lower than the expected frequency we draw a thin arc. The total number of graphical representations created equals the number of actions for which a significant Chi-square score is found (see line 25).

Now, we turn to the designed example. From the previous section, we know that the response *Terminate contract* results in a significant Chi-square score. To calculate which points are influential points we calculate the adjusted standardized residuals for each pair. To exemplify, we show the calculation of the ASR for the pair *Terminate contact = 1* and *VA*:

$$ASR = \frac{90 - 64.3}{\sqrt{64.3 * (1 - \frac{64.3}{300}) * (1 - \frac{64.3}{300})}} = 4.08$$

Given our Bonferroni correction gave us an alpha of 0.01 (see the previous section), we need to test on the 99 % confidence level. The critical absolute value for this is 2.57. Thus, if our ASR value is > |2.57| we mark it as an influential point and draw an

arc in the graphical representation. In the example of the pair Terminate contract = 1 and VA the ASR is larger than the critical score ($4.08 > 2.57$). Therefore, we draw a thick arc in the graphical representation of this example.

After conducting the above-described calculations for all actions, responses, and effects from the designed example, we obtain a total of three graphical representations (one for each action). In the next section, we evaluate the technique by applying it to a real-world data set.

3.5 Evaluation

The goal of this section is to demonstrate the capability of our technique to discover models that allow obtaining meaningful insights into action-response-effect patterns. To this end, we implemented our technique in Python and applied it to a real-world data set. The scripts are publicly available for reproducibility[1].

3.5.1 Data set

To evaluate our technique, we use a real-world data set related to the care process of a Dutch residential care facility. The event log contains 21,706 recordings of aggressive incidents spread over 1,115 clients. The process captured in this log concerns the aggressive behavior of clients in their facilities and the way client caretakers respond to these incidents. In the log, we can find an aggressive incident of the client, which fits in one of five action classes. This is followed by some measures taken by the staff as response to this incident, which fits in one of nine response classes. In line with the description of our technique, we transformed this log into suitable triples by adding the next aggressive incident of a client as the effect, given it took place within our ϵ. Thus, the effect can be one of five classes. As there are four different classes of actions, our technique will return four different graphical representations. Below, we present and discuss the results for one class of action: physical aggression against objects.

3.5.2 Results

Healthcare Case Results

After applying our technique to the data set, we obtain four graphs (one for each class of action). In Figure 3.2 we show the resulting graph when the initial action is physical aggression against objects ("po_s" in the figure). What we can see in the graph is the observed frequencies of the responses. For example, *terminate contact* has been observed 299 times in our data as a response to physical aggression against objects. Following this, the graph shows that in 43 events the effect to this response class is verbal aggression. From the data, we know that, in total, there are four action classes, nine response classes, and five effect classes. As such, the representation for one action could potentially contain 81 ($9 * 4 + 9 * 5$) arcs. In our graphical representation, we do not draw all these arcs, we only draw seven of them.

[1] Source code and results: github.com/xxlu/ActionEffectDiscovery.

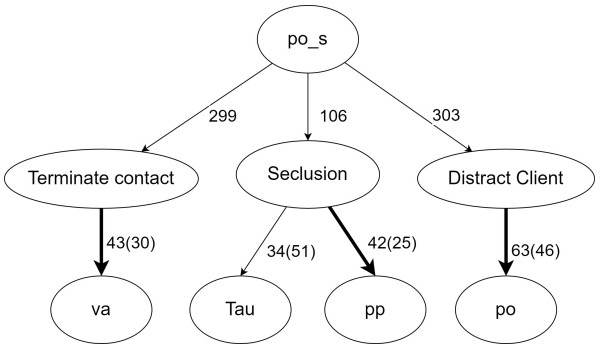


Figure 3.2: Graphical representation of applying our technique on the action-response-effect event log. The initial action is physical aggression against objects

Each arc represents a significantly higher (thick arc) or lower (thin arc) amount of observed compared to expected frequencies of interactions between the response and effect. As can be seen in the graph, this reduces the number of arcs substantially such that the impact of each individual response to a physical aggression against objects event can be studied.

Focusing on the insights we can obtain from the graphical representation in Figure 3.2. The figure shows that responding to a physical aggression against objects event with *seclusion* results in a significantly higher amount of physical aggression against people ("pp" in the figure). This can be seen by the thicker arc or by comparing the observed frequency (42) with the expected frequency (25). Studying the frequencies we can conclude that we observe that the response *seclusion* is almost 1.7 times as likely to have the effect equaling physical aggression against people compared to what is expected. In similar fashion, the response *terminate contract* and *distract client* lead to a higher likelihood of one class of effect. However, the response *seclusion* leads to a significantly lower likelihood of the effect being no next aggression incident (τ).

Comparison to Control-Flow Based Discovery

Figure 3.3, created using Disco[2], illustrates that a control-flow based discovery approach, such as the directly-follows approach, cannot provide such insights in the context of action-response-effect logs. The process model contains a large number of arcs. The number of arcs here increases exponentially with the number of responses observed. A possible solution to this could be to add information to the control-flow-based representation, such as the observed frequencies of the arcs or nodes. However, filtering based on the frequencies does not always have the desired result. This can also be seen in Figure 3.3. It could even be misleading since the data set is imbalanced. In this real-world scenario, a high frequency does not imply a significant pattern. This becomes obvious if we compare the approaches. From the figures, we can see that none of the significant response-effect pairs from Figure 3.2 are displayed in Figure 3.3. In order to understand the relations in the representation, we have

[2]Disco Tool Website: https://fluxicon.com/disco/.

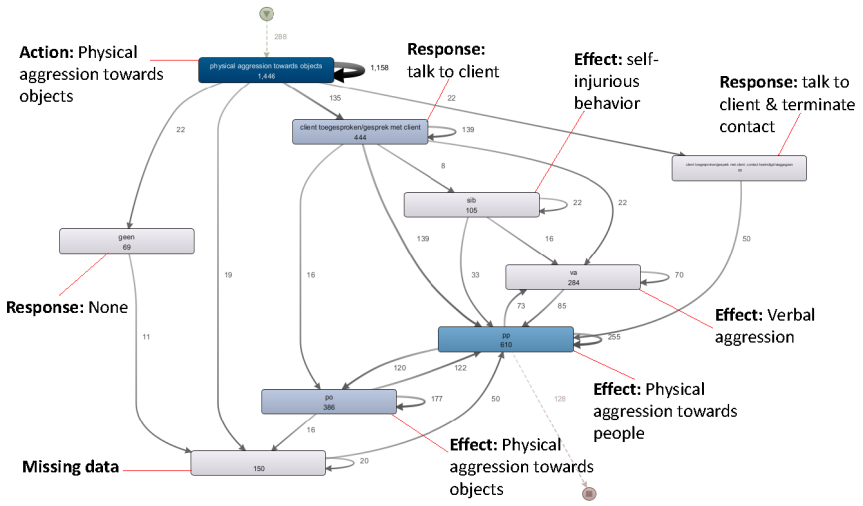


Figure 3.3: Directly-follows process model of the real-world event log for the initial action physical aggression against objects. This shows the process filtered on 5% of the possible activities and paths. This model is created using Disco

to account for the relative frequencies. These reveal the meaningful insights that are hidden in the representation of a discovery technique such as the control-flow. Hence, even after applying filtering mechanisms, Figure 3.3 does not provide the insights that are required to answer a question such as: If a client displays aggressive behavior of class X, which response is likely to lead to a (de-)escalation or future aggression?

3.5.3 Discussion

Insights

The key question identified at the start of this research addressed the desire to express insights into how a response to an action can lead to a desired or undesired outcome (effect). In our problem statement, we identified two main challenges associated with this that need to be overcome: (1) graphical representation, and (2) effective filtering mechanism. Studying the example of aggressive behavior highlights how the proposed technique addresses both of these challenges. Figure 3.2 shows that our technique creates a simple graphical representation that allows for insights into dependency relations that cannot be obtained using Figure 3.3. In addition, comparing the same figures we can see that the use of statistics reduces the number of arcs substantially. The filtering mechanism is effective in the sense that it filters those arcs that are meaningful, as opposed to those that are merely frequent.

As a result in Figure 3.2 we can see the effects of responses to aggressive incidents. Important to note is that physical aggression against people is seen as the most and verbal aggression as the least severe form of aggressive behavior. The figure shows that responding to a physical aggression against objects event with *seclusion* increases the likelihood of the next event being physical aggression against people. In other

words, this response leads to an undesired outcome: escalation of the aggressive behavior of the client. In contrast, we observe that the response *terminate contact* is more likely to lead to a verbal aggressive incident. Thus, this represents the desired outcome: de-escalation of future violence. Finally, the response *distract client* has the effect that the client is more likely to repeat the same class of action ("po" in the figure), indicating a circular relation.

Implications

One interesting implication of our technique is that the generated insights can be used to support decision-making processes. In our example, Figure 3.2 can be used to train existing and new staff members to ensure that appropriate responses are taken. Placing this technique in a broader medical context, the technique could help make informed decisions when different treatment options are considered. In a different domain, the technique could help a marketing organization understand the effectiveness of marketing strategies in terms of response to potential customers. In short, the discovery technique provides insights into action-response-effect patterns where the objective of analyzing the process is to understand possible underlying dependency patterns.

Limitations

It is worth mentioning that, in our technique, we assume the independence of the responses. This means that each response has a unique effect on the effect and there is no interfering effect when responses are combined. For example, if response r_1 is more likely to lead to c_1 and r_2 to c_2, then performing r_1 and r_2 are more likely to lead to follow-up effects c_1 or c_2, but not a different effect c_3. Statistical pre-tests can be performed to verify this assumption. A basic approach is to create a correlation matrix for the dummies of the responses. In our example, this matrix shows that the assumption holds. In other words, no responses are strongly and significantly correlated. If the assumption is violated then the technique should consider R' as input. R' is a set of all independent classes including those groups of responses that have a potential interfering effect.

3.6 Related Work

Over the last two decades, a plethora of process discovery algorithms was proposed (Augusto et al., 2018). The majority of these approaches generate procedural models such as Petri nets (Song et al., 2015; Verbeek et al., 2017), causal nets (Nguyen et al., 2017; Yahya et al., 2016), BPMN models (Augusto et al., 2017; Broucke & De Weerdt, 2017) or process trees (Buijs et al., 2012; Leemans et al., 2013). Some approaches also discover declarative models (Bernardi et al., 2014; Schönig et al., 2016) or hybrid models (i.e. a combination of procedural and declarative models) (De Smedt et al., 2015; Maggi et al., 2014). What all these techniques have in common is that they aim to discover the control flow of a business process, that is, the execution constraints among the process' activities. Our approach clearly differs from these traditional process discovery approaches by focusing on action-response patterns instead of the general control flow.

There are, however, also alternative approaches to process discovery. Most prominently, several authors addressed the problem of artifact-centric process discovery (Lu et al., 2015; Nooijen et al., 2012; Popova et al., 2015). The core idea of artifact-centric process discovery is to consider a process as a set of interacting artifacts that evolve throughout process execution. The goal of artifact-centric discovery, therefore, is to discover the lifecycles associated with these artifacts and the respective interactions among them. While artifact-centric discovery techniques move away from solely considering the control-flow of the process' activities, the main goal is still control-flow oriented. A related, yet different approach to process discovery was proposed in (Eck et al., 2016; Eck et al., 2017). This approach focuses on the different perspectives of a process and discovers and captures how their relations using Composite State Machines.

While the technique from (Eck et al., 2016; Eck et al., 2017) is potentially useful in many scenarios we address with our technique, the insights that can be obtained with our technique differ substantially. The technique from (Eck et al., 2016; Eck et al., 2017) allows us to understand how different artifact lifecycle states are related. For example, it reveals that a patient in the state "*healthy*" does no longer require a "*lab test*". The goal of our technique is to show what actually needs to be done (or should not be done) to make sure a patient ends up in the state "*healthy*". To the best of our knowledge, we are the first to propose a technique that discovers such action-response-effect patterns and allows the reader to develop an understanding of why certain events occur.

3.7 Conclusion

This chapter presents a technique to discover action-response-effect patterns. We identify two main challenges that we addressed in this research: (1) graphical representation, and (2) effective filtering mechanism. In order to address these challenges, we propose a novel discovery technique that builds on filtering influential relations using statistical tests. We evaluate our technique on a real-world data set from the healthcare domain. More specifically, we use our technique to study aggressive behavior and show that we gain valuable and novel insights from the representations discovered by our technique. The representations also show that the technique tackles both challenges by providing an easy-to-interpret representation that only displays meaningful relations.

In future work, we plan to further test the approach on real-world cases. In addition, we plan to extend this work in two ways: (1) by introducing more complex statistical tests to provide flexibility in the assumption of independence of the responses, and (2) by introducing statistical tests to approximate the optimal configuration of ϵ.

CHAPTER 4

From Action to Response to Effect: Mining Statistical Relations in Work Processes

Reading Guide. The previous chapter introduced a technique to detect, analyze, and visualize potential dependency patterns which are referred to as action-response-effect patterns. This chapter extends the proposed technique, the ARE miner, in a number of fundamental ways. First, the performance of the ARE miner is evaluated in an artificial data setting and a benchmark is made. Second, new insights are generated for the case study of aggressive behavior. Finally, the algorithm on which the technique builds is refined and parameters are fine-tuned and automatically determined. These extensions are described in this chapter and result in the state-of-the-art version of the ARE miner. This also concludes the first part of this dissertation.

This chapter is based on the following publication:

Koorn, J. J., X. Lu, H. Leopold & H. A. Reijers (2022a), "From action to response to effect: mining statistical relations in work processes", *Information Systems*, DOI: `https://doi.org/10.1016/j.is.2022.102035`.

4.1 Introduction

Process mining is a family of techniques that helps organizations to understand, analyze, and improve their work process (De Weerdt et al., 2013; Rojas et al., 2016). The basis for process mining techniques and their analyses are so-called event logs. These event logs are extracted from various information systems that are used within the organizations and, therefore, provide valuable insights into how work processes are actually executed (Van der Aalst, 2016).

Although the value of process mining has been demonstrated in various contexts, its application is still associated with a number of challenges (Rojas et al., 2016; Thiede et al., 2018). Two key aspects concern the way the process mining results are presented to the user. The representation of the results must be 1) easy to understand and 2) allow the user to obtain the required insights into the process execution. In the past, various process discovery techniques have been proposed for this purpose including the heuristic miner (Weijters & Van der Aalst, 2003), the fuzzy miner (Günther & Van der Aalst, 2007), and the inductive miner (Leemans et al., 2013). These techniques, however, all approach process discovery from a control-flow perspective, i.e., they discover ordering constraints among events.

Depending on the context, the control-flow perspective is not sufficient for understanding all relevant aspects of the process execution. Consider, for example, a care process in a residential care facility. In such a facility, clients with mental and/or physical disabilities reside and the care staff supports these clients in their daily lives. The main goal of most processes in this facility is to ensure the well-being of clients. To that end, the facility, among others, aims to minimize the aggressive behavior that is prevalent at the facility since it negatively impacts the well-being of the clients and the staff. Aggressive behavior can take many forms. For example, clients can become verbally aggressive or physically attack other clients, staff, or also themselves. When a client becomes aggressive, a staff member responds to the aggressive incident using one or multiple measures. These measures range from mild measures, such as verbal warnings, to severe measures, such as seclusion. The care facility is particularly interested in uncovering which of these measures lead to desired (i.e., de-escalation of aggressive behavior) or undesired (i.e., escalation of aggressive behavior) outcomes.

To understand and analyze such a process, we need to identify *action-response-effect* patterns. In our care process example, the aggressive incident of the client represents an *action*, the measure taken by the staff is a *response*, and the future behavior of the client is the *effect*. Existing process mining techniques cannot identify such action-response-effect patterns since their discovery requires analysis beyond the control-flow perspective. If we were to apply existing discovery techniques to an event log from such a care process, this would result in an unsatisfactory process representation for two reasons. First, the representation would be hard to read because it would contain too many connections. Second, the representation would not allow the organization to obtain the insights they require because the resulting representation would not show the effect of the behavior.

In light of these limitations of existing techniques, we propose a novel discovery technique in this work: the ARE miner. For this technique, we take advantage of well-established statistical tests to analyze event logs. The goal of this analysis is to dis-

cover and visualize understandable graphical representations of work processes. We achieve this by highlighting statistically significant and hiding the statistically insignificant relations that we discover through the statistical tests. In order to investigate the effectiveness of the technique, we evaluate it on an artificial data set and compare the results to a technique from the control-flow-perspective: the directly-follows graph. Furthermore, we demonstrate the applicability of the technique by conducting a case study in a Dutch residential care facility. We analyze a total of 21,384 aggression incidents related to 1,115 clients. Combining the insights from these two evaluations, we show that the ARE miner provides graphical representations that are 1) easy to understand, and 2) highlight informative insights.

This work is an extension of our earlier work that was published in the proceedings of the 18th International Conference on Business Process Management (Koorn et al., 2020). We extended the original research significantly in various ways. There are six main differences: 1) improved graphical representation of arcs, 2) automated determination of conceptual parameter epsilon, 3) introduction of a comprehensive quantitative evaluation, 4) extension of the qualitative evaluation, 5) revision of the related work, and 6) a thorough discussion of the limitations. We improved the *graphical representation* by including the strength of the identified statistical relations in the representation. As a further refinement of our technique, we introduced an approach to *automatically determine the parameter epsilon*, which was formerly done manually by domain experts. Now we included a data-driven approach to increase the generalizability and applicability of our technique. Besides these conceptual differences, we also extended the evaluation. We conducted a comprehensive *quantitative evaluation* based on an artificial data set to demonstrate the performance of the ARE miner in a broad spectrum of contexts. Furthermore, we extended the *qualitative evaluation* in two ways. First, we included and discussed all graphical representations from the case study to provide a more comprehensive view of the results. Second, we increased the depth by discussing additional types of relations. We revised the *related work* part by including a discussion on causal process mining techniques and by analyzing the differences and overlaps between existing techniques and the ARE miner. Finally, we expanded the *limitations* by critically reflecting on both limitations of the ARE miner itself as well as its evaluation.

The rest of the chapter is organized as follows. Section 4.2 describes and exemplifies the problem of discovering *action-response-effect* patterns. Section 4.3 introduces the formal preliminaries for our work. Section 4.4 describes the ARE miner for discovering *action-response-effect* patterns. Section 4.5 presents the evaluation of the ARE miner based on an artificial data set and a real-world event log. Section 4.6 elaborates on the insights, implications, and limitations of our work. Section 4.7 discusses related work before Section 4.8 concludes the chapter.

4.2 Problem Statement

Many processes contain action-response-effect patterns. As examples consider healthcare processes where doctors respond to medical conditions with a number of alternative treatments, service processes where service desk employees respond to issues with technical solutions, and marketing processes where customers may respond to

EID	CID	Timestamp	Action	Response
1	1	12-05 09:53	VA	Warning
2	1	13-05 13:35	PO	Distract Client, Seclusion
3	1	26-05 09:32	VA	Warning
4	1	26-05 11:02	PP	Distract Client
5	2	21-06 14:51	VA	Distract Client
6	1	23-06 21:23	VA	Distract Client
7	2	24-06 17:02	VA	-
8	3	29-08 11:22	VA	Warning
9	3	31-08 08:13	PO	Warning, Seclusion
10	3	31-08 10:48	PP	Distract Client

Legend: EID = Event identifier, CID = Client identifier, VA = Verbal Aggression, PP = Physical Aggression People), PO = Physical Aggression (Objects

Table 4.1: Excerpt from an action-response log of a care process

certain stimuli such as ad e-mails with increased demand. Let us reconsider the example of the healthcare process in a residential care facility in order to illustrate the challenge of discovering an understandable and informative process representation from an event log containing action-response relations. Of particular interest are the incidents of aggressive behavior from the clients and how these are handled by staff. Table 4.1 shows an excerpt from a respective event log. Each entry consists of:

1. an event identifier EID (which, in this case, is equal to the incident number),
2. a case identifier CID (which, in this case, is equal to the client identifier),
3. a timestamp,
4. an aggressive incident (action),
5. one or more responses to this event.

Figure 4.1 a) shows the directly–follows graph that can be derived from the events of this log. It does not suggest any clear structure of the process. Although this graph is only based on twelve events belonging to three different event classes, it seems that almost any behavior is possible. In addition, this representation does not provide any insights into certain hidden patterns (Van der Aalst, 2019). However, if we take a closer look, we can see that there are effects for a certain response. For instance, we can see that, over time, the aggressive incidents related to client 1 escalate from verbal aggression to physical aggression against objects and people. The verbal aggression event in June (EID = 6) is probably unrelated to the previous pattern since it occurs several weeks after. To gain an even deeper understanding, we need to take both the response and its effect into account. When we consider this, we see that both client 1 and 2 escalate from verbal aggression to physical aggression after the verbal aggression was only countered with a warning.

These examples illustrate that explicitly including the responses and effects in the discovery phase is important for answering the question of how to possibly respond to an action when a certain effect (e.g., de-escalating aggressive behavior) is desired. Therefore, our objective is to discover a model that: (1) shows the action-response-effect process, and (2) reveals the statistical patterns of which responses lead to a

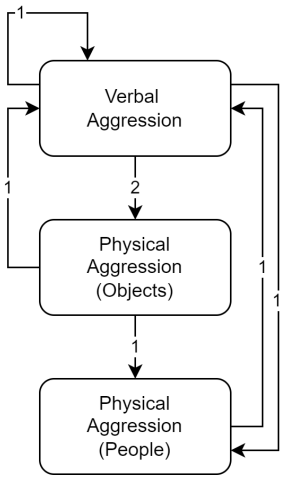


Figure 4.1: Directly-follows graph

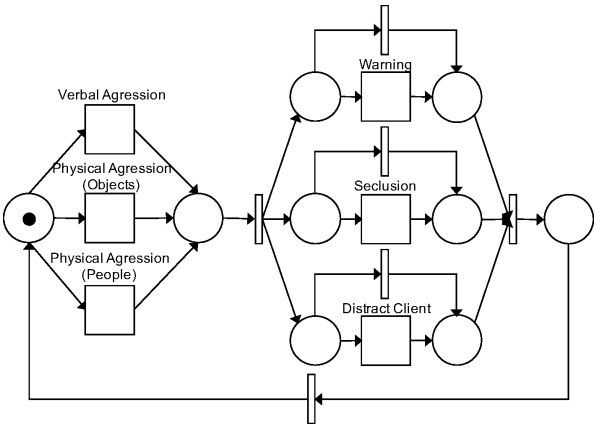


Figure 4.2: Petri Net

desired or undesired outcome (effect). There are two main challenges associated with accomplishing this:

1. *Graphical representation*: From a control-flow perspective, action-response relations are a loop consisting of a choice between all actions and a subsequent and-split that allows to execute or skip each response. Figure 4.2 b) illustrates this by showing the Petri net representing the behavior from the log in Table 4.1. Obviously, this representation does not allow understanding of which responses lead to a desired or undesired effect.
2. *Effective filtering mechanism*: The possible number of responses calls for a filtering mechanism that allows inferring meaningful insights from the model. In the example above, we only have three event classes and three event response classes (plus the “no response" class). This results in eight possible responses. In the case of 5 response event classes, we already face 32 ($=2^5$ possible responses classes). Including all these response arcs in a model will likely lead to an unreadable model that does not allow providing the desired insights.

In the next sections, we propose a novel technique, the ARE miner, that creates graphical representations of statistical patterns in event logs that contain actions, responses, and effects.

4.3 Preliminaries

As discussed, action-responses-effect patterns are observed in many processes and can provide diagnostic information regarding the follow-up effects of responses. To discover these patterns, we build on the well-established concept of event logs and introduce the concept *action-response-effect logs*. In this section, we first formalize the action-response-effect logs. We then discuss how the effects of events are defined.

4.3.1 Action-Response-Effect Logs

Starting from the event logs, we follow the definition that an *event log* is a set of sequences of events being recorded. Each *sequence* registers the execution of a case, also called a *trace*, and each *event* of the sequence represents an activity executed for the same case. Moreover, each event is associated with a set of *attributes*, which provides information such as who executed this event, when is the event executed, etc. To define action-responses-effect logs, we follow the concept of an event log and associate each event with the action (e.g., activities occurred), its response, and the effects by explicitly defining these attributes π_l, π_r, and π_{next}, respectively. We formalize action-response-effect logs as follows.

Let $\mathcal{E}$ be the universe of event identifiers. Let $\mathcal{C}$ be the universe of case identifiers. Let $d_1, ..., d_n$ be the set of attribute names (e.g., timestamp, resource, location). Let A be the set of actions and R a finite set of responses. An action-response-effect log L is defined as $L = (E, \pi_c, \pi_l, \pi_r, \pi_{next}, \pi_{d_1}, ..., \pi_{d_n}, <)$, where

- $E \subseteq \mathcal{E}$ is the set of events,
- $\pi_c : E \to \mathcal{C}$ is a surjective function linking events to cases,
- $\pi_l : E \to A$ is a surjective function linking events to actions,
- $\pi_r : E \to 2^R$ is a surjective function linking events to a set of responses,
- $\pi_{next} : E \to C$ is a surjective function linking events to the effects,
- $\pi_{d_i} : E \to \mathcal{U}$ is a surjective function linking the attribute d_i of each event to its value,
- $< \subseteq E \times E$ is a strict total ordering over the events.

Given an action-response-effect log L according to Definition 4.3.1, we shall use the shorthand notation $\sigma = \langle e_1, \ldots, e_n \rangle$ in the remainder of this chapter to refer to a trace that consists of n events with an identical case identifier. Furthermore, for any pair of events e_i and e_j with $i < j$, it holds that $e_i < e_j$ according to the strict total ordering of the events in log L.

The set of response events $\{r_1^e, \ldots, r_n^e\}$ of an event e is given by the function π_r; we write $\pi_r(e) = \{r_1^e, \ldots, r_n^e\}$. For each trace $\sigma = \langle e_1, \ldots, e_n \rangle$, the sequence of responses is $\langle \pi_r(e_1), \ldots, \pi_r(e_n) \rangle$. For example, in the action-response-effect log listed in Table 4.2, for event e_1: $\pi_c(e_1) = 1$ is the case of event e_1, $\pi_l(e_1) =$ "Verbal Aggression" (VA) is the action of e_1, and $\pi_r(e_1) = \{$"Warning"$\}$ is the set of responses of e_1.

4.3.2 Effects of Responses

As we discussed, we aim to investigate whether a certain response to an action has an effect on the follow-up event. As such, we measure the effectiveness of a response to an action by studying the effect. For this aim, we first define the effects of events by using the function π_{next} and introduce parameter ϵ for elapsed time. For each trace $\sigma = \langle e_1, ..., e_n \rangle$, we define the effect for each e_i, where $1 \leq i < n$ as follows: if the elapsed time to the next event e_{i+1} is less than ϵ, the effect $\pi_{next}(e_i)$ of e_i is the action of e_{i+1}, else we say that the effect is a silent action τ. Formally, if $\pi_{time}(e_{i+1}) - \pi_{time}(e_i) \leq \epsilon$, then $\pi_{next}(e_i) := \pi_l(e_{i+1})$, else $\pi_{next}(e_i) := \tau$.

To test the hypothesis of whether an effect is independent of the response to an action, the number of observed events is compared to the number of expected events of different responses and effects. To calculate the number of observed events, we

ID	Timestamp	Action	Response	Effect
1	12-05 09:53	VA	Warning	PO
1	13-05 13:35	PO	Distract Client, Seclusion	τ
1	26-05 09:32	VA	Warning	PP
1	26-05 11:02	PP	Distract Client	τ
2	21-06 14:51	VA	Distract Client	VA
1	23-06 21:23	VA	Distract Client	τ
2	24-06 17:02	VA	-	τ
3	29-07 11:22	VA	Warning	PO
3	31-07 08:13	PO	Warning, Seclusion	PP
3	31-07 10:48	PP	Distract Client	τ

Legend: VA = Verbal Aggression, PP = Physical Aggression (People), PO = Physical Aggression (Objects)

Table 4.2: Excerpt of the event log action-response-effect

create a matrix (table) where each cell is filled with the number of observed events of a response and an effect. Let $a \in A$ be an action, $R = \{r_1, \cdots, r_m\}$ be a set of responses, and $C = \{c_1, \cdots, c_n\}$ a set of effects. We define a $|R| \times |C|$ matrix, where each row represents a response r_i, each column represents an effect c_j, and each cell counts the number of observed events that have response r_i and effect c_j. We have

$$freq_{a,R,C} = \begin{pmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,n} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{m,1} & f_{m,2} & \cdots & f_{m,n} \end{pmatrix}$$

where

$$f_{i,j} = freq_L(a, r_i, c_j) = \quad |\{e \in L \mid \pi_l(e) = a \wedge r_i \in \pi_r(e) \wedge \pi_{next}(e) = c_j\}| \tag{4.1}$$

For instance, given a log L as listed in Table 4.2, $freq_L$("VA","Warning","PO") =$|\{e_1, e_8\}|= 2$. Considering Table 4.3 and omitting the column totals and row totals, it exemplifies a matrix $freq_{a,R,C}$. If the effects are independent of responses, then we should observe that the distribution of effects of a response is similar to the *total distribution*.

Each row r_i presents the distribution of effects $c_1, ..., c_k$ to the response r_i. To test whether each individual response r_i has an influence on the effects, we define $freq_{a,r,C}$ as a 2× |C| matrix:

$$freq_{a,r,C} = \begin{pmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,n} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,n} \end{pmatrix} \tag{4.2}$$

where $f_{1,j} = freq_L(a, r, c_j)$ and $f_{2,j} = \quad |\{e \in L \mid \pi_l(e) = a \wedge r \notin \pi_r(e) \wedge \pi_{next}(e) = c_j\}|$.

An example of $freq_{a,r,C}$ where r is "Terminate contact" is listed in Table 4.4. In the following section, we describe that in our ARE miner first a chi-squared test is carried out. This allows us to calculate the expected values and test the statistical dependency between responses and effects. The chi-square test compares the observed frequencies to the expected frequencies. If they differ significantly, then the null hypothesis

Observed	**PO**	**PP**	**VA**	τ	*Total*
Warning	250	400	200	50	*900*
Held with force	20	50	50	10	*130*
Seclusion	30	50	20	10	*110*
Terminate contact	100	100	90	10	*300*
Distract client	100	150	40	20	*310*
Total	*500*	*750*	*400*	*100*	**1750**
Expected	**PO**	**PP**	**VA**	τ	*Total*
Warning	257.1	385.7	205.7	51.4	*900*
Held with force	37.1	55.7	29.7	7.4	*130*
Seclusion	31.4	47.1	25.1	6.3	*110*
Terminate contact	85.7	128.6	68.6	17.1	*300*
Distract client	88.6	132.9	70.9	17.7	*310*
Total	*500*	*750*	*400*	*100*	**1750**

Legend: VA = Verbal Aggression, PP = Physical Aggression (People), PO = Physical Aggression (Objects)

Table 4.3: Excerpt of the tables used to perform high-level statistical tests; horizontal categories: effect, vertical categories: response

is rejected, which means we cannot rule out that there is a statistical dependency relation between the response and the effect.

The complete event logs containing action-response-effect are used in the ARE miner that is proposed in this chapter. The next section elaborates on this.

4.4 ARE Miner

Based on the formalization introduced in the previous section, we use this section to propose the ARE miner as a novel discovery technique. The goal of the ARE miner is to generate understandable process representations that provide the user with the required insights into the execution of the process. First, we describe the required pre-

Observed	**PO**	**PP**	**VA**	τ	*Total*
Terminate contact = 0	300	500	210	90	*1100*
Terminate contact = 1	100	100	90	10	*300*
Total	*400*	*600*	*300*	*100*	*1400*
Expected	**PO**	**PP**	**VA**	τ	*Total*
Terminate contact = 0	314.3	471.4	235.7	78.6	*1100*
Terminate contact = 1	85.7	128.6	64.3	21.4	*300*
Total	*400*	*600*	*300*	*100*	*1400*

Legend: VA = Verbal Aggression, PP = Physical Aggression (People), PO = Physical Aggression (Objects)

Table 4.4: Excerpt of the tables for an individual response used to perform statistical tests; horizontal categories: effect

Algorithm 2 Compute graph

Input: Event log L
Output: Graph $G = (V, \prec)$
1: {***STAGE 1: High-Level Statistics***}
2: **for** $a \in A$ **do**
3: Initiate matrix $O[a] \leftarrow freq_{a,R,C}$ {*see Equation 4.2, calculate the observed values*}
4: Compute matrix $E[a]$ {*calculate the expected values by following the chi-square test, see (Cochran, 1952)*}
5: Compute $\chi^2_a = \frac{(O[a]-E[a])^2}{E[a]}$ {*To test the dependence between responses R and effects* $A \cup \{\tau\}$}
6: **if** χ^2_a is significant **then**
7: {$O[a]$ *differs from* $E[a]$*, thus responses R have a statistically significant influence on the effects C*}
8: {***STAGE 2: Detailed Statistics***}
9: **for** response $r \in R$ **do**
10: Compute matrix $O[a]_r$, $E[a]_r$, and $\chi^2_{a,r}$
11: **if** $\chi^2_{a,r}$ is significant **then**
12: {***STAGE 3: Influential Points***}
13: Compute adjusted standardized residuals ASR_c {*see Section 4.4.4*}
14: **for** effects $c \in A \cup \{\tau\}$ **do**
15: **if** ASR_c is significant **then**
16: {*draw the arc from r to c*}
17: $V \leftarrow V \cup \{a_s\} \cup \{r\}$, $\prec \leftarrow \prec \cup \{(a_s, r)\} \cup \{(r, c)\}$
18: **end if**
19: **end for** {*effect*}
20: **else**
21: {$\chi^2_{a,r}$ *is insignificant, i.e., r has no significant influence on C. We do not draw node r or any arc from r to C*}
22: **end if**
23: **end for** {*response*}
24: **else**
25: {*Observed* $O[a]$ *follows the expected values* $E[a]$*, thus response R has no statistically significant influence on the effects C; thus, no arcs are drawn*}
26: **end if**
27: **end for** {*action*}
28: **return** G

processing steps. Then, we elaborate on the conceptual approach of the ARE miner, which consists of three main stages: 1) computing high-level statistics, 2) computing detailed statistics, and 3) identifying influential points.

4.4.1 Pre-processing the Event Log

We first pre-process the log to obtain the effects of responses. Since we are studying the effects of a response to an action, the duration between a response and its effects influences the likelihood of a statistical relationship between the two. Let us return to our example: if there is an aggressive incident, there is a given response to this incident. However, if the next incident takes place after a long time (e.g., a year) it is unlikely that this new incident is still dependent on the response to the initial action. Thus, we use the parameter epsilon (ϵ), see Section 4.3.2. ϵ represents the maximum duration between two events in which the first event is still considered to have an effect on the second event. We can define ϵ in two ways: (1) based on the data or (2) based on the knowledge of a domain expert. If we base the ϵ on the data, we define it as equaling the average duration of the events. To ensure that outliers do not influence the average, depending on the distribution of the data, a number of actions can be taken. For our specific example, the data is exponentially distributed. To account for this, we select 80% of the data when sorted on duration and take the mean of this subset of data. This results in an ϵ of 8.9 days. We round this up to full

days ($\epsilon = 9$) due to the granularity of our data. Other distributions in the data are likely to occur as well. One common example is that the data is normally distributed. In this case, we propose to define ϵ equaling the mean plus two standard deviations to the right (longest duration). This ensures that a subset of the data closest to the mean is captured.

We can also base the ϵ on the input of a domain expert. Consider the healthcare organization in our qualitative evaluation, see Section 4.5.2. We consulted with a behavioral expert in the organization. The domain expert from the organization defined ϵ as equaling seven days. This is what the domain expert indicates is the likely maximum duration between two events (aggressive incidents) where the response of the first incident still has an effect on the behavior of the client in the second incident. Based on ϵ, we introduce state τ. It represents the state where no next incident occurs within the defined duration of ϵ. In Table 4.2 we can see, for example, that distracting the client seems to be related to τ.

Another step in the pre-processing phase is to check the statistical assumptions. The Chi-square test, which we will elaborate on in the next sections, has a number of assumptions (McHugh, 2013): 1) the data should be frequencies or counts, 2) the categories of the variables are mutually exclusive, 3) each subject may only contribute to one cell (no repeated measures), 4) both variables are measured as categories, 5) the sample size should be sufficiently large. The first four assumptions are data requirements that need to be checked by the analyst that applies the ARE miner. The fifth assumption can be automatically checked by the ARE miner itself.

With regard to the fifth assumption (i.e., sufficient sample size), we implement a heuristic selection criterion: the value of the expected cells in each table should be 5 or greater in at least 80% of the cells (McHugh, 2013). If this criterion is not met, the action-response or response-effects pair is excluded from the analysis and an NA value is the output.

4.4.2 Stage 1: Computing High-Level Statistics

After pre-processing the event log, we investigate for each action the significant relationship between the responses and the effects. In our example, the client shows a certain type of aggressive behavior (the action). Given this, we are interested in how the response of a caretaker to that incident has an impact on the follow-up incident (effect). Hence, we will explain the ARE miner with a fixed initial action. The details are formalized in Algorithm 2. In the following, we will explain how the specific steps from Algorithm 2 are linked to the conceptual considerations.

In Table 4.3, we show an example of the observed and calculated expected frequencies for the action of physical aggression against objects (lines 3 & 4 in Algorithm 2). This allows us to perform a Chi-square test (Cochran, 1952) (line 5). Based on a confidence level α (usually 95%), the calculated Chi-square (χ^2) test value is compared to the Chi-square distribution to see if there is at least one pair of response-effect significantly different. If the Chi-square test value is insignificant, the action is excluded from the graphical representation (line 21). If the Chi-square is significant (line 6), this indicates that the effects may depend on the response. We then move to the second stage, see Section 4.4.3.

We demonstrate this first stage of the ARE miner by applying it to a designed

example based on our real-world data set presented in Table 4.3. Based on the observed values, we can calculate the expected values in the table, for example, the expected value for the first cell: response *Terminate Contact* and effects *VA* = $\frac{N_r \times N_c}{E[a][r][c]} = \frac{300 \times 400}{1750} = 68.6$. We know from Table 4.3 that there are five response classes and four effects classes, so the degrees of freedom: $c = (5-1) \times (4-1) = 12$. Given all this, we can calculate the Chi-square test value for the overall table:

$$\chi^2_c = \sum_{i=1}^{5} \frac{(O_{\text{Warning,PO}} - E_{\text{Warning,PO}})^2}{E_{\text{Warning,PO}}} + ... + \frac{(O_{\text{Distract client},\tau} - E_{\text{Distract client},\tau})^2}{E_{\text{Distract client},\tau}}$$

$$\chi^2_{12} = \frac{(250-257.1)^2}{257.1} + ... + \frac{(10-17.7)^2}{17.7} = 63.47$$

Now we need to determine if this value is significantly different from the mean of the Chi-distribution (Fisher & Yates, 1938). The formula for calculating the p-value is complex and will thus not be discussed in detail in this chapter. For more details, we refer to (Fisher & Yates, 1938). In the above example, the p-value (< 0.001) shows that our Chi-square value is significant. This indicates that for at least one pair of response-effects given action PO there is a significant difference from the expected frequency. Thus, we perform a Chi-square test for each individual response.

4.4.3 Stage 2: Computing Detailed Statistics

In the second stage of the ARE miner, we perform the Chi-square test again on each response class to determine for which response we need to perform post-hoc statistical tests (lines 8 - 10 in Algorithm 2). For this purpose, we create dummy variables. A dummy variable is made for each individual response, which takes the value of 0 or 1. The new table we create is a 2 x 4 table where the rows represent the response either taking a 0 or 1 value, see Table 4.4. Note that the degrees of freedom change to three. The same formulas are used to calculate the individual response Chi-square score and the corresponding p-value. A Bonferroni correction (Haynes, 2013) is made to correct the critical value for the fact that on the same table multiple sets of analyses are performed. The Chi-square test identifies for which responses there is at least one effect that is significantly different from the expected frequency. If the Chi-square value is significant, we create a node for the response and perform post-hoc tests to identify the exact pairs of response-effects that are significant (line 11).

We will demonstrate this stage in an example case. We test five times (one for each response). Thus, we apply the Bonferroni correction (Haynes, 2013) on a confidence level of 95% (meaning $\alpha = 0.05$): $\frac{0.05}{5} = 0.01$. If we take Table 4.4, we can use the same formulas as presented in the previous section to calculate the expected values. Note that we assume the independence of responses. Thus, if there are two responses, the action is counted twice: once for response 1 and once for response 2. Therefore, the observed frequencies in Table 4.3 are not necessarily equal to those in Table 4.4. If we perform the Chi-square test for the response *Terminate contact* we get a Chi-square score of 31.96 with a p-value < 0.001. Thus, for the response *Terminate contact* there is at least one effect that is significantly different from the expected frequency. In the next section, we describe how a post-hoc test will need to identify the exact pairs for which this is true.

4.4.4 Stage 3: Identifying Influential Points

In the last stage, the post-hoc tests are performed to test which exact pairs of response-effects have a significant contribution to the Chi-square test value. To do so, the adjusted standardized residuals (ASR) (Agresti, 2003) are calculated (line 13 in Algorithm 2). They represent a normalization of the residuals (observed - expected frequency). As the residuals can take either a positive or negative value we use two-sided testing. In order to improve the interpretability, we transform the α level into a critical value. We refer to (Fisher & Yates, 1938) for details on this approach. If $|ASR| > criticalscore$ the difference between observed and expected frequency is significant. A significant score means that a specific pair of response-effects has a significant impact on the overall test value. We refer to this as an *influential point*.

For each influential point, arcs are drawn in the graphical representation (lines 14-17). If the score is insignificant, no arc is drawn for that pair of response-effects. We first draw an arc from the action to the responses. Then, we draw an arc from the response to the effect for which we found a significant relation. If the observed frequency is larger than the expected frequency, i.e., the response leads to an increase in the frequency of effects, we draw a thick arc. Correspondingly, if the observed frequency is lower than the expected frequency we draw a thin arc. The total number of graphical representations created equals the number of actions for which a significant Chi-square score is found (line 25).

Now, we turn to the example from Table 4.4. From the previous section, we know that the response *Terminate contact* results in a significant Chi-square score. To calculate which points are influential points we calculate the adjusted standardized residuals for each pair. To exemplify, we show the calculation of the ASR for the pair *Terminate contact = 1* and *VA*:

$$ASR = \frac{90 - 64.3}{\sqrt{64.3 * (1 - \frac{64.3}{300}) * (1 - \frac{64.3}{300})}} = 4.08$$

Given our Bonferroni correction gave us an alpha of 0.01 (see the previous section), we need to test on the 99 % confidence level. The critical absolute value for this is 2.57. Thus, if our ASR value is > |2.57|, we mark it as an influential point and draw an arc in the graphical representation. In the example of the pair *Terminate contact* = 1 and VA the ASR is larger than the critical score (4.08 > 2.57). Therefore, we draw an arc in the graphical representation of this example.

To increase the readability of the graph, we use a variety of arcs. First, on the arc from an action to a response, we indicate the observed frequency of the behavior. This shows how often this specific action-response pattern is observed. Next, on the arc from a response to an effect, we display the observed frequency and the expected frequency in brackets. This shows whether or not the response leads to an increase or decrease in the behavior type of the effects. To also display the strength of the relation between the response and the effects in the graph, we adjust the thickness of the arc based on the adjusted standardized residuals. Recall that this value needs to be > |2.57| in order for an arc to be drawn. We introduce six classes of effect strength, i.e. three positive classes and three negative classes. We choose to use a total of six

classes to differentiate between the strength of the effect as this is a number that is easily comprehensible, yet allows for sufficient distinctions between effect sizes.

We determine the classes by identifying the maximum and minimum ASR scores for each action table. Subsequently, we create the range of ASR scores for each class by dividing the scores between the maximum ASR score and 2.57 equally into three classes. The same structure applies to the negative scores, but then we use the difference between the minimum ASR score and -2.57. As an example, assume the maximum ASR value is 8.30, and the minimum ASR value is -4.37. The three positive classes will be, from least thick to thickest; (1) [2.57:4.48], (2) [4.49:6.39], and (3) [6.40:8.30]. In line, the three negative classes will be, from least thick to thickest: (1) [-2.57:-3.17], (2) [-3.18:-3.77], and (3) [-3.78:-4.37].

After applying the last stage of the ARE miner to all actions, responses, and effects on the aforementioned example, we obtain a total of three graphical representations (one for each action). In the next section, we evaluate the ARE miner both on an artificial as well as a real-world data set to demonstrate that it indeed can generate understandable process representations that provide the user with the required insights into the process execution.

4.5 Evaluation

The goal of this section is to demonstrate the effectiveness of the ARE miner to discover models that allow obtaining meaningful insights into action-response-effect patterns. To this end, we implemented the ARE miner in Python[1] and conducted a quantitative as well as a qualitative evaluation. In the quantitative evaluation (Section 4.5.1), we use an artificial data set to systemically explore in a large range of constellations how the representations produced by the ARE miner compare to traditional process-oriented representations. In the qualitative evaluation (Section 4.5.2), we apply the ARE miner to a real-world action-response-effect log and investigate to what extent the discovered models are meaningful from a domain perspective.

4.5.1 Quantitative Evaluation

This section discusses the quantitative evaluation of the ARE miner. Our goal is to develop an understanding of how the representations produced by the ARE miner compare to directly-follows graphs, i.e., traditional process-oriented representations, in a variety of different settings. To this end, we generate an artificial data set that represents a broad spectrum of real-life scenarios. Using this set, we can systematically explore under differing circumstances how key characteristics, such as the number of arcs, develop. In Sections 4.5.1 and 4.5.1, we first elaborate on the artificial data set generation and the setup. Then, in Section 4.5.1, we present the results.

Data Set Generation

To obtain artificial data representing a broad spectrum of possible real-life situations, we generate a set different of frequency tables (see Table 4.3 for an example). The main rationale behind this approach is that frequency tables summarize all relevant

[1] Source code and results: github.com/xxlu/ActionEffectDiscovery

characteristics from the log that we build on in the context of the ARE miner. Hence, generating frequency tables instead of actual action-response-effect logs allows us to precisely control these characteristics.

To illustrate the details of the frequency table generation, recall that a frequency table captures the number of times a response $r \in R$ leads to an effect $e \in E$, given an action $a \in A$. For example, the first row from Table 4.3 describes how many times we observe a particular effect (i.e., *PO*, *PP*, *VA*, or τ) for the response *Warning*. Taking a look at the numbers, we see, for example, that the response *Warning* leads to *PO* in 250 cases and to *PP* in 400 cases. Intuitively, these absolute numbers can also be converted into probabilities. Given the total of 900 observations for the response *Warning*, we can determine that the probability of a *warning* leading to *PO* is approximately 0.28 (250/900). If we determine the probabilities for the other effects as well, we obtain the probability vector (0.28, 0.44, 0.22, 0.06). In statistical terms, this probability vector represents the probability mass function (PMF) of the underlying discrete distribution that we observe in the first row of Table 4.3. Since such a probability vector can be computed for every row, the frequency table can be described by using m probability vectors $(v_1, \ldots, v_m)$, where m is equal to the number of rows and, therefore, also to the number of responses.

To generate an artificial data set, we build on these probability vectors representing PMFs. The advantage of doing so is that they allow us to systemically capture a large range of possible real-life constellations. Intuitively, there are two extreme scenarios for an action-response-effect pattern. The first is if a considered response is only leading to a single effect as, for instance, described by the probability vector (1.0, 0.0, 0.0, 0.0). The second is if the likelihood for all effects is the same, as, for instance, described by the probability vector (0.25, 0.25, 0.25, 0.25). Besides these two extremes, there is an infinite number of alternative probability vectors for a given a number of n effects. Therefore, we introduce the parameter δ. By requiring that each probability p_i that is part of a probability vector $v = (p_i, \ldots, p_n)$ is a *multiple* of δ, we guarantee that the number of possible probability vectors is finite. Keep in mind that $\sum_{i=1}^{n} p_i = 1$ because we are dealing with vectors representing PMFs. Based on these considerations, we compute the set V of all possible probability vectors for a given number of effects n and the parameter δ. Given a number of m responses, the total set of possible frequency tables F is then given by the n-ary Cartesian power of V, i.e., $F = V^m = \{(v_1, \ldots, v_m) \mid v_i \in V \; for \; every \; i \in \{1, \ldots, m\}\}$. We use $f = (v_1, \ldots, v_m)$ to refer to an individual constellation from F.

To characterize the potentially large number of possible frequency table constellations, we introduce the complexity indicator *average standard deviation* (ASD). The ASD is the arithmetic mean of the standard deviations of the individual columns of a frequency table $f \in F$. As such, it quantifies to what extent the probabilities of different responses leading to the same effect differ from each other. The closer ASD is to zero, the smaller the differences across the responses. The closer ASD is to the maximum possible average standard deviation for f, the higher the differences across the responses. Note that this maximum value for ASD depends on the number of responses. If, for instance, $m = 4$, then the maximum ASD is 0.5. Figure 4.5 shows three possible frequency tables and their ASDs resulting from a generation run with n=4, $m = 4$, and $\delta = 0.05$. The left and the right tables show two extremes with

Response	Effect					Effect					Effect			
	n_1	n_2	n_3	n_4		n_1	n_2	n_3	n_4		n_1	n_2	n_3	n_4
m_1	0.25	0.25	0.25	0.25	m_1	0.80	0.00	0.00	0.20	m_1	1.00	0.00	0.00	0.00
m_2	0.25	0.25	0.25	0.25	m_2	0.25	0.25	0.25	0.25	m_2	0.00	1.00	0.00	0.00
m_3	0.25	0.25	0.25	0.25	m_3	0.00	0.00	1.00	0.00	m_3	0.00	0.00	1.00	0.00
m_4	0.25	0.25	0.25	0.25	m_4	0.30	0.10	0.30	0.30	m_4	0.00	0.00	0.00	1.00
	0.00	0.00	0.00	0.00		0.34	0.12	0.43	0.13		0.50	0.50	0.50	0.50
	ASD = 0.00					**ASD = 0.25**					**ASD = 0.50**			

Table 4.5: Three possible frequency tables with n=4, $m = 4$, and $\delta = 0.05$

the ASD being 0.00 and 0.50. The example in the center shows a rather mixed case. These three examples highlight the broad spectrum of frequency table constellations that may arise in practice and we, therefore, need to systematically consider them.

Based on the approach introduced above, we generated an artificial data set with $m = 4$, $n = 3$, $\delta = 0.2$. Note that higher values for parameters m and n as well as a smaller value for δ mainly affect the granularity of the results but not the results themselves. Therefore, we selected parameters that balance granularity and computational effort. In total, this choice of parameters results in a set F containing 175.616 different frequency tables. We will refer to these as *constellations*. To make sure that the generated constellations also include patterns with a low frequency (i.e., smaller than δ), we randomly add and subtract values of 0.01 up to 5 times per row in each f. Figure 4.3 visualizes the distribution of the resulting constellations with respect to the ASD. We can see that it represents a discrete representation of the Bell curve where constellations with an ASD of between 0.20 and 0.25 are most likely.

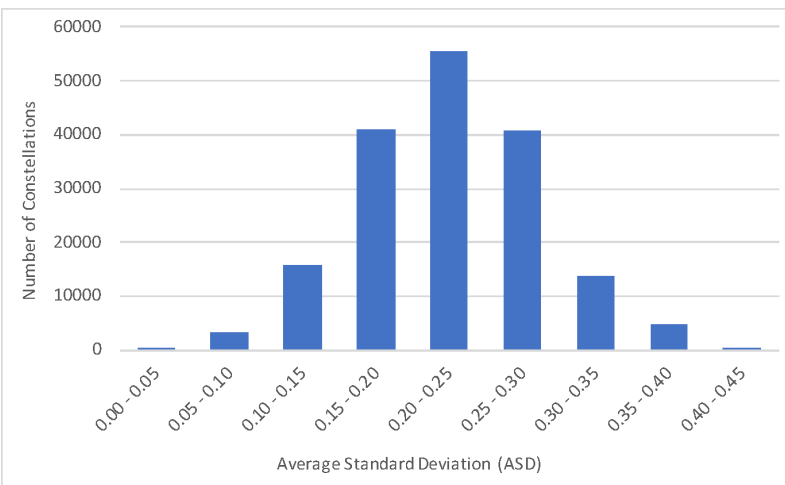


Figure 4.3: Distribution of constellations with respect to the ASD

Setup

In our evaluation experiments, we compare three different techniques:

- *ARE miner*: We implemented the ARE miner in Python[2] as described in Section 4.4. Our implementation also automatically performs all assumption checks. Note our implementation may return a graph without any edges in case no significant edges could be identified.
- *Naive DFG*: As a first baseline, we use a naive directly-follows graph implementation. This configuration returns an arc for every observed response-effect pattern, i.e., for every value in $f \in F$ that is above 0.
- *Filtered DFG*: As a second baseline, we use a filtered directly-follows graph implementation. This configuration returns an arc for an observed response-effect pattern if the relative frequency of that pattern with respect to the most frequent pattern from the considered constellation $f \in F$ is above the threshold τ. For the purpose of our experiments, we set $\tau = 0.8$. This means that if in a constellation f the most frequent pattern occurs with a probability of 0.5, then every pattern with a probability of less than 0.1 will be removed.

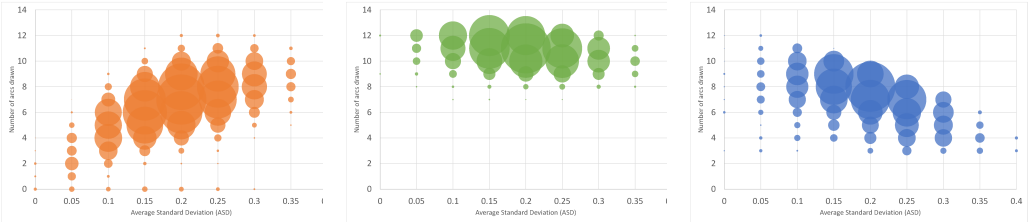
For each of the configurations above, we compute which arcs they generate for each $f \in F$. To quantify the results, we determine 1) the number of arcs generated for each f, and 2) the fraction of significant arcs generated for each f. The first performance measure allows us to understand to what extent the *total* number of arcs produced by the ARE miner compare to the two baselines. On the one hand, we expect to reduce the overall complexity and, therefore, generate fewer arcs. On the other hand, we want to demonstrate the applicability of the ARE miner, which means that we want to demonstrate that the ARE miner does not generate zero arcs or a single arc in the majority of constellations. The second performance measure helps us to understand how many of the *significant* arcs generated by the ARE miner is also covered by the two baselines.

Results

Below, we present the results from the quantitative evaluation. We first analyze *the number of arcs* generated by each technique. Then, we take a detailed look at the *fraction of significant arcs*.

Number of arcs. The results of our evaluation experiments with respect to the number of arcs are visualized in the bubble charts in Figure 4.4. The charts show how often a certain number of arcs (y-axis) were generated for a set of constellations from F with a particular ASD (x-axis). A first glance reveals that the representations produced by the ARE miner differ considerably from the respective DFGs. Most notably, on average, the number of arcs generated by the ARE miner is much lower than the number of arcs generated by the naive DFG-based approach. As for the number of arcs, the ARE miner seems to produce, on average, around the same number of arcs as the filtered DFG-based approach. Considering the ASD values, we see that the naive DFG-based approach produces a roughly equal number of arcs over the range of ASD values. If we take a closer look at the ARE miner, we see that it draws fewer arcs when the ASD is low and more when the ASD is high. We follow up on this observation below. Comparing this to the filtered DFG-based approach, we see that

[2]The code is publicly available for reproducibility: `github.com/xxlu/ActionEffectDiscovery`.



(a) ARE miner (b) Naive DFG (c) Filtered DFG

Figure 4.4: Overview of number of generated arcs for each $f \in F$

the DFG-based approach seems to do the opposite. It generates more arcs when the ASD is low and fewer when it is high.

Table 4.6 provides a detailed view on the results. It shows that the average number of arcs produced by the ARE miner in comparison to the naive DFG-based approach is about 4.10 lower (6.83 versus 10.93) and about 0.44 lower than the filtered DFG-based approach (6.83 versus 7.27). Note, however, that this number must be considered in the context of the chosen n and m. Since the maximum number of possible arcs is 12, filtering an average of 4 arcs has a notable effect on the resulting representations. If we take a closer look at the numbers, we can see that, as the ASD increases, the ARE miner draws more arcs and the filtered DFG-based approach draws less. The lower the ASD, the more arcs they contain. This suggests that, while the average number of arcs is almost equal, the filtered DFG-based representations may contain arcs that the ARE miner decided to suppress and tends to miss those the ARE miner decided to draw. This is caused by the fact that the notion of statistical significance is a relative consideration and not based on absolute numbers. Realizing this, the next section looks in more detail into which arcs each technique includes in the respective representations.

Fraction of significant arcs. To understand how the considered configurations differ on a semantic level, it is helpful to analyze which arcs the generated representations have in common. Building on the premise that statistically significant arcs provide the insights we are looking for from a semantic point of view, we are therefore interested in the fraction of significant arcs produced by the filtered DFG-based approach. Note that a comparison with the naive DFG-based approach is obsolete since the naive DFG-based approach will contain all possible and, therefore, also all significant arcs. Figure 4.5 visualizes the number of shared and non-shared arcs for both the filtered DFG-based approach and the ARE miner. More specifically, it shows how many arcs, on average, are produced for the different constellations.

For low values of ASD, we can see that both the number of arcs generated by the ARE miner and the number of shared arcs are very low. However, the number of arcs produced by the filtered DFG-based approach is quite high for low ASD values. Even in the lowest bin, from 0 to 0.05, the filtered DFG-based approach generates an average of 7.1 arcs. With an increasing ASD also the number of shared arcs increases. In general, this is in line with our expectations. The closer we get to an ASD of 0, the more equal the distribution of the data is. Hence, the ARE miner will identify only a few significant arcs, if any. The closer we get to the maximum ASD, the more

		Number of arcs														
	ASD	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**	**Avg.**
ARE miner with noise	0.00-0.05	58	20	36	6	2	0	0	0	0	0	0	0	0	122	0.97
	0.05-0.10	177	417	1129	588	642	179	23	1	0	0	0	0	0	3156	2.55
	0.10-0.15	126	45	459	2145	4829	4680	4577	1194	306	34	1	0	0	18396	4.82
	0.15-0.20	180	0	67	730	3439	8491	12335	11101	5460	1609	262	18	0	43692	6.21
	0.20-0.25	138	0	13	203	1415	5075	11257	15767	13553	6697	2241	318	59	56736	7.14
	0.25-0.30	84	0	8	34	319	1433	4458	8997	10969	7778	3132	688	80	37980	7.79
	0.30-0.35	18	0	0	0	11	137	733	2250	4010	3883	1951	440	43	13476	8.37
	0.35-0.40	0	0	0	0	0	10	44	197	516	644	406	73	0	1890	8.72
	0.40-0.45	0	0	0	0	0	0	6	16	41	65	40	0	0	168	8.70
	Total	*781*	*482*	*1712*	*3706*	*10657*	*20005*	*33433*	*39523*	*34855*	*20710*	*8033*	*1537*	*182*	*175616*	*6.83*
DFG with noise 100%	0.00-0.05	0	0	0	0	0	0	1	0	0	18	18	18	67	122	11.07
	0.05-0.10	0	0	0	0	0	0	0	0	18	108	558	852	1620	3156	11.25
	0.10-0.15	0	0	0	0	0	0	0	15	57	612	2724	6588	8400	18396	11.23
	0.15-0.20	0	0	0	0	0	0	0	21	156	1536	6927	17412	17640	43692	11.16
	0.20-0.25	0	0	0	0	0	0	0	9	288	2454	11631	26340	16014	56736	10.97
	0.25-0.30	0	0	0	0	0	0	0	9	342	3084	11859	16842	5844	37980	10.65
	0.30-0.35	0	0	0	0	0	0	0	0	174	1494	5340	5418	1050	13476	10.42
	0.35-0.40	0	0	0	0	0	0	0	0	42	432	942	456	18	1890	9.99
	0.40-0.45	0	0	0	0	0	0	0	0	6	90	72	0	0	168	9.39
	Total	*0*	*0*	*0*	*0*	*0*	*0*	*1*	*54*	*1083*	*9828*	*40071*	*73926*	*50653*	*175616*	*10.93*
DFG with noise 80%	0.00-0.05	0	0	0	10	0	0	54	9	9	30	3	3	4	122	7.13
	0.05-0.10	0	0	0	96	99	15	252	603	735	546	522	228	60	3156	8.15
	0.10-0.15	0	0	0	24	360	354	792	2865	4791	4938	3357	915	0	18396	8.40
	0.15-0.20	0	0	0	0	534	1245	2181	7440	12855	15255	4176	6	0	43692	8.14
	0.20-0.25	0	0	0	306	1404	2766	5877	15018	24468	6891	6	0	0	56736	7.38
	0.25-0.30	0	0	0	804	3057	5532	7884	14937	5766	0	0	0	0	37980	6.33
	0.30-0.35	0	0	0	354	2946	3498	4158	2520	0	0	0	0	0	13476	5.41
	0.35-0.40	0	0	0	372	870	510	138	0	0	0	0	0	0	1890	4.22
	0.40-0.45	0	0	0	78	90	0	0	0	0	0	0	0	0	168	3.54
	Total	*0*	*0*	*0*	*2044*	*9360*	*13920*	*21336*	*43392*	*48624*	*27660*	*8064*	*1152*	*64*	*175616*	*7.27*

Table 4.6: Number of arcs generated by each technique

we face a random distribution. In such a setting, the ARE miner is more likely to identify significant arcs and, therefore, generates an increasing number of arcs. Since the number of arcs produced by the filtered DFG-based approach is relatively stable, the number of shared arcs also increases when we move to the high end of the ASD value.

From a semantic perspective, Figure 4.5 highlights the importance of building on the statistical notion we use in the ARE miner. The filtered DFG-based approach generates a relatively stable number of arcs across all constellations although the number of statistically significant and, therefore, meaningful arcs differs considerably. Constellations with a very low ASD simply do not provide evidence that there are many meaningful patterns to detect. This, however, cannot be captured by filtering arcs based on frequency. It requires the statistical perspective exploited by the ARE miner.

In summary, the quantitative evaluation illustrates that the ARE miner performs well in a broad range of possible situations. We showed that 1) the ARE miner leads to a notable reduction in the number of arcs compared to the naive DFG-based approach, and 2) the ARE miner produces a different, and more meaningful, set of arcs than the

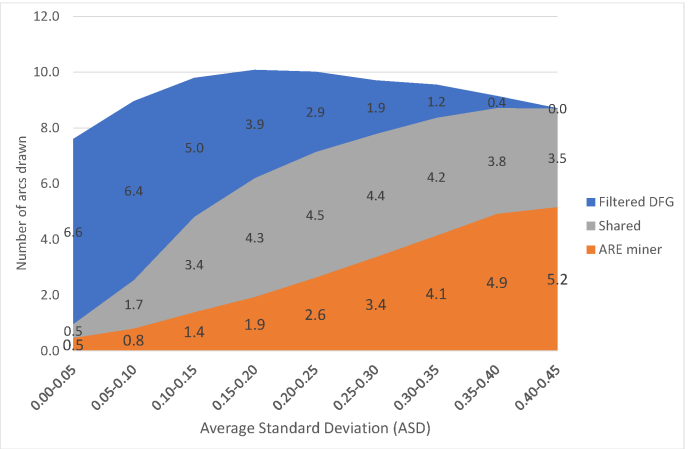


Figure 4.5: Average number of arcs drawn by the ARE miner and the filtered DFG-based approach, plus the number of shared arcs that are drawn when comparing the two techniques

filtered DFG-based approach. This highlights the value of building on the notion of statistical significance in this setting. Next, it is interesting to apply the ARE miner to a case study to investigate if the graphs produced by the ARE miner can provide relevant and meaningful domain-specific insights.

4.5.2 Qualitative Evaluation

This section discusses the qualitative evaluation for the ARE miner. Our goal is to demonstrate the effectiveness of the ARE miner to discover models that allow obtaining meaningful insights into action-response-effect patterns.

Data set

To evaluate the ARE miner, we use a real-world data set related to the care process of a Dutch residential care facility. The event log contains 21,384 recordings of aggressive incidents from 1,115 clients. The process captured in this log concerns the aggressive behavior of clients in their facilities and the way client caretakers respond to these incidents. The log consists of aggressive incidents of clients that belong to one of four different action classes. Each of these actions is followed by a number of measures from the caretakers as responses to the action. Each response belongs to one of nine different response classes. In line with the description of the ARE miner, we transformed this log into an action-response-effect log by defining the next aggressive incident of a client as an effect, given it occurred within an ϵ of 9 days as indicated based on the data. Otherwise, the effect is determined with τ. As a result, we obtain a total of five different effect classes. Table 4.7 summarizes the characteristics of our data set.

Actions	Physical aggression towards people	11,381
	Physical aggression towards objects	1,446
	Verbal aggression	5,778
	Self-injury	2,779
	Total	*21,384*
Responses	Talk to client	9,279
	Held with force	3,624
	Leave room	3,638
	Distract client	2,561
	Send away	3,169
	Seclusion	1,156
	Other measures	209
	None	783
	Ignore client	70
	Total	*24,489*
Effects	Physical aggression towards people	5,897
	Physical aggression towards objects	686
	Verbal aggression	2,369
	Self-injury	1,429
	No next incident (τ)	9,888
	Total	*20,269*
Clients	Minimum number of actions per client	1
	Maximum number of actions per client	449
	Average number of actions per client	19.2
	Total	*1,115*

Table 4.7: Overview of the characteristics of the real-world data set

Results

Below we present the results from the qualitative evaluation. We focus on three particular aspects: 1) the *interpretation* of the resulting graphs, 2) the *insights* we can obtain from these graphs, and 3) how the graphs *compare* to directly-follows graphs.

Interpretation. After applying the ARE miner to the data set, we obtain four graphs, one for each action class. In Figure 4.6 we show the resulting graphs for each action. Each arc in these graphs denotes an influential point representing a response-effect interaction. Recall that the number of observed instances for influential points is significantly higher (solid arc) or lower (dotted arc) than the statistically expected number of instances. In addition, the thickness of the arc visualizes the size of the effect, i.e., the thicker the arc, the stronger the effect. As illustrated by the graphs, the resulting number of arcs is, despite the complexity of the log, quite low since only a few arcs represent statistically significant interactions. This allows us to study the impact of each response to an action in detail.

To illustrate this, consider Figure 4.6a, where the initial action is *Physical aggression against objects* (po_s). Among others, this graph reveals that *Terminate contact* has been observed 299 times in our data set as a response to *Physical aggression against objects*. We can further see that in 48 instances, this response has led to the

effect *Verbal aggression*. In brackets, we can see that the expected value based on statistics for this arc is 32 instances. Thus, there are significantly more instances of *Verbal aggression* after a *Terminate contact* response than statistically expected, which is visualized using a solid arc. If we consider the response *Seclusion*, we can see that we observe an opposite effect for *Tau* (τ). Here, the observed number of instances (27) is significantly lower than the expected number of instances (45). Therefore, the interaction is visualized using a dotted arc. What both cases have in common is that they represent statistically significant interactions, which increase our understanding of what likely or unlikely effects a particular response will trigger.

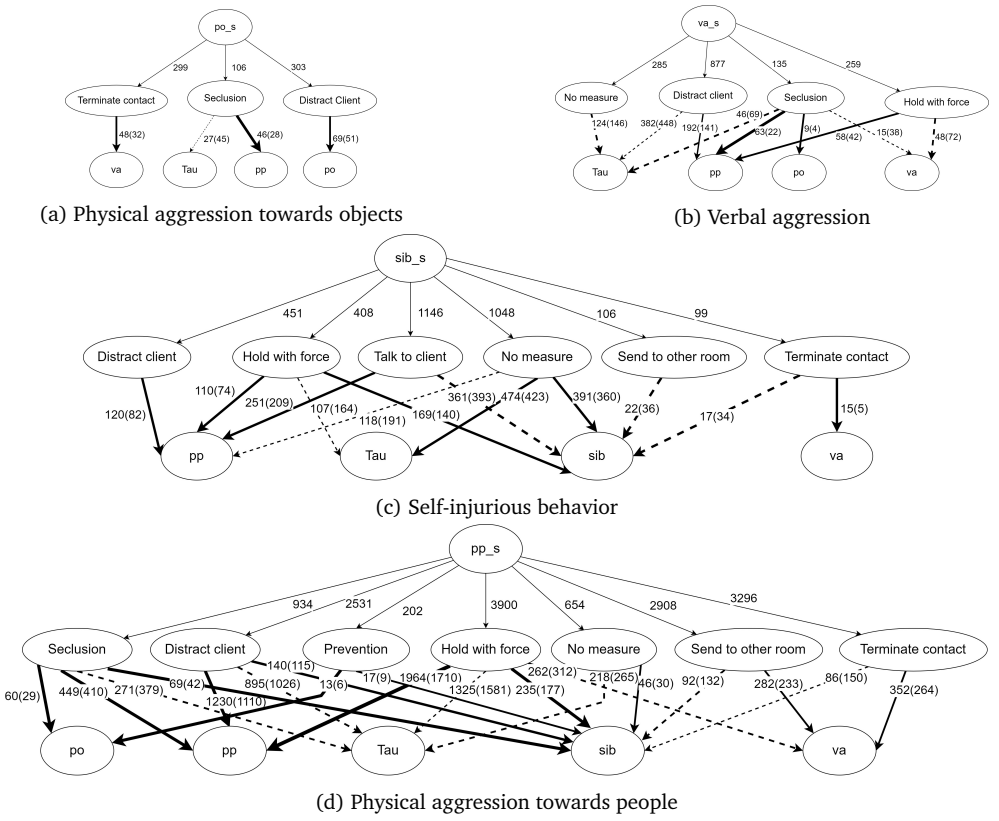
Insights. There are a number of relevant insights we can obtain from the graphical representations in Figure 4.6. For instance, Figure 4.6a reveals that *Seclusion* is not a good response to the action *Physical aggression against objects* since it results in a significantly higher number of instances of *Physical aggression against people* (*PP*). The frequencies show that the response *Seclusion* is almost 1.7 times as likely to result in effect *Physical aggression against people* than statistically expected. At the same time, the response *Seclusion* leads to a significantly lower likelihood of having no further aggression incident, as indicated by *Tau*. This highlights even further that *Seclusion* as a response to *Physical aggression against objects* is not a preferable choice.

In a similar fashion, we can interpret the resulting graphs for the other three actions. However, particularly Figure 4.6c and Figure 4.6d highlight that also focusing on statistically significant interactions may result in relatively complex representations. From an insight perspective, we can make three main observations: 1) depending on the action, the same response may lead to different outcomes, 2) some response-effect pairs are only significant for some actions, and 3) some response-effect pairs are significant across all actions.

The first observation is best illustrated by the response *No measures*. We can see that the response *No measures* leads to very different outcomes in Figure 4.6b and Figure 4.6c. If *No measures* is used as a response to the action *Verbal aggression* (Figure 4.6b), we observe fewer instances of *Tau* than statistically expected. In other words, it leads to an escalation of aggression since it is less likely that no next incident occurs. By contrast, if *No measures* is used as a response to the action *Self-injurious behavior* (Figure 4.6c), it leads to a significantly higher number of instances of *Tau* than statistically expected. What is more, it leads to significantly fewer instances of *Physical aggression towards people* than statistically expected. Both these outcomes can be considered as a de-escalation of aggressive behavior which are valuable insights for the management of this behavior.

For the sake of illustrating the second observation, consider the response *Talk to client*, which only occurs as part of a significant interaction in the context of the action *Self-injurious behavior* (Figure 4.6c). Here, we see that it leads to an escalation of violence, i.e., to significantly more *Physical aggression towards people*.

An example of the third observation is the response *Seclusion*, which has a similar effect across all four actions. We see that it leads to significantly more *Physical aggression towards people* and significantly less to *Tau*. This means that we see more of the most severe form of aggressive behavior and, at the same time, a lower likelihood of no next aggressive incident. Thus, we can globally speak of an escalation of the violence as a result of this response. Notice that for *Self-injurious behavior* there is no



(a) Physical aggression towards objects

(b) Verbal aggression

(c) Self-injurious behavior

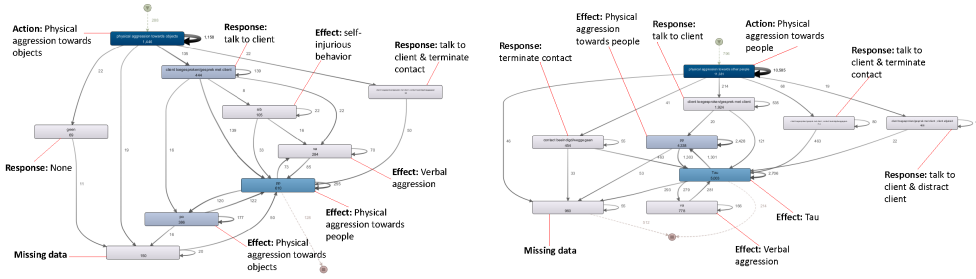
(d) Physical aggression towards people

Figure 4.6: Graphical representation of applying the ARE miner on the action-response-effect log for three initial actions

effect. This is logical considering that this response is used to restrain a client from being violent. As such, this response is rarely used in a setting where the victim is the client him/herself.

Comparison to directly-follows graph. Figure 4.7 shows the directly-follows graphs obtained for the actions *Physical aggression towards objects* and *Physical aggression towards people* using the process mining tool Disco[3]. For the sake of readability, the filtering settings are set to 5% and 1% respectively, i.e., only the 5% and 1% most frequent action-response-effect patterns are included. A brief analysis of the graph reveals that it does not allow us to obtain the same insights as the miner proposed in this chapter. Most notably, the process model contains a large number of arcs. Given that our data set contains four action classes, nine response classes, and five effect classes, the directly-follows graph can potentially contain 81 $(9 * 4 + 9 * 5)$ arcs. Already a single instance of a particular response-effect pattern for a considered action will result in an additional arc. The number of arcs increases exponentially with the number of responses observed. A possible solution to this could be to add information

[3]Disco Tool Website: https://fluxicon.com/disco/.



(a) Physical aggression towards objects

(b) Physical aggression towards people

Figure 4.7: Directly-follows process model of the real-world event log for the initial action physical aggression against objects (PO) and physical aggression towards people (PP). This shows the process filtered on 5% of the possible activities and paths for the initial action PO and 1% of the possible activities and paths for the initial action PP. The models were created using Disco

to the control-flow-based representation, such as the observed frequencies of the arcs or nodes.

However, filtering based on the frequencies does not always deliver the desired result. This is also illustrated in Figure 4.7. It shows that filtering could even be misleading since the data set is imbalanced. In this real-world scenario, a high frequency does not imply a significant pattern. This becomes obvious if we compare the techniques. From the figures, we can see that none of the significant response-effect pairs from Figure 4.6a are displayed in Figure 4.7a. In addition, for the most complex action, only one pattern (*PP - Distract client* - tau/τ) of response-effect can be observed in Figure 4.7b.

In order to understand the relations in the representation, we have to account for the relative frequencies. These reveal the meaningful insights that are hidden in the representation of a discovery technique such as the directly-follows approach. Hence, even after applying filtering mechanisms, Figure 4.7 does not provide the necessary insights. For example, even though we can observe a pattern for the action *Physical aggression towards people (PP)* in both our graph and the directly-follows graph, we cannot determine the meaning of the arc in the latter. We cannot assess whether the frequency that is displayed on the arc (22) is a statistically relevant effect. Hence, the directly-follows graph does not provide the insights that are required to answer a question such as: If a client displays aggressive behavior of class X, which response is likely to lead to a (de-)escalation or future aggression? If we consider the same pattern again for the action *PP*, we can see in our graphical representation that we show that the response *Distract client* leads to significantly fewer instances of *Tau*. This means that this particular response to the considered action seems to escalate violence, after all the chance of no next incident occurring is lower than we would expect based on statistics.

The qualitative evaluation in this section using a real-world data set strongly suggests that the representations generated by the ARE miner allow obtaining relevant domain-specific insights that can be directly translated into practical guidance.

4.6 Discussion

In this section, we discuss the implications as well as the limitations of the work presented in this chapter.

Implications. The key question identified at the start of this research addressed the desire to express insights into how a response to an action can lead to a desired or undesired outcome (effect). In our problem statement, we identified two main challenges associated with this that need to be overcome: (1) graphical representation, and (2) effective filtering mechanism. Our evaluation uses an artificial log covering a broad range of scenarios that highlight how the proposed ARE miner addresses both these challenges. In addition, we evaluate the ARE miner a real-world data set from the healthcare domain. Figure 4.6 shows that the ARE miner creates a simple graphical representation that allows for insights into statistical relations that cannot be obtained using Figure 4.7. In addition, we show that the use of statistics substantially reduces the number of arcs in comparison to a naive DFG-based approach. The filtering mechanism is also effective in the sense that it filters those arcs that are meaningful, as opposed to those that are merely frequent.

One interesting implication of the ARE miner-generated insights can be used to support decision-making processes. In our example, Figure 4.6 can be used to train existing and new staff members to ensure that appropriate responses are taken. For example, one could show that responding with *Seclusion* will likely escalate future violent behavior of the client. Placing the ARE miner in a broader medical context, the ARE miner could help make informed decisions when different treatment options are considered. In a different domain, the ARE miner could help a marketing organization understand the effectiveness of marketing strategies in terms of the response to potential customers. In short, the ARE miner provides insights into action-response-effect patterns where the objective of analyzing the process is to understand possible underlying statistical dependency patterns.

Limitations. The work presented in this chapter is subject to a number of limitations, which relate to the ARE miner itself as well as the experimental evaluation.

As for the *ARE miner*, there are three main limitations. First, we assume the independence of the responses. This means that each response has a unique impact on the effect and there is no interaction effect when responses are combined. For example, if response r_1 is observed to lead to effect c_1 and response r_2 is observed to lead to effect c_2, then only these independent patterns will be included even if the *combination* of r_1 and r_2 actually leads to c_3. Adjusting for this, would require a new formalization and introduce considerable additional complexity since the set of responses would be no longer R, but $R \times R$. Second, our formalization defines an effect as the next occurrence of an action after the response. In certain scenarios, it could be very interesting to consider generalizing this formalization by allowing the effect to be any type of event or activity. However, with such a generalization we can no longer compare the ARE miner with a directly-follows graph. As such, such a fundamental adjustment requires an entirely different means of comparison, which is left for future research. Third, we need to consider the data requirements of the ARE miner. The scenario in which the ARE miner is mostly applicable is when there is a choice to be made in the process. Hence, a form of categorical data needs to be available. In addition, the ARE

miner does not allow for continuous variables to be included. However, continuous variables can often be transformed into categorical variables.

The main limitation that needs to be considered for the *experimental evaluation* concerns the parameter settings for the artificial data. There are a variety of parameter settings that could be further explored. For example, changing the value of δ in the PMF generation phase from 0.2 to 0.1 would generate more precise insights into the mechanisms underlying the techniques. In addition, varying the number of responses and effects may lead to further interesting insights. There are two main reasons why we did not address these limitations in this research. First, the adaptation of these parameters cannot be expected to provide fundamentally different insights since they will mainly affect the granularity and size of the data. Second, the experiments for the presented setting already required substantial computing power. Therefore, we decided to stick to the chosen setting.

A limitation that is more of secondary nature is the variety of ways in which the frequency filter for the DFG-based approach can be implemented. Most filters aim to capture one of two key elements of the DFG-based approach: time or activities/paths. The ARE miner focuses on the filtering of activities (nodes) and paths (edges). Time, in the ARE miner, is represented via the definition of ϵ and thus represents a constant. Therefore, filtering based on the frequency of paths provides results that are best suitable for comparison.

4.7 Related Work

Over the last two decades, a plethora of process discovery techniques has been proposed (Augusto et al., 2018). The majority of these approaches generate procedural models such as Petri nets (Song et al., 2015; Verbeek et al., 2017), causal nets (Nguyen et al., 2017; Yahya et al., 2016), BPMN models (Augusto et al., 2017; Broucke & De Weerdt, 2017) or process trees (Buijs et al., 2012; Leemans et al., 2013). Some techniques also discover declarative models (Bernardi et al., 2014; Schönig et al., 2016) or hybrid models (i.e. a combination of procedural and declarative models) (De Smedt et al., 2015; Maggi et al., 2014). What all these techniques have in common is that they aim to discover the control flow of a business process, that is, the execution constraints among the process' activities. The ARE miner clearly differs from these traditional process discovery techniques by focusing on action-response-effect patterns instead of the general control flow.

There are, however, also alternative approaches to process discovery. We distinguish two prominent classes of techniques: artifact-centric process discovery and causal mechanism discovery. Several authors addressed the problem of artifact-centric process discovery (Lu et al., 2015; Nooijen et al., 2012; Popova et al., 2015). The core idea of artifact-centric process discovery is to consider a process as a set of interacting artifacts that evolve throughout process execution. The goal of artifact-centric discovery, therefore, is to discover the lifecycles associated with these artifacts and the respective interactions among them. While artifact-centric discovery techniques move away from solely considering the control-flow of the process activities, the main goal is still control-flow oriented. A related technique to process discovery was proposed in (Eck et al., 2016; Eck et al., 2017). This technique focuses on

the different perspectives of a process and discovers and captures how their relations change using composite state machines. While the techniques from (Eck et al., 2016; Eck et al., 2017) are potentially useful in many scenarios we address with the ARE miner, the insights that can be obtained with the ARE miner differ substantially. The techniques from (Eck et al., 2016; Eck et al., 2017) allow us to understand how different artifact life cycle states are related. For example, it reveals that a patient in the state "*Healthy*" does no longer require a "*Lab test*". The goal of the ARE miner is to show what actually needs to be done (or should not be done) to make sure a patient ends up in the state "*Healthy*".

The second, prominent set of discovery techniques studies the phenomenon of causality in process mining (Bozorgi et al., 2020; Brunk et al., 2020). In (Bozorgi et al., 2020), the authors investigate how treatment can have a (high) causal outcome for certain subgroups of cases. In their work, they propose an action-rule-based technique in which uplift trees are used to determine the subgroups for which the causal relations are relevant. In (Brunk et al., 2020), the authors rather look at the context in which the process takes place and run a Dynamic Bayesian Network model to determine causal relations. When we compare the ARE miner to previous work, we see that there are three main differences: (1) the definition of subgroups, (2) the transparency of the technique, and (3) the comprehensibility of the output. We discuss each of these differences in detail below.

First, the techniques differ in the way subgroups are defined. One strand of literature, the rule-based approaches, and related works, allow for the discovery of subgroups based on data. The main advantage of data discovery is that new subgroups can be discovered. By contrast, other techniques take subgroups that are defined by the user as input. The main advantage of the user-defined subgroups is that the subgroups intuitively make sense to the user of the approach. Therefore, the results of this user-defined approach provide results that are inherently meaningful and actionable to the user.

Second, the transparency of these techniques differs greatly. Previous research has shown that this plays a crucial role in the acceptance of and trust in new techniques. In (Shortliffe, 2012), the authors showed the crucial importance of understanding how a prediction is made for people in the medical domain in order to come to a decision about a patient's treatment. In (Martens et al., 2016), the authors expand on this by showing the same holds for other business domains. The main advantage that machine learning-based techniques have is that they are very powerful and can result in highly accurate results. However, the transparency of the process of obtaining the results is a known dilemma in machine learning techniques (Adadi & Berrada, 2018). As a result, the movement of explainable artificial intelligence prescribes that retrospectively additional models can be built to gain insights into this process. According to (Du et al., 2019), we can distinguish two counteractions to this: explainable by design or explainable post-hoc. The latter is used in the works of both (Bozorgi et al., 2020) (uplift trees) and (Brunk et al., 2020) (sensitivity analysis). Using a statistical approach, as we do in this chapter, the transparency of the technique is provided by design. All outcomes of the technique are traceable and can be recalculated manually. The main advantage of this is that the ARE miner is intrinsically transparent and thus

ensures insights into *why* and *where* questions, e.g., why do certain (causal) relations hold and where do these relations originate?

Lastly, the output of the approaches differs substantially. For rule-based approaches, the output is declarative in the sense that a set of textual rules are defined to which a case should hold in order to optimize the effect. In addition, in (Brunk et al., 2020), the authors provide the user with probabilistic parameters as output variables. The ARE miner also produces probabilistic parameters, but not as output. The ARE miner builds on the parameters by introducing an extra translation of the results into graphical representations with effect size indications. In this way, we provide the user with an understandable and actionable representation.

To the best of our knowledge, we are the first to propose a technique, the ARE miner, that discovers action-response-effect patterns and allows the reader to develop an understanding of why certain events occur. The ARE miner creates this understanding by having user-defined subgroups, which are used in a transparent technique to produce probabilistic visual output that is intuitive for the user.

4.8 Conclusion

This chapter presents the ARE miner to discover action-response-effect patterns within work processes. We identify two main challenges that we address in this research: (1) comprehensible graphical representation, and (2) effective filtering mechanism. In order to address these challenges, we propose the ARE miner that builds on filtering influential relations using statistical tests. We evaluate the ARE miner in two ways. First, we use an artificial data set to compare the performance of the ARE miner to traditional process-oriented representations. The results show that the ARE miner leads to both: (1) a reduction in the number of arcs drawn, and (2) a set of arcs that is different and more meaningful compared to the DFG-based approaches. Second, we evaluate the ARE miner on a real-world data set from the healthcare domain. More specifically, we use the ARE miner to study aggressive behavior and show that we can gain valuable and novel insights from the representations discovered by the ARE miner. The representations show that the ARE miner tackles both challenges by providing an easy-to-interpret representation that only displays meaningful relations such that it highlights informative insights.

In future work, we plan to undertake a number of steps to extend this work. In line with the limitations we presented previously on the ARE miner, we plan to conceptually extend this work by 1) developing an extension to the ARE miner that can estimate and incorporate the interaction effect that can arise when there are multiple responses to an action and 2) revisiting the concept of effects to see if we can relax the formalization to allow for different types of effects other than the next action of a process. Besides these conceptual extensions, we also plan to conduct additional evaluations. Most importantly, we intend to further test the ARE miner on additional real-world cases. What is more, we plan to compare the results of the ARE miner to machine learning-based approaches. In this way, we can obtain further insights into the applicability and the value of the ARE miner.

Part II: Moving Beyond the ARE miner

CHAPTER 5

Uncovering Complex Relations in Patient Pathways based on Statistics: the Impact of Clinical Actions

Reading Guide. In the first part of this dissertation we focus on developing the ARE miner, a novel process mining technique to detect, analyze, and visualize potential dependency patterns in work processes. The second part of the dissertation focuses on extensions to the ARE miner. In the following chapter, we generalize the ARE miner by proposing a state-action approach to studying potential dependency patterns. This state-action approach allows for a more relaxed definition of the input for the technique. Where the ARE miner is limited to studying activities, the state-action technique allows for less strictly defined input (e.g. the status of a patient). Again, the state-action technique focuses on potential dependency patterns. Thus, in this chapter we look into the statistical interplay between states and actions.

This chapter is based on the following publication:

Koorn, J. J., X. Lu, F. Mannhardt, H. Leopold & H. A. Reijers (2022c), "Uncovering complex relations in patient pathways based on statistics: the impact of clinical actions", in: *Proceedings of the 55th Hawaii International Conference on System Sciences*.

5.1 Introduction

Process mining is a popular family of techniques that uses event log data to visualize and analyze organizational processes (Van der Aalst, 2016). Also, in the field of healthcare, process mining techniques have become increasingly popular in recent years (Mans et al., 2015; Rojas et al., 2016). The healthcare field offers many opportunities for process mining to support data-driven analyses to optimize complex healthcare processes (Martin et al., 2020). Many healthcare case studies (e.g. (Mans et al., 2008)) have been performed and even new techniques and methodologies have been proposed (e.g. (Rovani et al., 2015)). As such, the healthcare domain and process mining techniques have proven to be a fruitful combination.

However, most process mining techniques focus on the control-flow perspective, i.e. they focus on the order in which activities are performed. To investigate processes from this perspective, a number of discovery techniques have been proposed over the years. Prominent examples include the heuristic miner (Weijters & Van der Aalst, 2003), fuzzy miner (Günther & Van der Aalst, 2007), and inductive miner (Leemans et al., 2013). In a healthcare setting, these techniques can help healthcare professionals, among others, to understand possible patient pathways (Prodel et al., 2015). So-called conformance checking techniques can also show whether medical guidelines and procedures are appropriately followed (Gatta et al., 2019).

Once such insights are obtained, natural follow-up insights are often required to deepen the understanding of a process. In that light, practitioners are often interested in finding out how particular actions impact the patient pathways and their well-being. The control-flow perspective does not reveal the impact that actions may have on the process. Recognizing this, existing process mining techniques have proposed several alternative ways to uncover these insights. From a causal mechanisms perspective, a number of techniques have been proposed that try to find statistically significant patterns that show the impact of treatments in processes (Bozorgi et al., 2020; Koorn et al., 2020). Composite state machine techniques also aim to tackle the problem by identifying state transitions (Eck et al., 2016).

Nonetheless, both strands of techniques face some fundamental limitations that prohibit their applicability in uncovering the impact of actions on the process. Firstly, the techniques related to causal mechanisms are often not generic in their applicability. To exemplify, the authors of (Koorn et al., 2020) propose a technique that can only analyze repetitive processes and single actions (rather than multiple actions). Secondly, the composite state machine techniques can capture the status of a patient state and the transition between patient states, but they cannot show how process activities influence state transitions. The limitations of the existing process mining techniques motivate the proposal of a novel approach to gain an understanding of the complex healthcare processes. Therefore, we propose a novel and generic state abstraction approach to generate new insights in terms of the impact of actions on the transition between states in organizational (healthcare) processes. In particular, we propose a statistical approach to provide insights into the complex relations that influence state transitions. We apply this technique to a public data set on sepsis and uncover previously hidden relations in the healthcare process. These new insights can form the basis for the hospital to improve their sepsis process in two ways: (1)

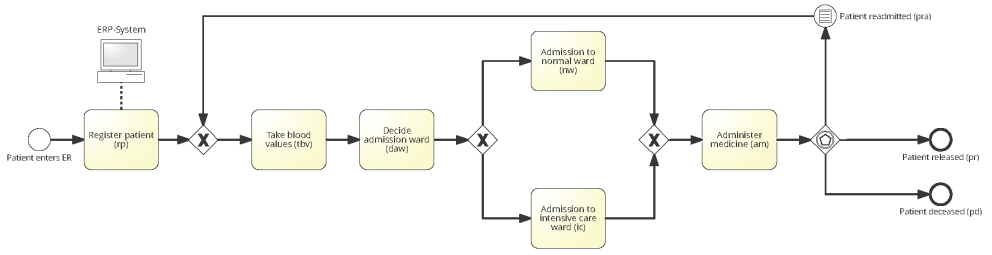


Figure 5.1: Simplified sepsis process as a BPMN model

to optimize their cost-benefit balance in patient care, and (2) to review the way they discharge patients to avoid early returns to the emergency room.

The rest of the chapter is organized as follows. In Section 5.2, we illustrate the problem we address in more detail. In Section 5.3, we introduce our approach from a conceptual point of view. In Section 5.4, we report on the results of our case study on data from an emergency room where we show that the proposed approach can be applied to real-world data and provides us with new insights on the impact of actions on state transitions. In Section 5.5 we discuss related work before in Section 5.6 we discuss the limitations and conclude the chapter.

5.2 Problem Statement

To illustrate the problem we address in this chapter, consider the scenario of a patient being admitted to the emergency room (ER) of a hospital. In Figure 5.1 we visualize such a healthcare process. What we see from the process model in Figure 5.1 is that the process starts when a patient is registered (*rp*). After registration, the medical team is required to perform initial blood tests to gain an understanding of the situation of the patient (*tbv*). Then, the medical team faces a decision to admit the patient to a normal care ward or an intensive care ward (*nw* and *ic*). Next, the medical team performs a number of activities that have to do with providing treatment to a patient in the form of administering medicine (*am*). For example, they can administer antibiotics through an intravenous line (IV). The end of the process, for most patients, is the discharge from the ER (*pr*). However, some patients return to the ER at a later stage (*pra*) or pass away (*pd*).

Process mining techniques can help healthcare practitioners and managers to obtain better insights into this process. Based on data logged by IT systems so-called process discovery techniques can detect and visualize which specific activities were performed and in which order these activities were executed. These activities are often a mix of medical activities (e.g. treatments) and logistic activities (e.g. transferring patients). Figure 5.1 represents the outcome of such a process discovery technique. We can see that the resulting process model gives an overview of the main process flow. This focus on activities and their order are commonly referred to as the *control-flow perspective*.

Visualizing the order in which activities are performed already aids in gaining an

overall understanding of the process. However, it does not reveal the impact of certain decisions on the process flow. Given the potential severity of decisions in an ER, it is very important to understand which decisions lead to desirable and undesirable outcomes. For example, doctors may want to understand if admitting a patient to an intensive care ward (*ic*) increases or decreases the chances of that patient needing to return to the ER at a later stage (*pra*). To understand such aspects, several process mining techniques complementing the control-flow perspective have been developed. We can identify two relevant research directions in this regard: (1) causal mechanism, and (2) artifact-centric process discovery.

Previous research has looked into this phenomenon from a *causal mechanism* perspective (Bozorgi et al., 2020; Koorn et al., 2020). One technique is proposed by the authors of (Koorn et al., 2020). In their approach, the authors study repetitive aggressive behavior of clients by trying to identify statistically significant patterns in actions of clients, responses of caregivers, and future aggression of clients. The technique provides insights into the process of aggressive behavior but is limited in its applicability to other scenarios for two reasons. First, the technique focuses on *repetitive* processes. As such, the outcome of the process (the effect) must equal the input (action) of the next iteration of the same process. Second, the technique only considers the impact of *singular* actions but ignores the possibility of sets of actions. Given these limitations, we cannot study our sketched problem setting. The sketched healthcare process is not a repetitive process as its input (the activities conducted in the context of the ER) and outcome (the discharge type) are not equal. Secondly, the activities performed in the process can also be sets of activities rather than a single activity, e.g. taking a number of blood tests rather than a single blood test.

Another research direction that relates to problems similar to ours are *artifact-centric* discovery techniques (Eck et al., 2016). These techniques use composite state machines to capture and visualize different perspectives on a process and how the relations in a process change. The advantage of such techniques is that, by using states, a generic approach is provided. The notion of states is interesting in this setting as we cannot capture the status of a patient in terms of activities, only in terms of states (e.g. *pr* or *pd*). In the sketched problem we are interested in how the process activities (e.g. *ic* or *mw*) influence the state of a patient. However, the composite state machine techniques do not provide insights into the activities that impact the state transitions. In the above problem scenario, we would be able to understand state transitions, but not the underlying causes for those transitions.

In sum, to obtain the insights required in our scenario, existing process mining techniques are either not sufficiently generic or do not provide insights into the complex relations. The combination of these two factors is crucial to aid practitioners and managers in gaining an in-depth understanding of the process and generating actionable insights. In this research, we introduce a novel approach that explicitly differentiates between *states* and *actions* and automatically discovers whether (sets of) actions lead to different states.

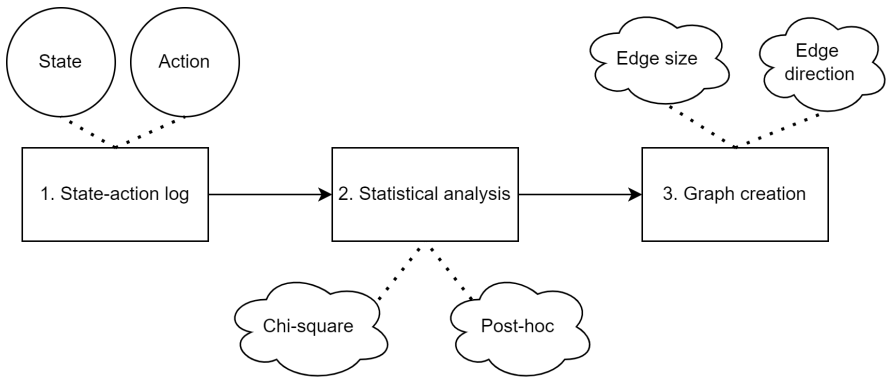


Figure 5.2: Proposed approach

5.3 Approach

In Figure 5.2 we show which steps we take in the approach. Below, we go into detail on every step.

5.3.1 Step 1: State-action log

Observed	**Normal discharge**	**Deceased**	**Readmission**	*Total*
Normal ward	350	50	90	*490*
Intensive care	350	150	60	*560*
Total	*700*	*200*	*150*	***1,050***

Expected	**Normal discharge**	**Deceased**	**Readmission**	*Total*
Normal ward	326.7	93.3	70.0	*490*
Intensive care	373.3	106.7	80.0	*560*
Total	*700*	*200*	*150*	***1,050***

Table 5.1: Running example sepsis chi-square table

The input for our approach is an event log L that captures how the considered process was executed. Formally, we can define L based on the universe of all events $\mathcal{E}$. The events recorded for a single execution (i.e., an instance) of the process are called a *trace*, which is modeled as a sequence of events. Therefore, we denote a trace with n events as $\sigma = < e_1, \ldots, e_n >$, where each event $e \in \sigma$ is part of $\mathcal{E}$. However, for the purpose of our analysis, we require an adapted notion of such a traditional event log. Therefore, we build the state-action log. Specifically, we need to distinguish which events from a trace σ represent *states* and which events from σ represent *actions* that lead to state changes. To this end, we introduce a function α, which returns for each $e \in \mathcal{E}$ whether e represents a *state* or an *action*. To illustrate this, consider the trace $\sigma_1 = < rp, tbv, daw, nw, am, pr >$, which represents a possible trace according to the process shown in Figure 5.1. As pointed out above, we are particularly interested in

the impact of the activities conducted in the ER on the discharge type (*pre*, *pra*, or *pd*). Therefore, we define α in such a way that it maps the three discharge types to *states* and all other events to *actions*. As a result, we can analyze whether and which activities lead to one or the other discharge type.

Due to the complexity of real-world processes, the definition of α requires input from domain experts. This means that we ask the user to define which events represent *actions* according to our definition. While this is certainly associated with some effort, it makes sure that our technique can detect practically relevant relationships.

5.3.2 Step 2: Statistical analysis

Once the state-action log is created, the approach moves to the analysis step in which we perform statistical tests to discover significant relations. Here, we perform chi-square tests, which are subject to a number of assumptions. This step is followed by post-hoc tests. The *Chi-square* test is a well-established statistical test proposed halfway in the last century (Cochran, 1952). The Chi-square test has many advantages (McHugh, 2013), the main reasons for its use in this approach are: (1) it is very robust due to its non-parametric nature (i.e. it does not make assumptions about data distribution), (2) the computation of the statistic is relatively easy and fully transparent, (3) the test provides rich information, and (4) the test is very suitable for in-depth post-hoc tests.

The test takes two inputs: (1) a set of observed behavior for two categorical variables and (2) the chi(χ) distribution. The categorical variables in the context of our approach are the state and the action. The data that is used in the chi-square test represents how often each combination of state and event is observed in reality. This is referred to as the *observed frequency*. The observed frequencies are converted into an $n \times m$ table where n represents the number of action sets and m represents the number of states. This is called the observed frequency table, see for example the top of Table 5.1. Based on this table the chi-square distribution is introduced. In our example scenario we have two actions; *admission to normal ward* and *admission to intensive care unit*, and we have three states; *normal discharge*, *deceased*, and *readmission*. Table 5.1 shows artificially generated data for illustrative purposes.

The chi-square distribution is used to calculate the *expected frequency* for each cell (i.e. combination of action and response). The expected frequency calculator takes the distribution and the observed frequency grand total and total of each column and row. Using these number the following formula is applied: $\frac{\mathrm{N_r} \times \mathrm{N_c}}{E[\mathrm{s}][\mathrm{a}]}$ where N_r equals the expected row total, N_c equals the column total and $E[s][a]$ equals the grand total.

Take for example the action *normal ward* and state *normal discharge*. The observed frequency is 350. The expected frequency is calculated based on the column total (700), row total (490), and grand total (1,050). Applying the formula on the example gives: $\frac{490 \times 700}{1050}$ = 326.7 The result of this exercise is represented in a separate table called the expected frequency table, see the bottom of Table 5.1.

Now, the chi-square test compares the difference between the observed and expected frequency for each cell. These differences are all summed and result in the chi-square score. The formula used for this is:

$$\chi_c^2 = \frac{(O_{\text{s1,a1}} - E_{\text{s1,a1}})^2}{E_{\text{s1,a1}}} + ... + \frac{(O_{\text{sn,an}} - E_{\text{sn,an}})^2}{E_{\text{sn,an}}} \tag{5.1}$$

In the formula, $O_{sn,an}$ describes the observed value and $E_{sn,an}$ describes the expected value for a combination of state and action where 1 to n denote the individual action or state. If we apply this formula on our example case for all the combination of states and actions, i.e. from state *normal discharge (ND)* to state *readmission (R)* and for action *normal ward (NW)* and action *intensive care (IC)*. This results in the following equation:

$$\chi_3^2 = \frac{(O_{\text{ND,NW}} - E_{\text{ND,NW}})^2}{E_{\text{ND,NW}}} + ... + \frac{(O_{\text{R,IC}} - E_{\text{R,IC}})^2}{E_{\text{R,IC}}} \tag{5.2}$$

We know from Table 5.1 that there are three states and two actions, so the degrees of freedom: $c = (3-1) \times (2-1) = 3$.

$$\chi_3^2 = \frac{(350 - 326.7)^2}{326.7} + ... + \frac{(60 - 80.0)^2}{80.0} = 51.56 \tag{5.3}$$

If the observed and expected frequency values are sufficiently different, a large test score is returned. The next step is to check which pairs of observed and expected frequencies are sufficiently different. This step is quite complex and describes how the chi-test score is compared to the chi-distribution, see (Fisher & Yates, 1938) for more details. The larger the chi-square test score, the more likely it is a significant score. In this study, we test on the alpha = 0.05 level. If the chi-square test score is significant, this indicates that there is at least one set of actions for which we observe a significantly different frequency of states than expected. In the example above, the p-value is $6.4 * 10^{12}$ which is well below the alpha level of 0.05. Thus, we can conclude that there is a significant relationship between at least one set of actions and states.

The chi-square test is subject to a number of data assumptions, which need to be checked (McHugh, 2013). Most assumptions have to do with the way the data is collected and stored. They can only be manually checked by the analyst. However, one assumption relates to the sample size and can be automatically checked. Specifically, the assumption states that in the expected frequency table in 80% of the cells a minimum value of 5 must be present. In this work, we apply a heuristic selection criterion to test this. If the assumption is not met, an NA value is returned and no further test results are returned.

The next step in the analysis is to perform *post-hoc* tests. The goal is to determine which specific combinations of sets of actions and states are significantly related. To that end, a detailed statistic called the adjusted standardized residual (ASR) is calculated (Agresti, 2003). The ASR standardizes the difference between the observed and expected frequency using the following formula:

$$ASR = \frac{O_{\text{s1,a1}} - E_{\text{s1,a1}}}{\sqrt{E_{\text{s1,a1}} * (1 - \frac{E_{\text{s1,a1}}}{N_{\text{c}}}) * (1 - \frac{E_{\text{s1,a1}}}{N_{\text{c}}})}} \tag{5.4}$$

To apply it to our example case scenario, let us consider the state action combination of *normal discharge* and *normal ward*:

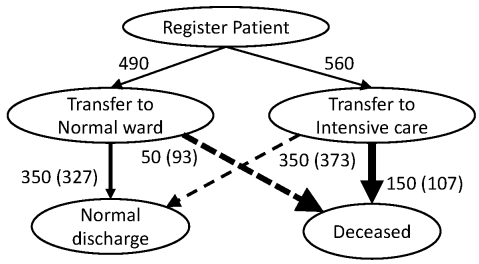


Figure 5.3: Resulting graph of the running example

$$ASR = \frac{350 - 326.7}{\sqrt{326.7 * (1 - \frac{326.7}{700}) * (1 - \frac{326.7}{700})}} = 5.70 \tag{5.5}$$

With the ASR score, three metrics are produced: (1) its significance, (2) its value, and (3) its size. In the analysis part of the algorithm, only the significance of the ASR is used. The ASR score is then normalized and compared to a standardized p-value score (critical score) (Agresti, 2003). The significance of the ASR score determines if it has a significant impact on the overall chi-square test score. In the example scenario above, the 5.70 score is significant. If the test result is significant, this means that the tested combination of state and set of actions is significantly related. Below, we will go into detail about how the graphical representations are created for all logs in which at least one ASR score is significant.

As we perform multiple statistical tests on the same data set we have to correct for the multiple comparisons problem (Miller, 1981). The multiple comparisons problem states that with each test we increase the chance of finding something by chance. To counteract this, we adjust the p-value we set as a significance barrier with the Bonferroni correction (Haynes, 2013). The test corrects the p-value according to the following formula: $p = \frac{alpha}{n}$ where *alpha* is the significance level, n equals the number of tests, and p equals the new p-value used for testing. Applying this to the running example, we perform tests on each cell (n=6) and test on a 0.05 alpha level. Thus: p-value $= \frac{0.05}{6} = 0.008$.

5.3.3 Step 3: Graph creation

The goal of the *graph creation* is to visualize the uncovered statistical relations from the analysis step. Existing work of (Koorn et al., 2020) proposes a technique to visualize statistical relations in process mining. In this work, we build on that particular visualization technique. The aim of graph creation is to increase the understandability of the graph to the end-user. To that end, the significance, value, and size of the ASR and the frequencies of the arcs are used to create graphical representations.

The *significance* determines which arcs need to be drawn for each node. If the ASR value is significant, an arc is drawn. If the ASR value is insignificant, no arc is drawn. The *value* of the ASR can be either positive or negative and determines its *direction*. The direction of the ASR cannot be negative as that would indicate a non-significant

value and thus no arcs would be drawn. If the ASR value is positive, a solid arc is drawn to indicate a positive relational direction. If the ASR value is negative, a dotted arc is displayed to indicate a negative relational direction. Returning to the example scenario described in the previous section, a 5.70 significant score would result in a solid arc to indicate a positive relational direction between a normal ward and normal discharge. In practical terms, after a normal ward, we see an increased likelihood that a normal discharge follows.

The *size* of the ASR value corresponds with the thickness of the arcs. In total there are six thickness classes, three for positive and three for negative values. These classes are based on the maximum and minimum ASR values and the critical value. A simple algorithm calculates the difference between the maximum and the critical score (usually $|2.57|$) and determines three equally large ranges for three classes. The classes reflect the effect size of the relation. A thicker arc means a stronger relationship. In our case, 5.70 is the largest ASR value. Thus, it would be represented as a thick solid arc in the graphical representation.

Note that a thicker dotted arc means a stronger negative relation. To exemplify, a thickly dotted arc between a set of actions and a state reflects that there is (highly) decreased likelihood that this state occurs after that set of actions. Finally, the observed and expected *frequencies* that are calculated in the second step are displayed on each drawn arc. This helps to gain an understanding of the prevalence of each combination of state and action.

If we apply this approach to the running example, as depicted in Table 5.1, we get the representation that can be found in Figure 5.3. Here, we see for example that admission to a normal ward (n = 490) (slightly) increases the chances of a normal discharge (observed = 350, expected = 350) and (largely) decreases the chances that the patient deceases (observed = 50, expected = 93). In addition, the same analysis can be applied to the cases that are admitted to the intensive care unit. Finally, we see that readmission is not included in the graph, which means that it has no effect on either the normal discharge or death of a patient.

5.4 Case study

The goal of this section is twofold: (1) to demonstrate the applicability of the proposed approach in a real-world scenario, and (2) to generate new insights into the complex relationship between state and action. To this end, the sepsis data from (Mannhardt & Blinde, 2017) is used in this research. Below, we will first go into detail on the data set and its characteristics. As will be explained below, the data can be analyzed in multiple ways. Therefore, we present two different ways to create a state-action log based on the sepsis event log in the pre-processing section. First, we present the results where we look into the effect of *continuous monitoring and substance administration* of a patient on the way a patient is discharged. Then, we present the results where we look into the impact of the *decisions of the medical team* on whether or not the patient returns to the ER room. Here, we study two main decision moments: (1) to which type of ward the patient is admitted, and (2) in what way the patient is discharged. The goal of looking into these challenges is to show that our approach can produce new work hypotheses that can be used to enhance the

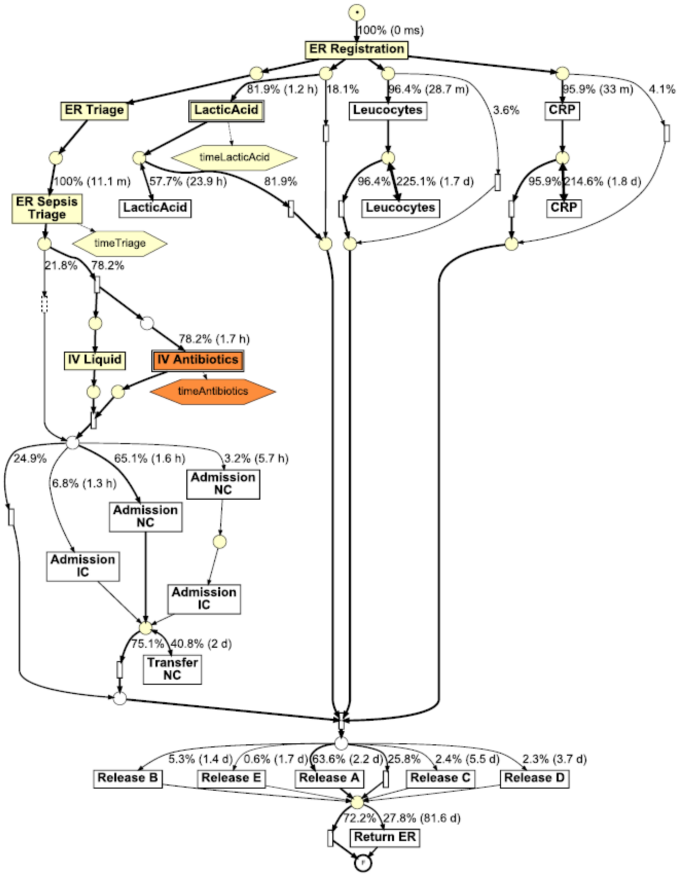


Figure 5.4: Multi-perspective process model used for decision mining on the sepsis process data in (Mannhardt & Blinde, 2017)

understanding of healthcare processes. Ideally, these insights form a starting point of a collaborative effort with healthcare practitioners and/or managers to uncover and understand the complex relations in their healthcare processes.

5.4.1 Case study data and context

The data for this project was collected in another study in which the researchers collaborated with a medium-sized regional hospital in the Netherlands (Mannhardt & Blinde, 2017). The project was aimed at studying the trajectories of emergency room patients, more specifically, those patients with symptoms of sepsis. Sepsis is a life-threatening condition where the body reacts to an infection by damaging its own tissue. Sepsis requires continuous monitoring and treatment with antibiotics. Data was taken from systems in three locations: emergency room, laboratory, and another ward. This data was anonymized and combined into one data warehouse after which an event log in XES format was created. In total, the log con-

	Activities mapped to states	**Activities mapped to actions**
Analysis 1	'ER Registration', 'Admission NC', 'Admission IC', 'Release A', 'Release B', 'Release C', 'Release D', 'Release E'	'Lactic Acid', 'IV Liquid', 'IV Antibiotics', 'CRP', 'Leucocytes'
Analysis 2	'ER Registration', 'return ER'	'Admission NC', 'Admission IC', 'Release A', 'Release B', 'Release C', 'Release D', 'Release E'

Table 5.2: Mapping from activities to states and actions of analyses

tains traces for 1,050 cases and 15,214 events that are recorded in 1.5 years of patient records (from November 2013 to June 2015). The log contains 16 unique activities and 846 distinct variants. More details on the data and its collection can be found in (Mannhardt & Blinde, 2017).

The event log contains sixteen activities: three activities for the registration and triaging, three activities for taking certain blood measures for patient monitoring (leucocytes, CRP, and lactic acid), two activities for the administration of substances to the patient (IV liquid and IV antibiotics), two activities for admission or transfer to normal ward or intensive care unit (admission NC or IC), five types of discharge (activities release A-E) from the hospital, and one activity capturing if patients returned to the ER at a later stage (within 28 days).

In their project, the authors of (Mannhardt & Blinde, 2017) mainly focus on three challenges: (1) conformance checking in terms of adherence to medical guidelines, (2) uncovering the various trajectories that patients can flow through, and (3) discover decision rules to detect returning patients (Mannhardt & Blinde, 2017). They create a process model depicted in Figure 5.4 to study these challenges. For our work, we aim to expand on this work by focusing on the latter two challenges. These two challenges are of interest as the impact of certain actions can influence both the trajectory of a patient as well as the patient return rate of patients to the ER. As for the *patient trajectories*, the authors *describe* the various types and frequencies of the paths that patients can take (Mannhardt & Blinde, 2017). They also indicate that some paths are more desirable than others. In this light, it is interesting to determine if there are certain actions that lead to a more desirable patient path. However, from the work of (Mannhardt & Blinde, 2017), no insights can be obtained from their findings as to which specific actions impact the path of the patient.

The first challenge addresses the patient trajectories, turning to the second challenge where the authors focus on *patients returning to the ER*. The authors of the original study try to apply decision mining techniques (De Leoni & Van der Aalst, 2013; Mannhardt et al., 2016) to discover if there were certain rules that could give insights into whether or not a patient would return to the ER within a certain time frame. Ideally, rules would be discovered that helped doctors gain insights into which actions to perform or decisions to take in order to reduce the number of returning pa-

tients. However, the authors could *not* find any such rules based on, amongst others, triage documentation and values from metrics taken. In this work, we will expand on the challenge of returning patients by studying if monitoring the patient continuously and/or administrating substances through IVs is correlated with a lower likelihood of the patient returning to the ER.

5.4.2 Pre-processing

As explained in our approach, we try and tackle two challenges: (1) patient trajectories, and (2) patients returning to the ER. For each challenge, we create a corresponding state-action log by mapping the activities to the set of states and the set of actions. For this cases study, we created two state-action logs, one for each analysis. For the *first analysis*, we mapped the ER registration, admission, and discharge activities as states and considered the tests and the medicines as the actions. This mapping allows us to find whether the lab tests and the medicines have any influence on the admissions (to IC or NC) and the type of discharge. For the *second analysis*, we mapped the admission and discharge activities to actions and the 'ER Registration' and 'return ER' activities as states. This mapping allows us to find whether the admission to IC or NC and the discharge types have any influence on the patient returning to the hospital or not. The concrete mapping is listed in Table 5.2. In the following sections, we discuss the results found.

5.4.3 Continuous monitoring and substance administration

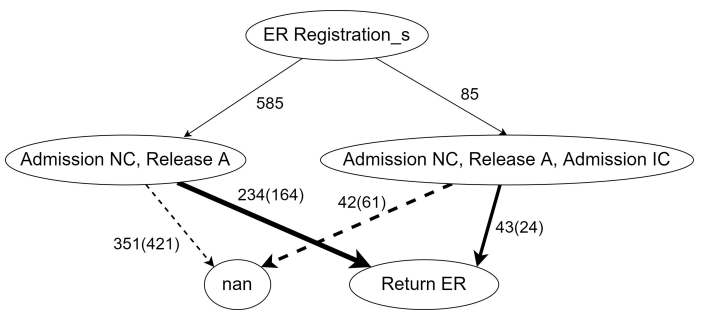


Figure 5.5: Sepsis analysis when we use metrics as actions and discharge type, admission type and registration as states

As mentioned previously, one of the challenges in the ER room is to find out how doctors can influence the trajectories of patients. In that respect, we developed a log where we are interested in the states: ward type and discharge type. *Ward type* can be either normal ward (Admission NC) or intensive care (Admission IC). *Discharge type* can be one of five types categorized as release A - release E. Finally, there can be patients still in the ER at the time of data collection; this should result in a nan (No return to ER). Next to the states, we are interested in the *actions* that can impact the state transitions: CRP, Leucocytes, Lactic acid, IV liquid, and IV antibiotics.

The first three actions are all decisions of the medical team to *(continuously) monitor* a patient's status through blood tests. The blood tests (CRP, leucocytes, and lactic

acid) are performed almost by default when a patient is admitted to the ER and has a slight fever. These tests are also performed regularly after a patient has been admitted to the hospital to check whether an applied treatment is working or not and may serve as a grounding to discharge patients. The last two actions regarding an IV refer to the decision of the medical team to *administer substances* to the patient. We can see in Figure 5.4 that these patient pathways exist, but we are interested in determining the *impact* of taking these blood tests and administering substances. The results from our approach can be split into two parts: (1) the relations visible in the graph, and (2) the lack of expected relations.

In Figure 5.5 we can see the statistically significant relations between the actions and the states. What we observe in the figure is that 656 cases are admitted to the normal care ward, roughly 60% of the population. In line with expectations, we can observe that two of the continuous monitoring actions are significantly related to the further trajectory of the patient: CRP and leucocytes. The CRP test refers to a c-reactive protein test; the leucocytes test is a white blood cell test. These are tests to check if the patient is reacting to inflammation.

In Figure 5.5 we can observe that performing these tests is related to a higher probability of release type A (normal discharge) and release type D (transfer to another care facility). Especially release type A is common, we observe that this follows CRP and leucocytes in 459 cases where we would only expect 372 cases based on the statistical tests. In addition, performing these two tests is related to a lower probability of readmission to the normal care ward, this is observed 121 times where we would expect 192 cases. In short, the results show that monitoring a patient's blood values is related to a higher probability of a normal discharge of that patient, which is the desired outcome for all parties involved. This is not so surprising, as the results of these tests are used to determine whether a patient can be discharged or not.

What cannot be observed from the figure is that there are a number of relations that are expected, but were *not* found. First, we would expect to see a similar pattern for all blood tests and substance administration. Interestingly, the other blood test (lactic acid) and the administration of liquid and/or antibiotics through an IV do not have a significant impact on the patient pathway. These three actions are performed when the medical team estimates that the patient's condition is severe. Therefore, we would expect it to have a relation with the admission to the IC and/or to other types of discharges. However, no such relationship is found in the data. It seems worthwhile to collaborate with domain experts to further investigate the effect of these actions on patient well-being. The goal of such a project could be to optimize the cost-benefit balance for the care of these patients.

5.4.4 Medical team decisions

The second challenge relates to the return of patients after they leave the ER. Previous work could not identify rules to determine what factors play a role in the return of patients (Mannhardt & Blinde, 2017). To follow up on this challenge, we investigate whether the decisions made by the medical team to admit a patient to a normal or intensive care unit and how the patient is discharged have an impact on the return of patients. As such, the *states* for this challenge are: registration and return to ER or nan (no return to ER). In turn, the *actions* are: admission type and discharge type.

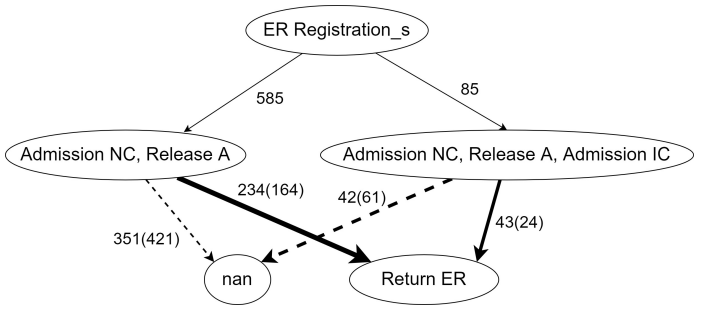


Figure 5.6: Sepsis analysis when we use registration and return as states, and admission type and discharge type as actions

From the original process model in Figure 5.4, we can see that the return of patients happens in almost a third of the cases (27.8 %), but we cannot infer what might cause a patient to return.

In Figure 5.6 we can see the statistically significant relations between the states and actions. What we can observe is that admitting the patient to the normal care ward and having a normal discharge (Release A) is a relatively frequent pathway for patients: it is observed for roughly half the patients (585 out of 1,050 cases). If we look at the impact of these actions, we see that they are related to a higher probability of a patient returning to the ER. We observe that this happens 234 times whereas we expect it to happen 164 times. In addition, these actions are related to a lower likelihood of patients not returning to the ER. This is observed 351 times versus the expected 421 times.

A similar pattern shows if a patient is not only admitted to the normal ward (Admission NC), but also the intensive care (Admission IC) followed by a normal discharge (Release A). The patient is, again, more likely to return to the ER and less likely to not return to the ER. This patient pathway is considerably less common but is still observed for roughly 10% of the patient population.

This is an interesting finding as it indicates that there is a group of patients that seem to be doing relatively well. This is indicated by the fact that they are admitted to a normal ward and discharged normally. However, a subgroup of this patient group might not be doing so well in reality as they return to the ER. This points to the fact that symptoms might be overlooked during the first admission that comes into play at a later stage. Further research should look into this patient population to identify which patient characteristics are associated with this to lower the return of patients. Gaining insights into this can result in better care for the patients as their pathology can be improved. In turn, this reduces both costs for the hospital and patients.

5.5 Related Work

As discussed in the problem statement, causal process mining techniques are related closely to the problem we set out to tackle in this chapter. To discover *causal mechanisms* in process models often decision rule or root cause approaches are proposed

(Bozorgi et al., 2020; Gupta et al., 2015; Suriadi et al., 2012). The authors of (Bozorgi et al., 2020) propose an exemplary case of such a decision rule mining technique. They develop a model in a financial context in which they use uplift trees to determine for which subgroups of clients a specified treatment has an effect. Although this technique produces valuable input, the technique heavily relies on the definition of subgroups. This is not always a suitable approach as it can be that there is a specific process for a (largely) uniform subgroup of patients. No insights can be generated into the effectiveness of treatments.

Another limitation that this approach poses becomes clear when we consider the output these techniques produce. The output of the decision rule mining techniques is often *declarative* in the sense that they produce a set of rules described in text format which serves as a guideline to optimize decision making in the process (e.g. (Bozorgi et al., 2020)). In other work on causal mining, the output of causal mining techniques is more imperative in the sense that the authors propose a more graphical style output (e.g. (Koorn et al., 2020)). Existing research has shown benefits for each type of output, but concludes that a hybrid form is optimal for maximizing the understandability of a process (Agresti, 2003). In turn, this should increase the effectiveness of turning process mining results into actionable insights (Martin et al., 2020).

In the present work, we present a novel approach that addresses these limitations. Our approach is generic in the sense that it is flexible in its formalization of state and action. In addition, our approach produces graphical representations (i.e. process models) with declarative elements (i.e. statistical relations and annotations) as output. This ensures that the approach is: (1) applicable to a wider spectrum of problem scenarios, and (2) that the results that are produced better can be used to support the interpretation and communication of the findings. To the best of our knowledge, no such approach or techniques have been previously proposed.

5.6 Conclusion

In this chapter, we proposed a novel approach to discover the statistical relations between states and actions. This work combined two main approaches to propose an approach that is both generic and identifies complex relations in organizational processes. The approach is generic in the sense that it can deal with a wide range of scenarios, as long as we can identify states and actions. The generic applicability stems from the flexibility in the concepts of states and actions. To exemplify, an action can consist of single or multiple activities. In the latter case, the action can be defined in different ways, for example, as a set or sequence of activities. The limitation of such a generic approach is that it requires manual work to define such actions and states. Ideally, this is done through a collaboration between a process analyst and domain expert.

We used a case study for two purposes: (1) to test the applicability of our approach, and (2) to generate new insights into complex relations. The case study concerned a public data set on 1,050 patients with sepsis that were admitted to the emergency room in a Dutch hospital. We applied the approach to tackle two remaining questions regarding patient pathways and patient return rates. First, the approach aimed to

identify those actions that have an influence and are related to the desired patient pathway. Thereby optimizing the cost-benefit balance in the care of patients. Second, the approach was used to determine what actions in the processes are related to higher or lower probabilities of patients returning to the ER at a later stage. Ultimately providing new insights into which patients are susceptible to early return to the ER room with sepsis symptoms.

There are two main limitations to the proposed approach: (1) domain knowledge is required to define actions and states, and (2) the approach cannot confirm that the relations are causal relations. As for the first limitation, the technique is not automated in the sense that it can detect states and actions. Ideally, this would be the case as it would decrease the burden on the involvement of domain experts. A brute force approach is an alternative, here one would try and map the activities in the process in all possible combinations of states and actions. However, this would severely escalate the multiple comparisons problem.

The second limitation concerns the fact that the approach discovers potential causal relations. It shows that there is a relation between a state and an action but cannot guarantee that this relation is causal. In order to do that, a number of others steps need to be taken. For example, possible confounding variables need to be checked and a number of other independence criteria need to be met (i.e. exchangeability, positivity, and consistency) (Bozorgi et al., 2020).

In future work, we will focus on the notion of causality by including additional tests and checks into the approach such that it can be automatically detected. In addition, we will address the problem of automatically detecting states and actions from a given event log.

CHAPTER 6

Mining Statistical Relations for Better Decision Making in Healthcare Processes

Reading Guide. In the previous chapter, we present a state-action extension to the ARE miner to make the technique more generic. In the second extension in this part of the dissertation, we present a work in which we look into confounding variables. Confounding variables are variables that interfere with the discovered dependency patterns by offering alternative explanations. In this chapter, we propose a technique that helps detect, display, and deal with potential confounding variables. Understanding and handling alternative explanations is a crucial step during the study of potential dependency patterns. This second extension also concludes the second part of the dissertation.

This chapter is based on the following work:

Koorn, J. J., X. Lu, H. Leopold & H. A. Reijers (2022b), "Mining statistical relations for better decision making in healthcare processes".

6.1 Introduction

In many professional settings, taking the right actions often results in whether the desired outcome can be obtained or not. In a healthcare context, such desired outcomes include improving symptoms of a patient, curing a certain disease, or plainly saving a patient's life. Given the considerable impact of certain actions in healthcare settings, there is a large desire to better understand how taking (or not taking) particular actions is linked to such outcomes (Brookhart et al., 2010; Deng et al., 2013). The data that is collected by modern Health Information Systems (HISs) provides a valuable basis to study and understand this link. Among others, HISs may record which actions have been taken, when these actions have been taken, and who was involved (Rojas et al., 2016; Van der Aalst, 2016). Nonetheless, the link between actions and outcomes is inherently complex and may depend on a large number of contextual factors (Deng et al., 2013).

To illustrate the complexity of the link between actions and outcomes, consider a residential care facility where patients with different intellectual disabilities live together. One of the main objectives of such a facility is providing the patients with the best possible quality of life and, therefore, preventing instances of aggressive behavior (Lloyd & Kennedy, 2014). If aggressive behavior occurs, caretakers face a number of different options for responding to such an incident. Potential responses range from mild measures, such as simply talking to the patient, to severe measures, such as isolating the patient. Whether the chosen response was effective with respect to preventing further aggressive incidents in the future is a complex question. The effectiveness of the response might be affected by certain patient characteristics, the type and severity of the incident, and other contextual factors. It is, for instance, well imaginable that a certain response can indeed lead to a lower number of aggressive incidents in the future, but only for incidents with a certain severity. If the severity is very high, it may not matter which response the caretaker chooses. This example illustrates how complex the link between actions and outcomes can be in real-life healthcare situations. Nonetheless, if it was possible to identify the complex relations in this setting and, in this way, reduce the number of aggressive incidents in the future, this could greatly improve the quality of life for patients in the facility.

The case of aggressive behavior highlights the importance of understanding the link between actions and outcomes in healthcare processes. Researchers have proposed various techniques that aim to discover such relations. The most prominent ones can be found in the literature on process mining in healthcare. They can be subdivided into two main groups. The first group of techniques leverages technology from the area of machine learning and artificial intelligence. For instance, Bozorgi et al. (2020) propose an action-rule-based technique in which they use uplift trees to determine for which client groups a particular treatment has a causal effect. The second group of techniques takes a statistical perspective. For example, Brunk et al. (2020) look at the context in which a process takes place and runs a Dynamic Bayesian Network model to determine causal relations. Koorn et al. (2020) propose a statistical mining technique in which they aim to uncover which measures from caretakers have a statistically significant effect with respect to desired outcomes.

While all these techniques provide insights into statistical relations between actions

and their outcomes, they do not properly address the problem of confounding variables, i.e., variables that explain the dependent variable better than the considered set of independent variables. Existing techniques either 1) do not account for confounding variables at all (Koorn et al., 2020) or 2) they do not provide insights into the confounding variables (Bozorgi et al., 2020). Not *accounting for confounding variables* can lead to a flawed understanding of the underlying process and, therefore, can lead to poor decision making (Nørgaard et al., 2017). In the example introduced above, the severity of an incident represents a confounding variable since it can explain the occurrence of aggressive incidents in the future (the dependent variable) better than the chosen response (the independent variable). From a decision-making perspective this means that, based on the severity, different responses need to be taken. Not *providing insights into confounding variables* can lead to similar problems since the user does not understand the possible effects of environmental or contextual factors on the discovered relations. As pointed out by Shortliffe (2012), this may become a considerable problem since decision-makers in the medical domain tend to only trust recommendations and predictions if they understand how they were generated.

Against the background of this research gap, we use this chapter to propose a novel relation mining technique for healthcare processes that 1) explicitly accounts for confounding variables and 2) transparently communicates the effect of the confounding variables to the user. To this end, we adopt a process mining perspective and consider healthcare processes as a series of events. More specifically, we consider a scenario where a decision-maker needs to choose from a set of responses given a particular action and we would like to understand whether these responses can be related to future actions, so-called effects. To develop and illustrate our conceptual ideas, we build on the problem of *aggressive behavior in residential care facilities*. As outlined above, in such a setting, we would like to understand whether a certain response from a caretaker (e.g., isolating a client) to a patient's action (e.g., aggressive behavior towards people) is linked to a patient's action in the future (e.g. no further aggressive behavior). On a technical level, we combine process mining techniques with advanced statistical testing. As a result, our technique can identify complex hidden relations that can be visualized in an understandable and transparent fashion. By doing so, we pave the way for a better understanding of the complex relations in healthcare processes and, in this way, make better decisions.

The rest of the chapter is organized as follows. In the background section, we introduce the technology of process mining and we elaborate on the technical and statistical aspects required for mining statistical relations. In the technique section, we define and explain the technical details of our technique. In the application section, we demonstrate the applicability of our technique using real-world data set relating to aggressive incidents in a residential care facility. Finally, we discuss the theoretical implications and limitations of our technique in the discussion section before we conclude the chapter.

6.2 Background

In this section, we discuss the technical and theoretical background of our work. First, we introduce the technology of process mining and explain why and how it is used.

Second, we introduce the notion of an action-response-effect log as the specific input our technique builds on. Third, we explain how we can detect statistically significant relations in action-response-effect logs.

6.2.1 Process mining

Process mining is a family of data analysis techniques that aims to discover, monitor, and improve organizational processes by analyzing data from so-called event logs (Van der Aalst, 2016). These event logs are generated by various information systems that are used in organizations and, therefore, capture how organizational processes are executed. Process mining provides companies with the means to achieve various objectives such as obtaining a better understanding and control of their processes (Li et al., 2008). Van der Aalst et al. (2011) distinguishes three forms of process mining: discovery, conformance checking, and enhancement. *Discovery* techniques are used to derive the control flow of a process. Their aim is to generate process models without any prior information. These process models often determine the starting point of process mining projects. With *conformance checking*, organizations can compare the intended process execution with the actual process execution to detect unknown structural deviations. Thirdly, *enhancement* extends existing process models with additional information by building upon process execution data. This includes the use of timestamps to highlight, for example, bottlenecks or throughput times. Process mining has been applied in a variety of settings, but a large strand of research focuses on the healthcare domain. Process mining has been applied in various healthcare settings (Rojas et al., 2016). Among others, process mining has been used to analyze patient care processes (Fei, Meskens, et al., 2010; Kim et al., 2013), dentistry processes (Bakhshandeh et al., 2017; Mans et al., 2012a), and cancer treatment processes (Binder et al., 2012). Given the frequent and successful use of process mining to analyze and understand healthcare processes, we also build on process mining technology to address the problem of relation mining in this chapter.

6.2.2 Action-response-effect logs

The *event logs* used by traditional process mining techniques are a collection of sequences of events. Each event is structured in the same way to record information such as the corresponding case, the corresponding task, the time of the execution, and the user (resource) who executed the task. A sequence of events that correspond to a particular case is called a *trace*. A trace is the lifecycle of one particular case. In the context of this chapter, we consider a specific type of event log: an *action-response-effect log*. It differs from a traditional event log in the sense that there are different types of events that are (casually) linked. To illustrate the notion of an action-response-effect log, we use the example of aggressive behavior in residential care facilities. Formally, an action-response-effect log L is a specific type of event log, where each event contains a case identifier (e.g., client id), an *action* (e.g., "Verbal aggression (VA)"), a *response* taken towards the action (e.g., "Distract Client"), and a follow-up *effect* (e.g., "physical aggression against object (PO)").

Table 6.1 exemplifies such an action-response-effect log, where each row records

ID	Event ID	Timestamp	Action	Response(s)	Effect	Severity
1	e_1	12-05 09:53	VA	Warning	PO	5
1	e_2	13-05 13:35	PO	Distract Client, Seclusion	τ	9
1	e_3	26-05 09:32	VA	Warning	PP	6
1	e_4	26-05 11:02	PP	Distract Client	τ	7
2	e_5	21-06 14:51	VA	Distract Client	VA	7
1	e_6	23-06 21:23	VA	Distract Client	τ	6
2	e_7	24-06 17:02	VA	-	τ	3
3	e_8	29-07 11:22	VA	Warning	PO	4
3	e_9	31-07 08:13	PO	Warning, Seclusion	PP	9
3	e_{10}	31-07 10:48	PP	Distract Client	τ	8

Legend: VA = Verbal Aggression, PP = Physical Aggression (People), PO = Physical Aggression (Objects)

Table 6.1: Excerpt of an action-response-effect log

an occurred event. The column "Action" indicates the action of the event, the column "Responses" lists the responses to the event, and the column "Effect" the follow-up effect. We define a function π_r to return the set of response events $\{r_1^e, \ldots, r_n^e\}$ of an event e; we write $\pi_r(e) = \{r_1^e, \ldots, r_n^e\}$. For each trace $\sigma = \langle e_1, \ldots, e_n \rangle$, the sequence of responses is $\langle \pi_r(e_1), \ldots, \pi_r(e_n) \rangle$. For example, in the action-response-effect log listed in Table 6.1, for event e_1: $\pi_c(e_1) = 1$ is the case of event e_1, $\pi_l(e_1)$ = "Verbal Aggression" (VA) is the action of e_1, and $\pi_r(e_1) = \{$"Warning"$\}$ is the set of responses of e_1.

For each trace $\sigma = \langle e_1, ..., e_n \rangle$, we define the effect for each e_i, where $1 \leq i < n$ as follows: if the elapsed time to the next event e_{i+1} is less than ϵ, the effect $\pi_{next}(e_i)$ of e_i is the action of e_{i+1}, else we say that the effect is a silent action τ. Formally, if $\pi_{time}(e_{i+1}) - \pi_{time}(e_i) \leq \epsilon$, then $\pi_{next}(e_i) := \pi_l(e_{i+1})$, else $\pi_{next}(e_i) := \tau$. An approach to convert a traditional event log into an action-response-effect log has been described in Koorn et al. (2020).

6.2.3 Mining action-response-effect patterns

Given an action-response-effect log, it is possible to mine relevant action-response-effect patterns (Koorn et al., 2020). Such patterns reveal, given a particular action, which relations between responses and effects are statistically significant. While the existing notion of action-response-effect patterns introduced by Koorn et al. (2020) does not take confounding variables into account, we build on this existing notion in this chapter. Below, we explain how we can mine action-response-effect patterns from an action-response-effect log using the chi-square test.

The chi-square test principally compares the observed frequencies to expected frequencies as calculated based on the chi distribution. The chi-square test is subject to a number of assumptions, which are described in detail in (McHugh, 2013). To test the hypothesis of whether an effect is independent of the response to an action, the number of observed events is compared to the number of expected events of different responses and effects. To calculate the number of observed events, we create a matrix (table) where each cell is filled with the number of observed events of a re-

Observed	**PO**	**PP**	**VA**	**SIB**	τ	*Total*
Warning	250	400	200	100	50	*1000*
Held with force	20	50	50	20	10	*150*
Seclusion	30	50	20	10	10	*120*
Terminate contact	100	100	90	60	10	*360*
Distract client	100	150	40	40	20	*350*
Total	*500*	*750*	*400*	*240*	*100*	**1990**
Expected	**PO**	**PP**	**VA**	**SIB**	τ	*Total*
Warning	252.5	378.8	202.0	116.2	50.5	*1000*
Held with force	37.9	56.8	30.3	17.4	7.6	*150*
Seclusion	30.3	45.5	24.2	13.9	6.1	*120*
Terminate contact	90.9	136.4	72.7	41.8	18.2	*360*
Distract client	88.4	132.6	70.7	40.7	17.7	*350*
Total	*500*	*750*	*400*	*240*	*100*	**1990**

Legend: VA = Verbal Aggression, PP = Physical Aggression (People), PO = Physical Aggression (Objects,)SIB = Self-Injurious Behavior

Table 6.2: Excerpt of the tables used to perform high-level statistical tests; horizontal categories: effect, vertical categories: response

sponse and an effect. Let $a \in A$ be an action, $R = \{r_1, \cdots, r_m\}$ be a set of responses, and $C = \{c_1, \cdots, c_n\}$ a set of effects. We define a $|R| \times |C|$ matrix; where each row represents a response r_i, each column represents an effect c_j, and each cell counts the number of observed events that have response r_i and effect c_j. We then have

$$freq_{a,R,C} = \begin{pmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,n} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{m,1} & f_{m,2} & \cdots & f_{m,n} \end{pmatrix}$$

where

$$f_{i,j} = freq_L(a, r_i, c_j) = \quad |\{e \in L \mid \pi_l(e) = a \wedge r_i \in \pi_r(e) \wedge \pi_{next}(e) = c_j\}| \tag{6.1}$$

For instance, given a log L as listed in Table 6.1, $freq_L$("VA","Warning","PO") =$|\{e_1, e_8\}|= 2$. Considering Table 6.2 and omitting the column totals and row totals, it exemplifies a matrix $freq_{a,R,C}$. If the effects are independent of responses, then we expect to observe that the distribution of effects of a response is similar to the *total distribution*.

Each row r_i presents the distribution of effects $c_1, ..., c_k$ to the response r_i. To test whether each individual response r_i has an influence on the effects, we define $freq_{a,r,C}$ as a $2\times |C|$ matrix:

$$freq_{a,r,C} = \begin{pmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,n} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,n} \end{pmatrix} \tag{6.2}$$

where $f_{1,j} = freq_L(a, r, c_j)$ and $f_{2,j}$ is the frequency distribution of effects of the responses other than r, i.e., $f_{2,j} = \quad |\{e \in L \mid \pi_l(e) = a \wedge r \notin \pi_r(e) \wedge \pi_{next}(e) = c_j\}|$.

Observed	**PO**	**PP**	**VA**	**SIB**	τ	*Total*
Terminate contact = 0	300	500	210	180	90	*1280*
Terminate contact = 1	100	100	90	60	10	*360*
Total	*400*	*600*	*300*	*240*	*100*	*1640*
Expected	**PO**	**PP**	**VA**	**SIB**	τ	*Total*
Terminate contact = 0	312.2	468.3	234.1	187.3	78.0	*1280*
Terminate contact = 1	87.8	131.7	65.9	52.7	22.0	*360*
Total	*400*	*600*	*300*	*240*	*100*	*1640*

Legend: VA = Verbal Aggression, PP = Physical Aggression (People), SIB = Self-Injurious Behavior, PO = Physical Aggression (Objects)

Table 6.3: Excerpt of the tables for an individual response used to perform statistical tests; horizontal categories: effect

An example of $freq_{a,r,C}$ where r is "Terminate contact" is listed in Table 6.3. Performing a chi-square test allows us to calculate the expected values and test the statistical dependency between responses and effects. The chi-square test compares the observed frequencies to the expected frequencies. If they differ significantly, then the null hypothesis is rejected. This means we cannot rule out that there is a statistical dependency relation between the response and the effect.

In line with the preliminaries, existing techniques aim to uncover statistical relations in a more process-oriented manner. These techniques uncover interesting patterns that are useful for understanding a healthcare process such as aggressive behavior. However, the techniques suffer from an important limitation. In order to better understand these potential causal relations, one needs to consider and explain the potential effect that a confounding variable can have on the discovered relationships. This chapter proposes a technique that can detect and visualize confounding variables.

6.3 Technique

In this section, we present our technique for relation mining in healthcare processes. As illustrated in Figure 6.1, our technique consists of three main steps: (1) statistical tests, (2) scenario determination, and (3) graph creation. As input, it requires an action-response-effect log, a set of potentially confounding variables, and two specific parameters. As output, it provides the user with a number of graphical representations illustrating the identified relations. Below we elaborate on the inputs, the three conceptual steps, and the output in detail.

6.3.1 Input

The inputs for our technique consist of three main elements: (1) an *action-response-effect* log, (2) data on the candidate confounding variable, and (3) parameters k and M. Since we already introduced (1) in the background, we now focus on (2) and (3). In general, the data that is used for the candidate confounding variable can be any

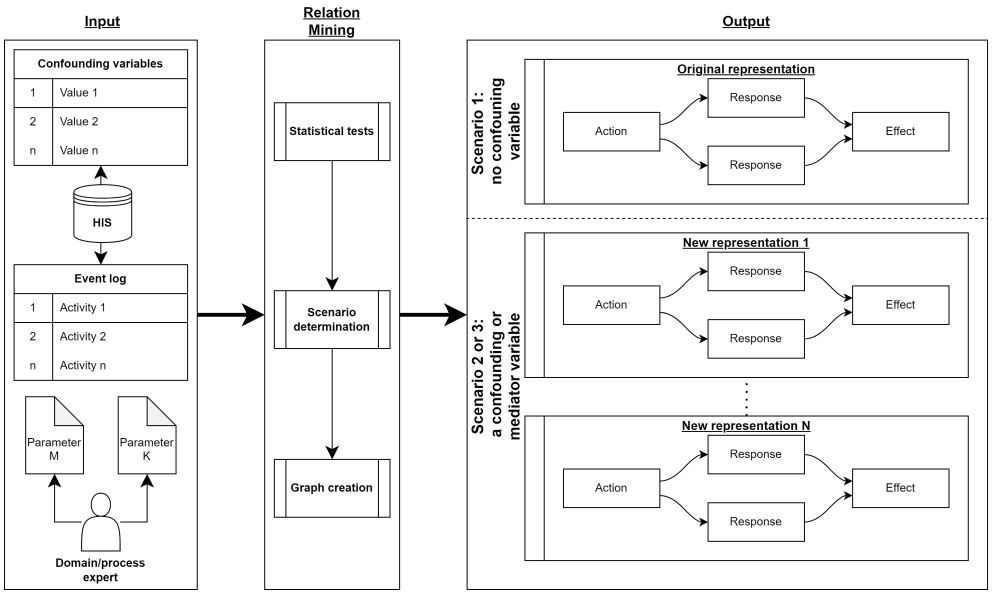


Figure 6.1: A visual representation of the proposed technique

attribute in the data set that is measured in the same fashion as the action, responses, and effects. In Table 6.1 this is showcased by the last column *Severity*. The data can be of both categorical or continuous nature. In the proposed technique, we use a number of parameters to determine the strategy for detecting a possible confounding variable. First, we use k to denote the number of stratified data sets that we create. By default, this parameter is set to a value of 2. Simply put, if the candidate variable consists of two categories (e.g. *aggression history* where we have yes or no), we split the data into two data sets. One set only contains clients that have a history of aggressive behavior and one set only contains clients that were never aggressive before.

Next to that, we need a strategy to split the data set into multiple subsets. This holds for candidate variables that have more than 2 categories and for numerical variables. Parameter M determines this strategy. Two main approaches can be taken: (1) use of domain knowledge, or (2) use of an automated approach. The first case is preferred since the interpretation of the outcomes is highly dependent on domain knowledge. However, if this is not feasible, there are approaches to automate the splitting of data into subsets. For this technique, it is most suitable to split the data such that the observations are distributed equally over the categories (i.e. subsets). This minimizes the chances that the assumptions for the statistical test are violated. The maximum number of data subsets (k) can then be expanded in an iterative way, keeping the size of each k equal as long as the test assumptions are met.

6.3.2 Relation mining

The relation mining is the core of our technique and consists of three specific steps: (1) statistical tests, (2) scenario determination, and (3) graph creation.

Statistical tests

The goal of performing the statistical tests is three-fold: (1) determine for which actions there exist statistical relations between response and effect, (2) determine which exact responses relate to the effect, and (3) test how the exact relations hold when we introduce the confounding variable.

When we look for confounding variables, we first check the chi-square test score for the original data set. The original data set here contains the action-response-effect log that is described in the preliminaries section. The chi-square test takes this log and calculates for each entry the expected values based on the chi-distribution. Then, the observed frequencies from the original log and expected frequencies based on the chi-distribution are compared to each other and a test score is calculated. This score indicates how large the difference between these scores is on an aggregate (log) level. If the test score is significant, this means there is at least one significant combination of response type and effect type. As such, we can conclude that there is a relation between the response and effect variable. If the test score is insignificant, this means that there is no relation between the response and effect variable. The exact mathematical approach of the chi-square in the action-response-effect context is described in the preliminaries section.

If the chi-square test score for the original log is insignificant, no further testing is performed. If the chi-square test score for the original log is significant, each individual response variable is tested in a similar fashion to determine which response type has a relation with the effect variable. This is done by performing a chi-square test for each response type. If these scores are significant that means that the specific response type is related to the effect. If the test score is insignificant, no relation between that particular response type and the effect exists.

Finally, we introduce an extra variable, the possible confounding variable. We refer to it as the *candidate variable* from here on. The candidate variable is used to stratify the original data into k smaller data sets. We refer to these as *subsets*. Following this, we perform the chi-square test on each of the subsets. We refer to respective results as the *subset test scores*. The test scores for each subset can be significant or insignificant, again indicating whether or not there is a relation between the response and the effect variable.

Scenario determination

At this point, there are three possible scenarios that can occur: (1) the original test score and subsequent subset test scores all have the same results, (2) the original chi-square test has the opposite results in comparison to each individual subset test scores, or (3) the original chi-square test equals at least one and opposes at least one subset test score. In Table 6.4 we present the different scenarios.

Below, we go into detail on each of the scenarios. First, we explain the theoretical rationale. Next, we exemplify the interpretation using a running example. For the running example, we take the type of response taken by the nurse (e.g. terminate contact) as the independent variable and the type of future aggressive behavior of the client (e.g. physical aggression) as the dependent variable. The candidate variable we investigate is *severity of the incident* (score between 1 and 10), where we split the category into two: severe incidents (score ≥ 7), and mild incidents (score 1-6). In

Original data	Subset	Confounding variable
Significant	All significant	No confounding variable
	Mixed results	Mediator variable
	All insignificant	Confounding variable
Insignificant	All significant	Confounding variable
	Mixed results	Mediator variable
	All insignificant	No confounding variable

Table 6.4: Scenarios for test results for confounding variable. The (in)significant values refer to the chi-square test score(s)

Observed	**PO**	**PP**	**VA**	**SIB**	τ	*Total*
Terminate contact = 0	100	160	75	60	30	*425*
Terminate contact = 1	30	40	30	20	3	*123*
Total	*130*	*200*	*105*	*80*	*33*	*548*
Expected	**PO**	**PP**	**VA**	**SIB**	τ	*Total*
Terminate contact = 0	100.8	155.1	81.4	62.0	25.6	*425*
Terminate contact = 1	29.2	44.9	23.6	18.0	7.4	*123*
Total	*130*	*200*	*105*	*80*	*33*	*548*

Legend: VA = Verbal Aggression, PP = Physical Aggression (People), SIB = Self-Injurious Behavior, PO = Physical Aggression (Objects)

Table 6.5: Excerpt of the tables for an individual response used to perform statistical tests. This table represents the group of data with a mild severity score (1-6); horizontal categories: effect

Table 6.5 and Table 6.6 we display how the data from Table 6.3 could be stratified based on the severity score. As becomes clear, the original table is split based on the severity score of each incident into two tables. Note that the numbers of both Table 6.5 and Table 6.6 add up to the numbers in Table 6.3.

Scenario 1: no confounding variable

In the first scenario, we find that all test results are either significant or insignificant. If all test results are significant, there is a relation between the independent and dependent variable, regardless of the candidate variable. If all test results are insignificant, there is no relation between the independent and dependent variable, regardless of the candidate variable. In both cases, the candidate variable has *no effect* on the relationship between the independent and dependent variable. Thus, we can conclude the candidate variable is not a confounding variable.

In terms of the running example, suppose we initially find a significant relationship between the type of response and type of future aggression. For example, responding to aggression by terminating contact with a client leads to more verbal aggression in the future. If we then split the data set based on severity and we get significant scores for both subsets, we can conclude that regardless of the severity of the incident this relation holds. If the initial finding is that there is no significant relation, then we find no effect of terminating contact with a client on the type of future aggression. If all subset test scores are also insignificant, this shows that the relationship does not change if we consider only severe or only mild incidents. Based on this, we

Observed	**PO**	**PP**	**VA**	**SIB**	τ	*Total*
Terminate contact = 0	200	340	135	120	60	*855*
Terminate contact = 1	70	60	60	40	7	*237*
Total	*270*	*400*	*195*	*160*	*67*	*1092*
Expected	**PO**	**PP**	**VA**	**SIB**	τ	*Total*
Terminate contact = 0	211.4	313.2	152.7	125.3	52.5	*855*
Terminate contact = 1	58.6	86.8	42.3	34.7	14.5	*237*
Total	*270*	*400*	*195*	*160*	*67*	*1092*

Legend: VA = Verbal Aggression, PP = Physical Aggression (People), SIB = Self-Injurious Behavior, PO = Physical Aggression (Objects)

Table 6.6: Excerpt of the tables for an individual response used to perform statistical tests. This table represents the group of data with a severe severity score ($\geq$ 7); horizontal categories: effect

can conclude that the severity of the incident is *not* a confounding variable in both examples. Apparently, the severity of the incident does *not* influence the relation between type of response and type of future aggression in this scenario.

Scenario 2: a confounding variable

In the second scenario, we find that the original test result shows the opposite (in terms of significance) compared to each of the subset test scores. In this scenario, there is a significant effect of the candidate variable on the relationship between the independent and dependent variables. In the case that the original test score is significant and the subset test scores are insignificant, the test results show that the hypothesized relation between the independent and dependent variables disappears. In the opposite case, where the original test score is insignificant but the subset test scores are significant, the results show that under certain circumstances (i.e., those captured in the candidate variable) there is a relationship that is hidden when all data is combined. Hence, the hypothesis that there is no relation between the independent and dependent variables is falsified. Thus, in both cases we can conclude that the candidate variable is a *confounding variable*, after all, the candidate variable presents *an alternative* explanation to the independent variable for the dependent variable.

Returning to the running example, we would find a significant relationship between the type of response and type of future aggression. Let us consider talking to the client as a type of response and verbal aggression as a future aggression type. If we split the data using the severity score, and run the chi-square test on each subset, we then find insignificant chi-square test scores for both subsets. The interpretation is as follows: There seems to be a relation between talking to the client and future verbal aggression. However, if we separately consider severe and mild aggressive incidents, the relation disappears. In the opposite case, we would initially find no relation between talking to the client and future verbal aggression (insignificant chi-square test score). However, when we split the data based on severity score we would see that for mild and severe incidents separately, we do see the relation. In both cases, we see that the severity score has an impact on the relationship between talking to the client and future verbal aggression. As such, we conclude that the severity

of the incident is a *confounding variable*; it explains the dependent variable (future aggression type) better than the independent variable (response type).

Scenario 3: a mediator variable

In the final scenario, we find mixed results in the subset test scores, regardless of the original test score. In this scenario, if the original test score is significant we hypothesize that there is a relation between the independent and dependent scores. However, the subset test scores show that this relation only holds under certain, yet not all, circumstances captured in the candidate variable. In other words, the candidate variable mediates the relation between the independent and dependent variables. Thus, we can conclude that the candidate variable is a specific type of confounding variable. It has an influence but does not provide a complete alternative explanation for the relation between the independent and dependent variables. If the original test score is insignificant, we hypothesize that there is no relation between the independent and dependent variables. However, based on the sub-test scores we can conclude that in certain circumstances, captured in the candidate variable, the relation does exist in some contexts. The relation disappears when all data is combined. Hence, the candidate variable has an effect on the relation between the independent and dependent variables. Thus, when we find mixed results in the subset test scores, we can conclude that the candidate variable is a specific type of confounding variable which we refer to as the *mediator variable*.

In this scenario, we would initially find a significant relation between terminating contact with a client and future verbal aggression. This is the case in Table 6.3. However, when we introduce the severity score again, the chi-square test for incidents with a high severity score results in an insignificant test score (Table 6.6), but for mildly severe incidents we do find a significant test score (Table 6.5). What we can also see from the expected frequencies in both tables is that the distribution between applying the treatment or not has changed. For example, for the effect PP in Table 6.5 we see a 20%-80% distribution (n = 160/40) for applying the effect or not. In Table 6.6 this distribution shifts to 15%-85% (n = 340/60). As a result, we find mixed results in the subset tests.

From the mixed subset results we conclude that if the severity score of an incident is high, it does not matter if we terminate contact with the client in terms of the type of future aggressive behavior. By contrast, if the severity of an incident is mild, terminating contact with the client has an effect on the type of future aggressive behavior. The same reasoning goes when the initial finding is insignificant. As a result, we can conclude that the severity score is a confounding variable in both cases as it *influences* the relation between type of aggression and type of future aggression. Note that this is different from scenario two in the sense that it only influences the relation under specific circumstances, as such we refer to this "type" of confounding variable as the mediator variable.

Graph creation

The goal of the graph creation is to visualize the mined relations. To this end, we build on the visualization mechanisms proposed in the context of the ARE miner (Koorn et al., 2020). In principle, a detailed statistic called the adjusted standardized residual (ASR) is calculated for all possible combinations of actions, responses, and

effects. The ASR is used in three ways: (1) its significance, (2) its value, and (3) its size. The *significance* of the ASR determines which edges need to be drawn out of all possible combinations of nodes. Then, based on the *value* of the ASR, either a solid arc (positive values) or dotted arc (negative values) is drawn. Thirdly, the *size* of the ASR value determines the thickness of the arcs, there are six size classes in total, three for positive and three for negative. The reasoning is, that the thicker the arc, the larger the effect size it represents. Finally, the observed and expected frequency are displayed on each arc to provide the user with additional concrete insights into the statistical relations.

6.3.3 Output

The *output* of our technique is a set of mined relations visualized in the context of graphical representations. When the technique detects a confounding variable, there are three scenarios: (1) The original test score is significant and the subset test scores are significant. In this case, this technique produces the same graphical representation as to the ARE miner, see Figure 6.2 as an example. (2) The original test score is insignificant, but the subset test scores are significant. The technique only returns the graphical representations for each of the stratified data sets with a significant test score. In this case, the severity of the aggressive incident is tested as a confounding variable. We produce one graphical representation for highly severe incidents (e.g. score ≥ 7) and one graphical representation for less severe incidents (e.g. score < 7). (3) There are mixed results in the subset test regardless of the original score. In this case, the technique produces only the graphical representations for the subsets with a significant test score (see Figure 6.3 and Figure 6.4 as examples of the output). Finally, if the original score is insignificant and all sub-test scores are also insignificant, no graphical representation is generated as there are no statistical relations.

6.4 Application

In this section, we show the applicability of the proposed technique. First, we introduce the data from a real-world case study. Then, we introduce the candidate variable: severity of the incident. Finally, we elaborate on the results and insights we can obtain when we apply the proposed technique to the data.

6.4.1 Case study: data set

To evaluate the applicability of our technique to detect confounding variables we use a real-world data set related to the care process of a Dutch residential care facility. The event log contains 21,384 recordings of aggressive incidents from 1,115 clients. The process captured in this log concerns the aggressive behavior of clients in their facilities and the way client caretakers respond to these incidents. The log consists of aggressive incidents of clients that belong to one of four different action classes. Each of these actions is followed by a number of measures from the caretakers as responses to the action. Each response belongs to one of nine different response classes. We transformed this log into an action-response-effect log by defining the next aggressive incident of a client as an effect if it occurred within 9 days. Otherwise, the link is not

Actions	Physical aggression towards people	11,381
	Physical aggression towards objects	1,446
	Verbal aggression	5,778
	Self-injury	2,779
	Total	*21,384*
Responses	Talk to client	9,279
	Held with force	3,624
	Leave room	3,638
	Distract client	2,561
	Send away	3,169
	Seclusion	1,156
	Other measures	209
	None	783
	Ignore client	70
	Total	*24,489*
Effects	Physical aggression towards people	5,897
	Physical aggression towards objects	686
	Verbal aggression	2,369
	Self-injury	1,429
	No next incident (τ)	9,888
	Total	*20,269*
Clients	Minimum number of actions per client	1
	Maximum number of actions per client	449
	Average number of actions per client	19.2
	Total	*1,115*

Table 6.7: Overview of the characteristics of the real-world data set

considered and the effect is defined as τ, i.e. none. This transformation procedure is in line with the approach followed by Koorn et al. (2020). As a result, we obtain a total of five different effect classes. Table 6.7 summarizes the characteristics of our data set.

6.4.2 Case study: the candidate variable

To put the results produced by our technique into a context, it is interesting to compare them to those from the ARE miner (Koorn et al., 2020), the technique that is most similar to ours. Among others, the ARE miner uncovered that responding to verbal aggression with physical restraints (i.e. seclusion of client) leads to an increased chance of escalation of future violence (i.e. more violence towards persons). However, the ARE miner does not account for candidate variables that can influence the discovered relations. In this work, we build on this notion and check for possible confounding variables. Based on insights from the healthcare organization and aggression literature one promising candidate variable is identified: the severity of an aggressive incident.

Previous research has linked the severity of an incident to the well-being of both client and caretakers (Hastings, 2002). Well-being in the context of aggressive behav-

ior is strongly determined by the stress that is experienced by the people involved in the aggressive incidents. Aggressive behavior can have serious negative consequences for clients, e.g. more coercive measures, physical injuries, or exclusion (Nieuwenhuis et al., 2017; Van den Bogaard et al., 2018). These factors all increase the stress experienced by clients. Research shows that clients who experience more stress display different kinds of aggressive behavior (Brosnan & Healy, 2011).

On the other hand, research shows that support staff who experience high levels of stress due to aggressive behavior of their clients are likely to respond in different ways (Hastings, 2002; Zijlmans et al., 2012). The negative experience of an aggressive incident can translate to both under- or overestimating the severity of an incident. Underestimating the severity can result in a minimized perception of the risks and attention required to deal with the behavior. Overestimating can lead to the unnecessary seclusion/isolation of clients with all negative consequences associated with these measures such as violation of autonomy and respect (Huckshorn, 2004; Noda et al., 2012).

In summary, we see that the severity of incidents influences aggressive behavior in two ways. First, it affects the behavior of clients. Second, it has an impact on the way caretakers handle aggressive behavior. Given this, we hypothesize that the severity of an incident is a confounding variable for the relationship between the response of the caretaker and the future aggression of clients.

6.4.3 Case study: results

The first step of the technique is to determine how the data is stratified. In our case, we stratified the data based on domain knowledge. *Severity* refers to the gravity of the incident as indicated by the caretaker on a 1-10 scale. Based on this knowledge, two subsets are created. Thus, our k parameter equals the default value ($k = 2$). One subset is created for mild incidents where we define mild as an incident with a score between 1 and 7. Another subset captures severe incidents, where we define severe as a score ≥ 7.

Recall that the relation miner produces results for response-effect combinations for a given action. As such, a substantial amount of tables and graphical representations were produced when checking for the confounding variables. To keep the results digestible we selected an exemplary case. In Table 6.8 we highlight the results for the initial action self-injurious behavior where we check for the candidate variable severity of the incident. In Figure 6.2 we display the original results for the action self-injurious behavior. Subsequently, Figure 6.3 and Figure 6.4 show the results when the data is stratified based on the severity of the incident. Here, Figure 6.3 shows the results for mild incidents (score 1-6) and Figure 6.4 shows the results for severe incidents (score ≥ 7). In the next section, we go into detail on the findings and show how each of the scenarios described in the technique section can be found in the results of this case study. As can be seen for one response (seclusion) no test could be performed on the subsets; this has to do with the prerequisites of the chi-square test. We will get to this in the limitations.

Response type	Original set	Subset 1: severity <7	Subset 2: severity ≥ 7	Scenario
Terminate contact	Significant	Significant	Significant	1
Send to other room	Significant	Insignificant	Insignificant	2
Distract client	Significant	Significant	Insignificant	3
Talk to client	Significant	Significant	Significant	1
Seclusion	Insignificant	Not enough data	Not enough data	-
Hold with force	Significant	Significant	Significant	1
No measure	Significant	Significant	Significant	1

Table 6.8: Results of candidate variable tests for the variable severity of incident. The action is self-injurious behavior. The significance scores refer to the result of the chi-square score(s)

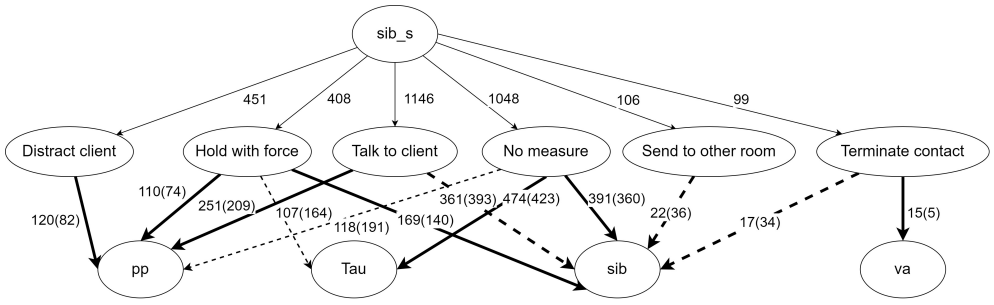


Figure 6.2: Graphical representation for action self-injurious behavior for original data set

No confounding variable

Recall that scenario 1 from the previous section refers to the category where there is no confounding variable. In Table 6.8 we see that a number of responses fit this description. For example, *Terminate contact* as a response has a significant effect on the future aggression after a self-injurious behavior incident. We can observe this finding in the graphical representation based on the original results in Figure 6.2, the node in the middle row with *terminate contact*. Here, we can see that based on the original data, terminating contact with the client increases the chances of a next

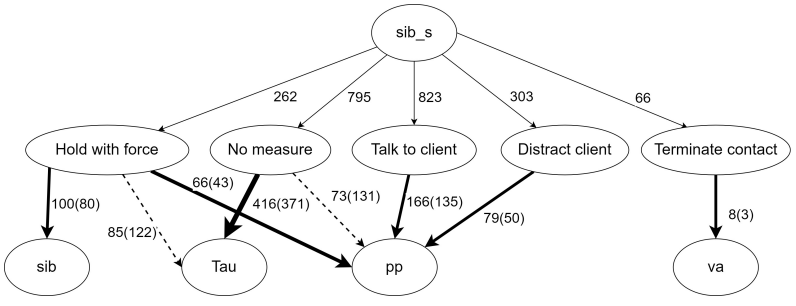


Figure 6.3: Graphical representation for action self-injurious behavior for mild incidents (score 1-6)

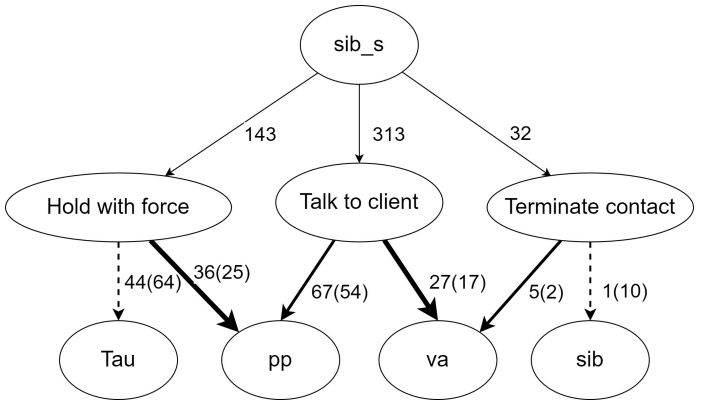


Figure 6.4: Graphical representation for action self-injurious behavior for severe incidents (score ≥ 7)

incident of the type of verbal aggression (VA) and reduces the chances of another self-injurious behavior type of future incident (SIB).

When the data is then stratified and tested again, we see that the chi-square results of the subset tests match that of the original test (see Table 6.8). Thus, the pattern of response and effect that we observed in the original data *remains the same* for each of the subsets. We can see this in the graphical representations wherein the mild incidents (Figure 6.3) the response *terminate contact* increases the chances of verbal aggression as future aggression. In addition, in Figure 6.4 we observe the same relation to verbal aggression and we see that terminating contact reduces the chances of future self-injurious behavior. As such, we can conclude that the relationship between terminating contact with a client and the effects (i.e. future aggressive incidents) holds regardless of the severity of the incident. In other words, severity score is *not* a confounding variable in this context.

A confounding variable

In the second scenario described in the previous section, we find a confounding variable where the subsets all return the opposite test score compared to the test score based on the original data set. In Table 6.8 we see that the response *send to other room* fits this description. In the original data set, we find a significant relationship between this response and the effects. This is visualized in Figure 6.2 where we can observe from the node *Send to other room* that this response reduces the chance of a future repetition of the self-injurious behavior.

When we stratify the data and check for the impact of sending a client to his/her room, we can see that there is no effect of this response on the future aggressive behavior of the client. Visually, we can also observe this in both Figure 6.3 and Figure 6.4. There is no node for sending a client to their room anymore in either figure. This means that the original finding that sending a client to his/her room does not seem to have an effect when we consider the severity of the incident. As such, the severity of the incident is a *confounding variable* in this context and we should disregard this relation from the original finding.

A mediator variable

The final scenario described in the previous section refers to the case where the tests results of the subsets are mixed. In that case, there is a mediator variable. In our specific case, the response *distract client* meets these criteria. The test score for original data shows a significant relation. In Figure 6.2 (most left node *Distract client*) we see that distracting a client results in a higher chance of future violence against another person.

When we then stratify the data based on the severity score we see that this relationship holds for mild incidents, but not for severe incidents. This is also visually clear as we observe the node *distract client* in (Figure 6.3) with the same relation to the effect of physical aggression towards people (PP). In contrast, there is no such node in the graphical representation in Figure 6.4 for severe incidents. The interpretation of the effect of distracting a client is a bit more complex. The results show that if an incident is severe, distracting a client does *not have an effect* on future aggression. In other words, distracting a client is neither harmful nor helpful in this context. By contrast, if incidents are mild in terms of severity, then distracting a client does have an effect. In particular, distracting a client increases the chance of future aggression towards another person in this context. We can conclude that the severity score is a mediator variable as it influences the initially found relation between distracting a client and future aggression. As such, it is very important to use the stratified graphical representations as support when making decisions on whether or not to distract a client.

6.5 Discussion

In this section, we discuss the results of the application. First, we discuss the theoretical implications of this work by relating it to broader (IS) literature. Then, we highlight the two main limitations of this approach: (1) data distribution, and (2) interaction effects.

6.5.1 Theoretical implications

The phenomenon of confounding variables is not unique to the aggressive behavior setting (Nørgaard et al., 2017). It is a concept that emerges whenever statistical relations come into play. Data analytics projects aim to uncover relations and patterns in data that inform academics, practitioners, or the general public on the underlying mechanisms of a certain phenomenon (Ghasemaghaei et al., 2016). The ultimate goal in these projects is often to not only find these relations and patterns but to understand the mechanisms in the data (Bygstad & Munkvold, 2011). Confounding variables form a crucial step in this path from observation to understanding (Greenland et al., 1999). As such, the technique proposed in this research can be used to strengthen research that aims to uncover insights into statistical relations and dependency patterns. Thus, the technique proposed in this research can be generalized in two ways: (1) decision-making, and (2) domain. First, the technique can provide a valuable addition to data science approaches that work on decision-making in health-

care. Second, it can be applied beyond the specific healthcare setting of the case study.

The technique can provide further value to decision-making techniques that do not account for possible confounding variables. Decision-making has been an important topic of discussion on the IS literature agenda for a long time. Gallupe et al. (1986) already pointed out that decision-making processes will increasingly have to be made in a shorter amount of time and based on a larger amount of information. Recent research confirms that decision-making is increasingly leaning on a data-driven approach (Rhyn & Blohm, 2019). One example of current research that works on data-driven decision-making processes in the healthcare realm is Deng et al. (2013). They take a data-driven approach to healthcare decision-making. What becomes apparent when we study these examples in more detail is that decision-making is best supported by having a good data-driven understanding of the (statistical) relations that exist among the factors based on which decisions are made. The technique proposed in this chapter helps understand and visualize these relations and patterns by considering and handling alternative explanations that can influence the outcome.

Besides the healthcare domain, there are other fields in IS that are working on similar problems. Prominent examples from the IS field where statistical relations are also sought are the digital commerce and data science movements. In line with previous work of Bozorgi et al. (2020), uplift trees have gained traction in these fields. For instance, in the digital commerce literature uplift trees have been used to study the way (digital) enterprises can approach their customers best (Gubela & Lessmann, 2020). Another example stems from the analytics and data science field where authors used uplift trees to increase the cost-efficiency of experimentation (Haupt et al., 2019). The uplift tree techniques generally do not provide insights into the mechanisms related to confounding variables. The technique proposed in this research can support users of these techniques to gain a better understanding of the underlying effects that confounding variables have on the discovered relations. With these insights, the user can make better-informed decisions on which actions to perform.

6.5.2 Limitations

There are two main limitations to the proposed technique. First, we observe that in the testing of our case study, a good distribution of the data is key to performing the statistical tests proposed in our technique. Data distribution becomes a limiting factor when the original data is stratified based on a candidate variable that heavily skews the data. The second limitation regards an underlying assumption that is made regarding the independence between the studied (candidate) variables. To give a practical example, the technique builds on the premise that the severity of an incident does not influence the duration of an incident, e.g. incidents that take longer are more severe. These assumptions have been manually checked prior to performing the statistical tests using correlation matrices and showed no signs of (linear) relations.

6.6 Conclusion

In this chapter, we investigate the detection of statistical relations between the responses of an action and the effects of these responses on the process outcome. In particular, we address the challenges of (1) detecting confounding variables, and (2) obtaining concrete insights into their influence on the statistical relations between the actions and the actual outcomes of a care process.

We proposed a novel technique to detect confounding variables and use graphical representations to show their influence on the process outcomes. We applied our technique to a real-world data set from the healthcare domain. The results show that we detect three kinds of confounding variables and explain their impacts on the process outcome in a graphical representation. For example, we not only detect that the severity of an incident is a confounding variable when responding to self-injurious behavior. In addition, we show that when an incident is severe, distracting a client does not have an effect on future aggression, whereas when an incident is less severe, distracting a client does help. In future work, we plan to validate and apply our technique to other healthcare cases. In addition, we will expand on this work by looking into automating the check for interaction effects amongst the confounding variables.

Part III: Addressing Process Mining Evaluations

CHAPTER 7

Bringing Rigor to the Qualitative Evaluation of Process Mining Findings: An Analysis and a Proposal

Reading Guide. The first two parts of this dissertation focus on more in-depth approaches to the discovery of work processes. In the third part, we take a step back and shift our attention to the evaluation of process mining projects. In particular, we are interested in understanding how people perform evaluations in process mining projects especially when external experts (i.e. domain experts) are involved. On top of that, in this chapter, we aim to help people who perform these types of evaluations by proposing a set of guidelines that help improve the validity of these evaluations with domain experts. This chapter concludes the main parts of this dissertation and is followed by the concluding remarks.

This chapter is based on the following publication:

Koorn, J. J., I. M. Beerepoot, V. S. Dani, X. Lu, I. Van de Weerd, H. Leopold & H. A. Reijers (2021), "Bringing rigor to the qualitative evaluation of process mining findings: an analysis and a proposal", in: *2021 3rd International Conference on Process Mining (ICPM)*, IEEE, pp. 120–127.

7.1 Introduction

Process mining is widely used to discover, analyze, and improve business processes in various industries. Given its popularity, several methodologies have been developed to guide both practitioners and academics in performing process mining projects (Emamjome et al., 2019). An important factor in process mining methodologies is the interaction between process analysts and domain experts. This interaction often takes place at the start of a project, during data extraction and pre-processing, and at the end of a project, i.e., during the *evaluation*. The evaluation is an important step in any process mining project as it is concerned with making sure that the findings are actually valid. The domain experts are essential in this context since they are able to assess and interpret the findings and translate them into actionable insights and recommendations. In this way, they make sure that a process mining project eventually results in organizational value (Van Eck et al., 2015).

While existing process mining methodologies generally recognize the importance of evaluating with domain experts (Emamjome et al., 2019), they do not provide specific guidelines as to how such an evaluation should be performed. The main evaluation focus of most existing process mining research is determining the effectiveness of proposed techniques using established metrics such as precision, recall, etc. Projects with this evaluation aim typically use a *quantitative* research approach, which commonly uses methods such as surveys and experiments. These are methods that aim to numerically test theories and models by examining the relationship between variables (Creswell, 2013). While this is often sufficient to evaluate the technique itself, the translation of the findings of a process mining project into actionable insights and recommendations requires an additional evaluation step that involves domain experts (Van Eck et al., 2015). In these evaluations often a *qualitative* research approach is used in which methods such as semi-structured interviews and focus groups are common. This approach is used to explore the 'why' and 'how' behind discovered models or theories (Creswell & Poth, 2016).

Overall, we observe an abundance of projects that involve organizational partners where evaluations take place at the end of a project, but the field varies widely in its approach to the evaluation with these partners. Based on this observation we hypothesize that informal ways of evaluating have been applied over time. With that in mind, we perform a literature study to describe current practices in process mining projects. Additionally, we look into strategies from the qualitative research field, which provide more guidance during qualitative evaluations with domain experts in process mining projects.

This research aims to help move process mining research forward by offering support for process mining experts that seek to perform a qualitative evaluation in their project. More concretely, we propose a list of six validation strategies from the qualitative research field. These strategies should be considered when performing process mining evaluations in which domain experts are involved.

The rest of the chapter is structured as follows. First, we describe the methodology behind the literature study. Then, the results section describes the goals and methods of these case studies. In the proposal section, we describe six existing valida-

tion strategies from the qualitative research domain. Finally, we conclude our work, discuss the limitations, and sketch directions for future research.

7.2 Research method

We performed an in-depth, systematic literature review aimed at reviewing the existing literature on process mining projects in which domain experts are involved. We followed the guidelines presented by (Kitchenham & Charters, 2007). We illustrate the search and selection process in Fig. 7.1. In total, four of the authors actively performed the literature study, we refer to this as the literature team. We explain the process in more detail below.

7.2.1 Extraction and Abstract Screening

Our objective was to include papers that describe the process of performing a process mining project in practice. Therefore, we used the Scopus database to collect a broad sample of papers. Gheasemi & Amyot (2016) showed that Scopus provides the best balance between relevance and quantity when researching process mining papers. The literature team extracted the papers in October 2020 using the keywords "process mining" AND "case stud*". The search resulted in 244 potential candidates being judged on abstract and title. Three members of the literature team were involved with this next phase of the extraction process. To ensure uniformity amongst the team members, the first five papers were individually screened after which the decision to include/exclude was verified by one other team member. Note that conference proceedings are excluded to avoid duplicates as the individual conference papers are already included. The result of this phase was a set of 191 papers that described one or more process mining case studies.

7.2.2 Full Text Coding

Next, each member of the literature team coded five papers in NVivo (Jackson & Bazeley, 2019), a qualitative analysis software by QSR. The code 'expert' was used whenever we came across a piece of text in which a domain expert was mentioned in the context of a process mining project.

In the second part of the coding phase, two members of the literature team, whom we refer to as the core team, performed another round of full-text reading, each on half of the papers mentioning in any way the involvement of domain experts in process mining case studies. We started with five papers before synchronizing and discussing the proposed sub-codes (Goal and Method), after which we read another ten papers before synchronizing a second time. The result of this phase was a set of quotations related to the two topics, collected from a total of 80 papers.

7.2.3 Analysis and Synthesis

In the analysis and synthesis phase, the core team revisited their quotations for the two sub-codes. Each team member created an overview with a reflection on the information found for both the goal and method of the evaluation, for each of the papers.

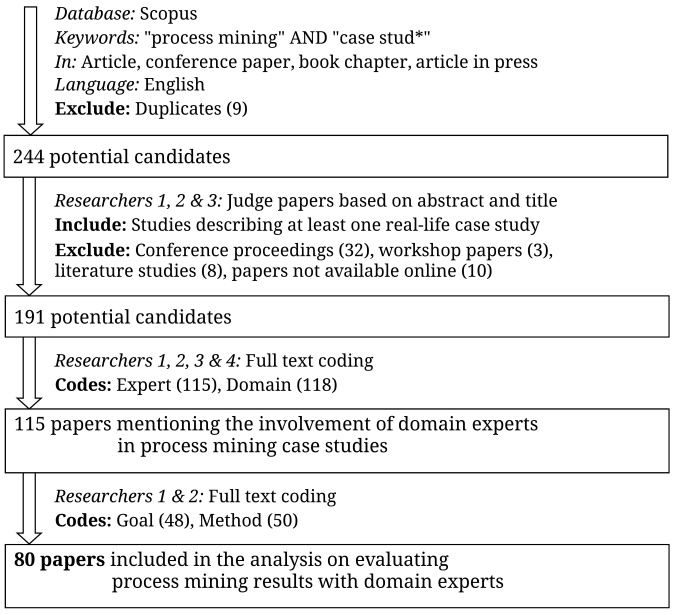


Figure 7.1: Literature study process

After both team members were finished creating the overview and discussing common themes together, the overviews were merged. Finally, the core team performed two rounds of abstraction, the first round analyzed the lower level codes. These were grouped together in higher-level codes that were given by both team members. For example, the description 'conformance checking' and 'identifying non-conforming cases' were combined, this resulted in 14 description codes. Then, the second round of abstraction took a more holistic perspective. In the second round, the core team considered fundamental differences between the higher-level codes. The outcome of these rounds of abstractions is presented in the results. If a paper did not include a quote for either code, the paper was excluded from further analysis (n=35).

7.3 Results

In what follows, we present the results from the coding of the literature review. Recall that we look into process mining case studies in which a domain expert is involved in the evaluation. In these projects, we refer to the authors as *process analysts* to avoid confusion. We coded in two main themes: (1) goals of the evaluation and (2) method of evaluation. Having established this, we can cross-reference the results to see the overlap between the codes. Below, we first go into detail on each of the themes, followed by the cross-reference analysis.

Goal	Sub-goal	Papers
Artifact	Understandability	Abo-Hamad, 2017; Andrews et al., 2020a; Benevento et al., 2019; De Alvarenga et al., 2018; De Weerdt et al., 2013; Deokar & Tao, 2020; Epure et al., 2015; Gorissen, 2017; Jalali, 2014; Johnson et al., 2018; Lehto et al., 2016; Pini et al., 2015; Rojas et al., 2019; Sirgmets et al., 2018; Song et al., 2009
	Usability	Andrews et al., 2020a; Deokar & Tao, 2020; Gorissen, 2017; Pini et al., 2015; Rojas et al., 2019; Sirgmets et al., 2018; Suriadi et al., 2017; Swinnen et al., 2011; Van Langerak et al., 2017
	Quality	Baier et al., 2014; Benevento et al., 2019; Ciccio & Mecella, 2015; De Weerdt et al., 2013; Dees et al., 2017; Deokar & Tao, 2020; Duma & Aringhieri, 2017; Epure et al., 2014; Ferreira et al., 2007; Gorissen, 2017; Knoll et al., 2019; Lee et al., 2016; Ly et al., 2012; Mans et al., 2008; Marazza et al., 2019; Pérez-Castillo et al., 2012; Reijers et al., 2009; Sirgmets et al., 2018; Song et al., 2009; Taylor et al., 2011; Thabet et al., 2018; Van Eck et al., 2016; Van Langerak et al., 2017; Yang et al., 2018
Insights	Findings confirmation	Alvarez et al., 2018; Andrews et al., 2020b; Astromskis et al., 2015; Baier et al., 2014; Cho et al., 2014; Ciccio & Mecella, 2015; Denisov et al., 2018; Elhadjamor & Ghannouchi, 2019; Ferreira et al., 2007; Gorissen, 2017; Gunnarsson et al., 2019; Hessey & Venters, 2016; Hompes et al., 2017; Knoll et al., 2019; Kurniati et al., 2019; Li et al., 2011; Mahendrawathi et al., 2015a; Mahendrawathi et al., 2015b; Mahendrawathi et al., 2017; Mans et al., 2013; Mans et al., 2008; Maris et al., 2016; Mueller-Wickop & Schultz, 2013; Park et al., 2015; Partington et al., 2015; Pérez-Castillo et al., 2011; Prathama et al., 2019; Reijers et al., 2007; Reijers et al., 2009; Rojas et al., 2019; Rovani et al., 2015; Stuit & Wortmann, 2012; Suriadi et al., 2017; Syamsiyah et al., 2017; Van den Ingh et al., 2021; Van Eck et al., 2015; Vogelgesang & Appelrath, 2016; Wang et al., 2014
	Relevance	Gunnarsson et al., 2019; Jalali, 2014; Mahendrawathi et al., 2015b; Pérez-Castillo et al., 2011; Suriadi et al., 2017; Swinnen et al., 2011; Thabet et al., 2018; Van den Ingh et al., 2021
	Generalizability	Hessey & Venters, 2016; Kurniati et al., 2019; Partington et al., 2015; Wang et al., 2014
	Conformance checking	Cho et al., 2014; Erdem & Demirörs, 2017; Mahendrawathi et al., 2017; Roubtsova & Wiersma, 2018

Table 7.1: All citations per goal category

7.3.1 Goal of evaluation

By studying the various process mining case studies in which domain experts are involved, we found that there were substantial differences in the objectives that were set out in the evaluation phase of the individual papers. In Table 7.1 we present the goal and sub-goal categories found in the literature study. In particular, we found the most fundamental difference in the focus of the evaluation. Here, two main focuses can be distinguished: (1) focusing on the evaluation of the artifact itself, or (2) focusing on the evaluation of the insights that can be generated from the artifact.

The first type of evaluation focuses on evaluating the artifact that is created. Slightly less than half of the articles belong to this type of evaluation. We find three measures that highlight which aspects of the artifact are evaluated: (1) understandability (n = 15), (2) usability (n = 9), and (3) quality (n = 25). Usually, a process analyst creates an artifact which is consequently presented to a business domain expert. The *understandability* of the artifact refers to the level of comprehension with which the business expert can examine the artifact. Concrete examples of evaluation criteria for understandability are complexity of the artifact (e.g. (Benevento et al., 2019)) or readability of an artifact (e.g. (Syamsiyah et al., 2017)). The *usability* captures the ease of use of the artifact. In one of the case studies, the process analyst collaborated with domain experts to determine specific KPIs for the process, also in light of conformance checking (Wang et al., 2014). Finally, the *quality* of the artifact describes how well the artifact is constructed. The largest part of the studies focuses on this latter aspect (n = 25). Typically, the quality of an artifact consists of several dimensions: correctness (the artifact is true and correct), completeness (the artifact contains all appropriate elements), conciseness (the artifact represents the same information in a similar fashion repeatedly), and consistency (the artifact does not contain contradictions). Note that quality here only refers to the internal quality of an artifact; the external quality (i.e. clarity) is captured in the usability and understandability features.

The slight majority of the studies focus on the second type of evaluation: the insights that one can generate based on the artifacts that are created during a process mining project; we refer to these evaluations as *insights evaluation*. The largest part of research that focuses on insights evaluation considers the *confirmation* of findings as to the most important aim for the evaluation (n = 41). Typically, insights evaluations focus on the interpretation and explanation of findings represented in the artifacts. For example, some studies compare the expectations of business experts to the insights generated from the artifact (e.g.(Mueller-Wickop & Schultz, 2013)). Another goal is to regard the *relevance* of the findings, where the process analysts try to determine how valuable the produced artifact is to the organization. In one case study, the evaluation focused on four questions, one of which concerned determining the relevance of their proposed framework for designing data visualization for process mining diagrams (Sirgmets et al., 2018). Some studies also consider the *generalizability* of the produced artifact. To exemplify, in one case study a multi-perspective approach was tested in one context, where the evaluation focused on gaining insights into the applicability of the approach in other contexts (Kurniati et al., 2019). Finally, some studies focus on *conformance checking* based on the produced artifact. Here, the

goal is to perform a "reality check". These studies compare the artifact that is created based on real data (event logs) to the existing protocols within an organization as indicated by the domain experts (e.g. (Cho et al., 2014)).

7.3.2 Method of evaluation

We now turn our attention to the *method* that is used in the evaluation of process mining case studies in which domain experts are involved. In Table 7.2, we present the coding hierarchy for the method of the evaluation. We can distinguish two main types of methods: qualitative and quantitative. Within each type of method, we identify a number of data collection techniques (e.g. survey). On an aggregate level, we observe a strong tendency to perform the evaluation using a qualitative method (n = 64).

What stands out within the qualitative methods is that we find a clear pattern as to the lack of structure in the evaluation phase of a process mining project. The vast majority of the studies perform an *undefined discussion* to evaluate their results (n = 39). These undefined discussions are usually described using the following terminology: “The results were presented to a business expert.” or “We discussed the results of the project with a domain expert.”. In these studies, no guidelines, rules, or protocols are described that support the evaluation phase of the research.

Other data collection techniques are also employed in the qualitative method group, such as *interviews* (n = 19), *focus groups* (n = 2), and *workshops* (n = 4). These studies all describe a setting in which the domain expert and process analysts interactively discuss the study findings following some sort of protocol. Often, a rigorous approach and reflection on the chosen method and technique lack in these works as well.

By contrast, we observe that in the quantitative method group, studies follow a more rigorous approach. For example, a strand of studies collects data through the *annotation* of process mining results (n = 12). In one case study, the authors ask experts to manually annotate the complexity of so-called attack models for an intrusion detection system (De Alvarenga et al., 2018). Furthermore, some studies use *surveys* to evaluate their findings (n = 6). This is generally done when the evaluation criteria are determined *a priori*. For example, conjoint analysis is used to determine the weights of various performance dimensions (Van den Ingh et al., 2021). Finally, we see a handful of studies setting up an experiment (n = 4) to study their results. For example, in one research the authors simulate a hospital setting to validate their algorithm (Cho et al., 2014).

7.3.3 Cross-reference analysis

As presented above, process mining project evaluations can differ fundamentally in two respects: (1) their goal, and (2) their method. It is important to note that the divisions presented above are not mutually exclusive: a study can pursue both types of goals or employ multiple types of methods. The goal of the present study is not to provide a normative perspective on when to use which method or define what goal. Rather, we describe the current practices in process mining projects. There are two things that stand out when we do this: (1) a *qualitative approach* is often taken to

Method type	Methods	Papers
Quantitative	Survey	Andrews et al., 2020a; Cho et al., 2014; Epure et al., 2014; Gorissen, 2017; Suriadi et al., 2017; Van den Ingh et al., 2021
	Manual annotation	Baier & Mendling, 2013; Benevento et al., 2019; Calvanese et al., 2017; De Alvarenga et al., 2018; Deokar & Tao, 2020; Dixit et al., 2018; Lee et al., 2016; Pérez-Castillo et al., 2012; Pérez-Castillo et al., 2011; Thabet et al., 2018; Van Eck et al., 2016; Yang et al., 2018
	Experiment	Deokar & Tao, 2020; Epure et al., 2014; Pérez-Castillo et al., 2012; Van den Ingh et al., 2021
Qualitative	Focus group	Gunnarsson et al., 2019; Johnson et al., 2018
	Undefined discussion	Andrews et al., 2020b; Astromskis et al., 2015; Benevento et al., 2019; Cho et al., 2014; Cho et al., 2019; Ciccio & Mecella, 2015; De Weerdt et al., 2013; Denisov et al., 2018; Elhadjamor & Ghannouchi, 2019; Engel et al., 2012; Erdem & Demirörs, 2017; Hompes et al., 2017; Lehto et al., 2016; Li et al., 2011; Mans et al., 2008; Marazza et al., 2019; Maris et al., 2016; Martin et al., 2019; Montali et al., 2014; Mueller-Wickop & Schultz, 2013; Park et al., 2015; Partington et al., 2015; Pérez-Castillo et al., 2012; Pérez-Castillo et al., 2019; Pini et al., 2015; Prathama et al., 2019; Reijers et al., 2007; Rojas et al., 2019; Rovani et al., 2015; Samalikova et al., 2014; Song et al., 2009; Stuit & Wortmann, 2012; Swinnen et al., 2011; Syamsiyah et al., 2017; Taylor et al., 2011; Van den Ingh et al., 2021; Van Eck et al., 2015; Van Langerak et al., 2017; Vogelgesang & Appelrath, 2016
	Interviews	Baier & Mendling, 2013; Benevento et al., 2019; Calvanese et al., 2017; De Alvarenga et al., 2018; Deokar & Tao, 2020; Dixit et al., 2018; Epure et al., 2014; Erdem & Demirörs, 2017; Gorissen, 2017; Lee et al., 2016; Ly et al., 2012; Mahendrawathi et al., 2017; Pérez-Castillo et al., 2012; Pérez-Castillo et al., 2011; Sirgmets et al., 2018; Thabet et al., 2018; Van Eck et al., 2015; Van Eck et al., 2016; Wang et al., 2014
	Workshop	Epure et al., 2015; Kurniati et al., 2019; Roubtsova & Wiersma, 2018; Sirgmets et al., 2018

Table 7.2: All citations per method category

	Qualitative	**Quantitative**	**Qualitative & Quantitative**	**Total**
Artifact	7	3	5	**15**
Insights	28	1	2	**31**
Insights & artifact	7	2	2	**11**
Total	**42**	**6**	**9**	**57**

Table 7.3: Cross-reference results when combining the highest level codes in method (columns) and goal (row) of evaluation

evaluate process mining findings, and (2) the majority of the projects aim to generate insights based on their artifacts, i.e. *insights evaluation*.

To gain a better understanding of the relation between method and goal in the evaluation of a process mining project, we cross-referenced the higher-level codes. In Table 7.3 the results are visualized for all 57 studies for which a goal and a method were defined. What we can see from the table is that most studies use a qualitative approach (n = 42), especially when the goal is to gain insights (n = 28). Another apparent trend shows that there are quite some studies that use a mixed-method approach (n = 9) or have mixed goals (n = 11).

What we can infer from this is that a qualitative approach with the goal of insights evaluation is most frequently used. Recall that the most frequently used method in the qualitative approach is an undefined discussion. In addition, note that almost a third (n = 23) of the original studies (n = 80) are not included in the cross-reference analysis (n = 57) as they do not define both a goal and a method for their evaluation. Thus, a large portion of studies are lacking a structured method to perform this insights evaluation. Most existing case studies do not follow specific guidelines to properly evaluate their findings in a qualitative manner. Thus, we observe that a systematic approach to these types of evaluations is missing. We observe that individual studies sometimes use good practices. We believe this requires a look into existing best practices from the qualitative research literature.

7.4 Proposal

From the literature study, we can conclude that a qualitative approach is often taken, but a systematic approach for determining the accuracy and meaning of findings is mostly lacking. In the wider scientific literature, there are many ways in which normative support is offered to researchers. Such ways of support can be divided into different layers, which Saunders et al. (2009) illustrate by means of a so-called 'research onion'. When designing a study, researchers peel off the individual layers one by one, going from broad research philosophies all the way down to specific data collection and analysis techniques. A process mining study can be designed in a similar way. Researchers may start by deciding on a broad research paradigm, such as design science (Hevner et al., 2004), before deciding on particular research methods, such as a case study. Within such a case study, they may choose specific data collection techniques such as interviews or focus groups. Up to this point, the literature on eval-

uation with domain experts in process mining is quite explicit. However, the *inner layers* are not as clear. There is little discussion on how data is best collected and analyzed or how accuracy is ensured. To propose a way to fill this gap, we look into literature from qualitative research. Based on literature from the qualitative research field, we propose six validation strategies that should be considered in qualitative process mining evaluations. Below, we will go into detail for each strategy and reflect on the extent to which these strategies can be observed in current process mining practices.

7.4.1 Validation Strategies

Strategy	**Effect**	**Practical guideline**
Engage with the field of research	An open and honest evaluation	Carefully select domain experts; include data quality issues in presentation of the results
Triangulation	Completeness and consistency of the results	Use multiple quantitative (e.g. simulations) and qualitative evaluation methods (e.g. interviews)
Peer review or external audit	Credibility of the analysis and interpretation	Plan peer reviews to reflect on research design, approach, and results on a regular basis and keep notes of these meetings
Refine work hypothesis	Transparency and soundness of the results	Keep detailed notes on hypotheses, how they are tested, and the final results and use these to guide the evaluation with domain experts
Clarify and normalize bias	Transparency and reliability of the results	Discuss different types of biases in the evaluation or limitation section
Perform member checking	Credibility of the results	Ask interviewees to check the correctness and authenticity of a summarized report of the interview results and interpretation

Table 7.4: All validation strategies, effects, and practical guidelines

In their seminal book on qualitative research methods, the authors discuss validation strategies (Creswell & Poth, 2016), see Table 7.4. They propose a number of strategies to perform in qualitative studies: (1) engage with the field of research, (2) triangulation, (3) peer review or external audit, (4) refine work hypothesis, (5) clarify bias, (6) perform member checking. These strategies are all potentially relevant for process mining projects that qualitatively evaluate their findings with domain experts. In their work, the authors recommend researchers always follow at least two strategies when engaging in qualitative research (Creswell & Poth, 2016). Below, we first elaborate on each strategy, explaining how they aim to improve qualitative research in general. Second, we provide guidelines to show how the strategy can be applied in the context of an evaluation with domain experts in a process mining project.

Engagement and understanding of the field

This strategy refers to the relation a researcher builds with the study participants, the understanding the researcher builds of the (organisational) culture, and the ability of the researcher to spot how misinformation might influence the study (Creswell & Poth, 2016; Lincoln et al., 1985). In a process mining project, the study participants are the domain experts involved in the study. Building a trust relationship with the experts ensures that there is an honest and open evaluation at the end of the project. Honesty and openness contribute largely to the value generation for both the research and the organization. Next to that, the researcher gains an understanding of the culture of the organization to interpret the data correctly. Finally, it is of vital importance that a researcher gets a feeling for where misinformation, such as bad data quality, can stem from to account for this during the interpretation and generation of the findings. This prevents the researcher from drawing wrongful conclusions.

A good practice is presented in the process mining literature by Alvarez et al. (2018). The authors start the qualitative evaluation by presenting the final findings, assumptions made, and the interpretation of the findings. In light of the strategy of engagement and understanding of the field, this can be expanded in two ways. First, to *carefully select* the domain experts that are involved in the qualitative evaluation. The domain experts must possess the required knowledge of the project and hold a central position in the organization to collaboratively interpret the results. Second, to *Standardize and discuss the presentation*. Presentations might often be held in process mining projects, but are rarely discussed explicitly in the research article. As a result, the content of such presentations can vary widely. Therefore, we propose to explicitly mention if a presentation is given to domain experts in the evaluation and discuss the content of the presentation. We propose to include at least three points in the presentation as proposed by Alvarez et al. (2018): (1) final findings, (2) assumptions made, and (3) interpretation of findings. In addition, we advocate including a fourth part: discussing data quality issues that arose and were tackled during the project. Making this an explicit part of the presentation allows for the domain experts attending the qualitative evaluation to check if all potential data quality issues are addressed, this is in line with the identification of misinformation as discussed previously.

Triangulation

This strategy refers to the use of multiple data sources to study the research problem (Creswell & Poth, 2016; Patton, 1999). Ideally, a *mixed-methods* approach using qualitative and quantitative techniques is taken to increase the validity of the findings. Another possibility is to use multiple data sources within one type of method. In our literature study sample, no example could be found that applied the triangulation strategy to one evaluation goal. In one research, the authors apply a mixed-methods approach, but they do so by applying one method for each goal (Epure et al., 2014). Their quantitative evaluation describes how a confusion matrix, as proposed by Provost & Kohavi (1998), is used to measure the performance of the proposed process mining algorithm. The qualitative evaluation aims to validate the quality of the recommendations that the algorithm produces through structured interviews. In order to apply the triangulation strategy in full, the interviews would need to be complemented with another data source to evaluate the quality of recommendations.

Peer review or external audit

This strategy refers to reviewing the research process with a reviewer or auditor (Creswell & Poth, 2016; Miles & Huberman, 1994). The difference between the peer review and the external audit is the connection to the research. An external auditor cannot have any connection to the research, whereas a peer reviewer can have some connection to the research. This strategy was proposed by Lincoln et al. (1985) who describe the reviewer as a 'devil's advocate' that checks with the researcher (in process mining, process analyst) how the research is performed. The authors stress the importance of doing these peer reviews on a regular basis during the research and keeping notes of each meeting. In process mining, this would be a good strategy to: (1) critically reflect on the research design, (2) ensure the data is handled in a compliant manner, and (3) offer opportunities to discuss interpretations of results.

In practice, peer-reviewing or external auditing would not be included in an academic article. As such, it is hard to determine the extent to which this strategy is already applied in the process mining community. We propose that the content produced through this strategy can serve as complementary material that should be provided upon request. The external audit works similarly to a peer review of an academic article. However, this audit only focuses on the result generation and interpretation. The external auditor is given access to the data and notes, including the data collected from domain experts. The audit focuses on answering the question: are the findings, conclusions, and interpretations, supported by the data? This is particularly useful when domain experts are involved and data generated by them is interpreted by the process analyst. Performing such an audit or review increases the validity of the research as it gives an outsider a chance to check the interpretation of the process analyst.

Refine work hypothesis

This strategy describes how a researcher can use negative case analysis (Creswell & Poth, 2016; Lincoln et al., 1985; Patton, 1999). This is an analysis in which a researcher formulates a hypothesis, and changes it every time a case is encountered that the hypothesis cannot explain. Keeping detailed notes on each hypothesis, the way it is tested, and the final result increases the transparency and soundness of the final results, and insights generated. These notes can help guide the qualitative evaluation in the sense that they can help scope and structure the process of translating the findings into insights with domain experts.

The cyclical nature of hypothesis refinement is a well-known and established approach in process mining analysis (see for example (Van Eck et al., 2015)) and in certain paradigms (see for example (Hevner et al., 2004)). The concept is usually applied throughout the project. However, to the best of our knowledge, no research has described how to use it during the qualitative evaluation. This strategy can be applied to the work of Alvarez et al. (2018). In their work, they define two hypotheses for the analysis. An additional hypothesis for the qualitative evaluation would be a good first step to help guide this phase of the project. In addition, transparency into the evolution of the hypotheses provides valuable insights into the research. It makes the considerations of the process analyst explicit. This allows domain experts to validate the process of insight generation that the process analyst has conducted.

Clarify biases

Furthermore, another strategy describes the standard practice of reflecting on the possible *biases* from a qualitative perspective in process mining research. Many biases can be present in a study, especially when domain experts are involved. One illustrative example is the familiarity of domain experts with process mining (Pérez-Castillo et al., 2012; Sirgmets et al., 2018; Van Eck et al., 2015). Depending on the goal of the process mining project, this might influence the qualitative evaluation in different ways. For example, in recommendation systems, the experience of the domain experts influences the quality of the output of an algorithm, a domain expert more experienced with process mining can better understand recommendations displayed in a process model, whereas a less experienced domain expert might benefit from recommendations in text form. We know that the more experienced domain experts are with process mining, the better they can interpret the models and derive insights. It should be standard practice to discuss these types of biases in the evaluation or limitations of a process mining project. To exemplify, consider one research that already does so: "the assessment of this study would probably be carried out through surveys involving people that possibly do not have the same expertise level about business process management, which may imply biased results." (Pérez-Castillo et al., 2012).

Member checking

This strategy originated in studies that use interviews as a data collection technique (Creswell & Poth, 2016; Lincoln et al., 1985; Stake, 1995). It describes how researchers, after they have interviewed the participants and analyzed the data, return to the interviewees to confirm that their findings and the interpretation thereof are credible. Looking at the literature study, we can see that a comparable approach is partially deployed by some researchers who aim to confirm their findings through interviews. The advantage of introducing member checking as an approach in process mining is that it introduces rigor by providing a set of guidelines and procedures. Consider for example a specific member checking technique proposed by Birt et al. (2016). The authors describe a five-step plan to increase the validity of studies that use interviews: (1) prepare a synthesized summary of raw data and the interpretation of the data, (2) formulate criteria to select participants for eligibility for a member check, (3) send member check data and feedback form, (4) gather response data, and (5) integrate response data with raw and interpreted data.

We can apply the member checking strategy to the work of Epure et al. (2015). After the interviews have been conducted and analyzed the process analysts would return to (a part of) the group of interviewees (i.e. members) and provide them with a report on their findings of the interview. The members would then be asked to check the authenticity and their comments can be used as input to check for the validity of the interpretation of the interview data. The strategy can also be used when focus groups are used as a data collection technique.

7.5 Conclusion

The involvement of domain experts in process mining projects is essential in translating results into actionable insights. We performed a systematic literature study of

recent process mining case studies where domain experts were involved in the evaluation. We found that such evaluations are performed widely, but that there is a lack of structure in *how* they are performed. With that in mind, we present six strategies from the qualitative research field. We show good examples of existing process mining projects where aspects of these strategies are already applied. These six strategies contribute to the process mining community by offering a set of guidelines to perform a more rigorous qualitative evaluation of process mining results with domain experts.

The literature study has been performed using an established systematic review checklist and four researchers were involved in the process. However, the setup may pose some *limitations*. First, our sample of papers purely consists of case studies in which the involvement of a domain expert is mentioned. Although different terms were used, it excludes studies in which an expert was involved but not mentioned explicitly. Second, the sample of papers was exclusively drawn from the Scopus database. Although Scopus has been shown to contain the most relevant process mining papers, a small number of additional papers may be found in other databases. Third, the use of our search terms might exclude studies that use case studies but do not refer to them as such. Last, a large number of studies do not explicitly describe how they evaluate their findings. As such, we cannot infer anything from their practices. Explicit reporting on the qualitative evaluation aspects is highly recommended to increase the transparency, replicability, and validity of a study.

In future work, we want to focus on filling in each layer of research (i.e. from philosophies to data collection and analysis) as defined by (Saunders et al., 2009) for qualitative evaluations in process mining projects. The insights on each of these layers contribute to a more rigorous, and ultimately better, qualitative evaluation of process mining projects.

Concluding the Research

CHAPTER 8

Conclusion

In this chapter, we reflect upon and conclude the work of this dissertation. First, we *reflect* on the contributions that were outlined at the start of this dissertation. Then, we look beyond these contributions to discuss the *practical implications* of this dissertation. After this, our attention turns to a more elaborate visitation of the *Challenges and Open Issues* that are posed to the works in this dissertation. Finally, we look to the future and discuss directions for *future development* of the work in this dissertation.

8.1 Reflection

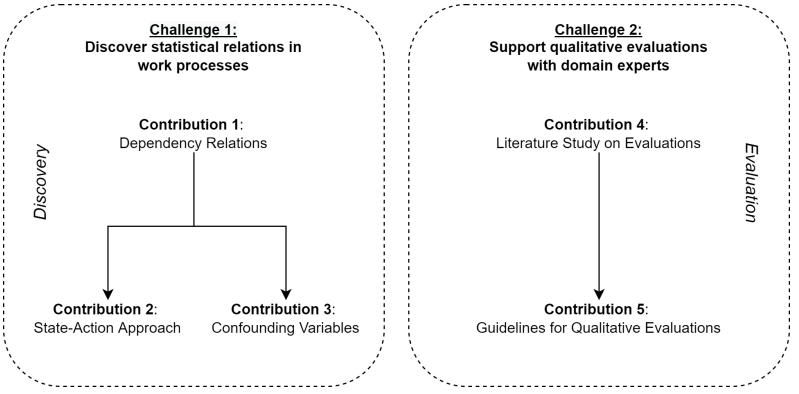


Figure 8.1: Overview of the main challenges and contributions

At the start of the dissertation, we defined two main challenges and five main contributions. In Figure 8.1 we visualize the key challenges and their related contributions. In this section, we will revisit each contribution and summarize how they are addressed. Then, we reflect on the contributions by considering their impact on the existing body of knowledge.

Contribution 1: A process mining technique that uses well-established statistical mechanisms to detect, analyze, and visualize statistical relations within work processes.

One of the process mining techniques proposed in this dissertation is the ARE miner, which is developed in Part 1 (Chapter 2 - 4). The ARE miner contributes to addressing the challenge of identifying, analyzing, and visualizing statistical relations in work processes. In Chapter 3 & 4, we propose an algorithm that combines a number of statistical tests, and a visualization in graphical form of the statistical relations. The ARE miner is a technique that helps organizations to understand the impact of their activities on their future activities. The potential dependency relations analyzed and visualized can inform an organization on how they could optimize a process to respond in the best way possible.

Contribution 2: A novel technique to detect, analyze, and visualize statistical relations between actions and states in a work process.

In Chapter 5 we propose a technique that moves beyond the notion of activities to states and actions. In the technique, we define a state-action event log, which allows us to capture information beyond that of an activity. For example, we can study the state of a patient (e.g. *healthy*) and how actions influence this state. The technique is built around a similar set of statistics as the ARE miner, like the Chi-square, and produces enhanced process models. This extension incorporates the first challenge of using statistical mechanisms to discover dependency relations in work processes. The technique proposed in Chapter 5 allows for a broader perspective of what elements a relation can consist of. While the ARE miner is restricted to analyzing activities; this extension allows for states and actions, which can both be loosely defined.

Contribution 3: An extension to existing process mining discovery techniques to help identify confounding variables in processes.

In Chapter 6 we propose an extension that can detect and deal with confounding variables. Confounding variables are alternative explanations for a potential dependency relation. The ARE miner produces such potential dependency relations. We take a causal perspective to address the main challenge of discovering statistical relations in work processes. One of the main concerns, when the term causal relations is brought up, is: Is there a check for alternative explanations? This extension is important as it helps to check for alternative explanations. Specifically, it helps to check if the potential dependency relation discovered by the ARE miner can be *better explained by alternative variables*. Once this is checked, the relation is more likely to be one of a causal nature.

Contribution 4: An overview of current practices in process mining case study evaluations in which domain experts are involved.

In Chapter 7 we move past the discovery phase of process mining into the evaluation phase. Here, we perform a literature study to understand how external experts from organizations (*domain experts*) are involved in this step of a process mining project evaluations. Specifically, we are interested in how these evaluations are performed (method) and what the goal of these evaluations are. To understand this we studied process mining case studies of the past years. We discover that most of these studies focus on generating insights, but lack a defined method as to how they perform the evaluation. These insights help us address the second main challenge of this dissertation. The second challenge describes how there is a need for more *actionable support for evaluations with domain experts*. This chapter helps us create an understanding of exactly what kind of methodological support is required in process mining projects.

Contribution 5: A set of guidelines to support qualitative evaluations in process mining projects in which domain experts are involved.

Finally, in Chapter 7 we build on the insights generated in contribution 4 to propose a set of six guidelines to support process mining evaluations. Here, we specifically focus on the qualitative aspect of the evaluations. The guidelines help address the second main challenge of this dissertation. This challenge describes the need for actionable support in process mining evaluations in which domain experts are involved. Specifically, the guidelines provide *actionable support* to process analysts as to how to increase the validity of their evaluation if they include domain experts. In other words, the guidelines are a first step to help process analysts accurately measure what they intend to measure in an evaluation with domain experts.

8.2 Practical Implications

This dissertation, in practical terms, can be approached from two perspectives: *technique applicability* and *aggressive behavior*. First, there are implications concerning the use of the presented techniques when we implement them. Second, in the presented work the focus of the studies is largely on presenting work where we study aggressive behavior of clients that are physically and/or mentally challenged.

8.2.1 Insights into the Applicability of the Techniques

In this section, we will reflect on the lessons learned by applying the techniques presented in this dissertation in real-life scenarios. First, we elaborate on the practical implications of the notion of *statistical relations*. Then, we look into the *use cases* of the proposed techniques. Finally, we discuss the use of *domain knowledge* in the techniques.

Statistical Relations

In this dissertation, we propose new techniques to discover a different type of relation in work processes. In addition to sequential order relations, we propose to look into *statistical relations*, which we refer to as potential dependency relations. The limitation we address is that just looking at frequencies and order in a process does not always give the insights required by an organization. Especially when there are moments in the process where an actor or organization can make a choice about which activity to perform. In such a context it is interesting to gain an understanding of the impact such a choice has on the outcome of the process. In this section, for readability purposes, we only refer to activities and outcomes. These concepts can be replaced with actions and states as well.

What makes the notion of a statistical relation interesting is that both its discovery and its absence have interesting implications. In the case that a potential dependency pattern is *discovered*, an organization gets a first understanding of which of its activities have an impact on the outcome. We refer to the relation as potential, as we cannot be certain that there is causality. This remains an open issue which we elaborate on further in Chapter 8.3. What we learned from the application of this technique in our case studies is that it can be equally interesting the *absence* of a de-

pendency relation where one would be expected. For example, you find there is no dependency relation between a certain activity (e.g. administering medicine) and an outcome (e.g. patient cured). Finding no relation indicates that this activity has little to no effect on the outcome. These insights can help organizations think about where to invest their resources.

Let us return to the scenario where we do discover a potential dependency relation. Here, the techniques proposed in this dissertation also provide three extra features that help interpret the relation: direction, strength, and frequency. One of them is the *direction* of the discovered potential dependency relation. The direction of the relation tells us if an activity is more or less likely to lead to a certain outcome. To exemplify, dependency tells us if applying a treatment (e.g. talking to a client) is related to future verbal aggression. Direction then tells us how they are related, i.e. talking to the client leads to more future verbal aggression. This helps organizations understand how the activity and outcome are related.

Another interesting feature is the *strength* of the discovered relation. The strength of a relation describes how large the impact of an activity is on the outcome. The techniques use a three bracket system: small, medium, and large to indicate the strength. Let us return to the example where the activity is talking to a client and the effect is future verbal aggression. We learned that talking to a client leads to more future verbal aggression. Strength tells us this is a strong relation, i.e. it is much more likely that after talking to a client the client becomes verbally aggressive in the future. This helps organizations get a feeling for how powerful the discovered potential dependency relation and its direction are. Important to note is that all techniques only return potential dependency relations that are statistically significant. Finally, the techniques provide a feature similar to that of a traditional discovery technique: *frequency*. This helps understand how frequently the activities, actions, or states are performed or occur. In general, it is interesting because it helps organizations distinguish between common and uncommon process variants.

Use Cases

On a domain level, we have focused on the healthcare domain in this dissertation. We have applied it in two specific contexts, one of aggressive behavior for which we collected new data through a collaboration with a healthcare organization in the Netherlands. The other context is one from a public healthcare data-set from the Technical University of Eindhoven on sepsis (see (Mannhardt & Blinde, 2017) for more details). Healthcare is an interesting use case for this technique as there is a large interest in understanding the effect of treatments. In our respective examples, the effect of the treatment on how to respond to an aggressive incident. In the case of sepsis, the effect of different decisions in a sepsis process, such as to which ward a patient is admitted to or which blood tests to perform. Looking beyond the value the techniques bring to the domain of healthcare, we see a lot of potential to apply this technique to other domains as well. Examples of other domains include, but are not limited to: finance, logistic, customer behavior, and marketing. It is important to note that the applicability of the techniques is not bound to a certain domain. The applicability of the techniques is bound to a certain type of process. In the rest of this section, we illustrate how the type of process differs per technique.

In the ARE miner, we propose a technique that can discover statistical relations in work processes. The ARE miner takes a specific event log, the ARE log, in which we define three types of activities: an action, a response, and an effect. For applicability purposes, it is important to understand that we define the effect to be the next action for a case. In the context of aggressive behavior, the action is an aggressive incident, where we specifically look at the type of aggressive behavior. So the effect is the next aggressive incident of that same client, for which we record the type of aggressive behavior again. This is interesting as it helps us understand how the response has an impact on the behavior of the client. However, the ARE miner approach has limits to what kind of processes we can study. The ARE miner focuses on processes for which the end and start are similar (i.e. the action (start) and effect (end) are the same activity). Therefore, the ARE miner cannot look at other process outcomes. Following up on the aggressive behavior example, the ARE miner does not allow us to study the well-being of a patient after an aggressive incident.

Therefore, we introduce an extension in Chapter 5 that allows for a more generic approach to studying potential dependency patterns. By introducing the concepts of states and actions this approach is more flexible and allows for more types of processes and their related event logs to be studied. States and actions are fairly freely defined and can consist of one or many activities or even attributes that can be used to define them. In Chapter 5 we showcase this, for example, we use the location of the patient (the ward to which the patient is admitted) as a status of the patient. As such, we can now study a large variety of process outcomes and the impact of more process-related elements, rather than just activities, on that outcome. An important note that we have not addressed is how these actions, responses, effects, actions, and states are defined. In the next section, we elaborate on this when we discuss the domain knowledge required by the techniques proposed in this dissertation.

Domain Knowledge

All techniques require some form of *domain knowledge*. Two ways in which domain experts help in these techniques stand out: defining input data and evaluating results. First, this knowledge is required to *define* the variables, activities, states, and actions. In this dissertation, we do not provide a fully automated technique in the sense that it can detect, from raw data, potential variables that can serve as input for the techniques. A domain expert is always involved to help identify potential variables and, if necessary, help identify categories within the variables. Second, at the end of the project, a domain expert is often involved to help *evaluate* and interpret the results. This is common practice, but it is easy to underestimate the time that is required for the domain expert to understand the outcomes of these techniques. Not only does the domain expert need to understand process mining in general, but also the more technical implications of the statistical results need to be conveyed and discussed. This last step is vital in helping organizations understand the value of the output of the techniques. It is also a good moment to reflect and brainstorm on potential ways to improve a work process that is analyzed.

8.2.2 Insights into Aggressive Behavior

This dissertation, in practical terms, largely focuses on presenting work where we study the aggressive behavior of clients that are physically and/or mentally challenged. Much of this work has been published in more technical outlets, such as Business Process Management, Process Mining, and Information Systems outlets. As such, in the publications, we focus on assessing and explaining the performance of the techniques proposed. As a result, the practical insights that are generated during these exercises are disseminated throughout the chapters. The goal of this section of the dissertation is to provide a central point where we revisit, summarize, and reflect on the most prominent practical findings and insights. In this respect, we put a focus on the insights related to aggressive behavior. Here, we first discuss insights related to the client. Then we look at how caretakers can respond to these aggressive actions of clients. Regarding the responses of caretakers, a detailed overview can be found in Appendix A. Finally, we reflect on these findings and consider how these insights can be translated to be used in practice.

Client Action

To understand aggressive behavior, it is important to know that the different types of outwardly directed aggressive behavior can be ranked according to their general perceived severity. In that way, from least to most severe, the order is: verbal, physical aggression towards objects, and physical aggression towards people. In addition, self-injurious behavior is studied, but that is inwardly directed aggression and thus it cannot be ranked with the other types of aggression in this way. Finally, in most techniques, we introduce the aggressive behavior type tau (τ) which indicates that no new aggressive incident has been observed for the same client within a certain period (e.g. 14 days). For the healthcare organization, it is interesting to understand how they can de-escalate the violence of their clients. This means moving towards tau as much as possible or at least minimizing the number of incidents where there is physical aggression towards people. It is with this mindset that we study the output of our techniques and summarize the practical implications below.

In Chapter 2, we take a more client-oriented perspective and we make two important observations: (1) we find that clients tend to repeat the same types of behavior, and (2) we find that aggression towards people is (by far) the most frequently displayed type of aggressive behavior by clients. It is interesting to see that clients tend to repeat the same types of behavior. This poses the challenge of understanding how we can move clients to deviate from their 'normal' behavior. Especially if we take into consideration the fact that clients mostly tend to display the most severe type of aggression, the urge to understand and change how they behave becomes larger.

Caretaker Response

Let us adopt the perspective of the caretaker to try to understand how we can approach the need for a change of aggressive behavior in clients. After all, the goal of most techniques is to help understand the impact of activities on future activities and the process as a whole. In the case of aggressive behavior, we are mostly interested in how the response of a caretaker impacts the behavior of a client. This is also the activity on which we have the most influence from a practice point of view, we cannot

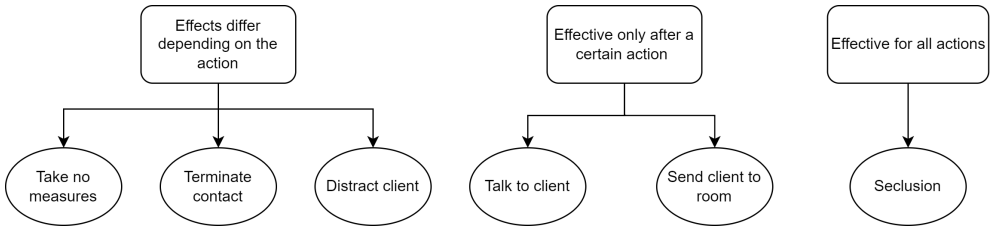


Figure 8.2: Overview of caretaker response observations

change the initial action of the client in this model. Important to understand when we talk about the response of caretakers is that we identified eight types of responses which can also be ranked from less to more severe: no measure, talk to client, distract client, terminate contact with client, send client to another room, preventive measures, held with force, and seclusion (isolation with a locked door). We have not found insights for all responses, but below we go into detail about a number of them. A high-level overview can be found in Figure 8.2, and a more detailed overview can be found in Appendix A.

If we adopt the caretaker response view, we can work from three grand observations: (1) depending on the action of the client, the same response of the caretaker may lead to different outcomes, (2) some responses of the caretaker may only be effective for a certain action of the client, and (3) some responses of the caretaker have a similar effect regardless of the action of the client.

The *first observation* is that a response may effective in different ways depending on the action that preceded it. Most of the practical insights we obtained can be categorized under this observation. To exemplify, we identified that *not taking a measure* as a response has a very different effect after verbal compared to self-injurious behavior incidents. In the case of verbal aggression, not taking a measure is related to a decreased likelihood that the client will not become aggressive in the future. In other words, not responding to a verbally aggressive incident escalates future violence. In contrast, not responding to the incident after a client has shown self-injurious behavior leads to an increased likelihood that the client will not become aggressive in the future. In addition, in the context of self-injury, no response also decreases the chances of the client becoming aggressive towards another person. Both effects show that not responding to self-injurious behavior de-escalates future violence. More insights on other responses can be found in Appendix A.

The *second observation* is that a response may only be effective after a particularly aggressive action, whereas it is ineffective after other aggressive actions. Consider the response of a caretaker to *talk to a client*. In the studies, we find that this response has an effect when a client has shown self-injurious behavior. In that case, it escalates violence as it is related to an increased likelihood of future aggression in terms of physical harm towards another person. However, when a caretaker responded by talking to a client after an incident of the other types of aggressive behavior, no effect was found on future aggression.

The *third observation* is that some responses will have the same effect, regardless of the preceding aggressive action of the client. We observe this specifically for the

most severe type of response: *seclusion*. In short, we find that seclusion always leads to an escalation of future violence. It is more likely that clients become physically aggressive towards people after a caretaker has secluded them. In addition, it becomes less likely that the client will not display any type of future aggression.

Translating Findings

A question that might arise after reading these insights might be: what can we do with them? One promising direction for translating our research findings into practical relevance is to use the insights as a basis for evidence-based training programs. These techniques can be used to test and verify potential strategies for dealing with aggressive behavior. They can show the impact of activities on behavior and help (new) caretakers understand in which circumstances which responses might be more or less desirable.

Another way of looking at the translation of findings of this work is to see how the developed techniques might help discover new insights in the future. To that end, the publication of Chapter 4 was used as a blueprint to successfully apply to a tech startup program. This program is run by UtrechtInc, a startup incubator for the Utrecht region. Within the program, the ARE miner is used to explore ways to further develop the techniques and relate them to market needs. The program is a 6-month program and at the time of writing, the program is still running.

8.3 Challenges and Open Issues

Throughout the dissertation, several challenges and open issues are addressed. In this section, we focus on one main challenge and one important open issue that deserves a more elaborate background than what is provided in the chapters. The main challenge regards the *data requirements* for the various techniques. In addition, we address an important open issue regarding the distinction between potential dependency patterns and *causality*.

8.3.1 Data Requirements

The main challenge originates from the common use of the same statistical test (Chi-square) in all techniques. The chi-square tests impose several *data requirements* to the data that is used as input. Most notably, the test requires the data to be of the categorical form. This means that each numerical variable must be adjusted accordingly. Practically, this means that a variable such as age (let us assume a whole number between 0 and 120) cannot be used without adjustments. To modify age to fit the Chi-square test one could, for example, define three categories (e.g. minors (0-18), adults (19-65), and seniors (66-120)). The categories must be mutually exclusive in the sense that one data point cannot be in multiple categories.

In addition, the data cannot be repeated measures. Repeated measures are data for which we register data for the same category at two different periods in time, usually, we call them a pre-test and post-test. For example, data on a pain score that is registered before and after treatment cannot be analyzed. The category is the same (i.e. pain score), only the time point of measurement is different (i.e. before and after treatment). The chi-square test in this case cannot be used to study the variable pain

score. Most of these data requirements can be checked by either the person using the technique or a domain expert. Some other data requirements, such as the size of the data-set, are automatically checked by the techniques proposed in this chapter, see Chapter 4.

A final note regarding the data concerns the data distribution. Although the Chi-square test is a non-parametric test, which means it is quite robust to distributions in data. Nonetheless, the distribution of data does have an impact on its performance as we show in the experiments in for example Chapter 4. What is important to note is that when it comes to the check for confounding variables, the performance of the technique improves as the data is distributed more equally (see Chapter 6).

This section shows the challenges of the techniques proposed in this dissertation in terms of data requirements. They stem from the specific requirements related to the use of the Chi-square test. These requirements need to be kept in mind when the techniques are used. Keeping the requirements in mind helps to ensure that the results produced by the techniques are valid.

8.3.2 Causality

Causality is an important topic in terms of the open issues related to the techniques proposed in this dissertation. There are three important aspects of causality: (1) association, (2) temporal order, and (3) alternative explanations (Allen, 2017). The techniques that are put forward in this thesis can identify *potential* causal relations. Below, we delve into the debate surrounding causal relations. As the debate is ongoing, we refer to most discovered relations in this dissertation as *potential causal relations*.

The *association* aspect is addressed by the existing techniques. Specifically, what the Chi-square and the post-hoc test can determine is if the observed data is significantly different from the expected data. This determines that there is an association between two variables, or, in other words, the two variables are related. Even more specific, the post-hoc tests show exactly which categorical pairs within both variables are associated with one another.

Due to the temporal nature of process mining we know the order of events. From that, we can infer that one variable precedes another variable. Therefore, we approximate adherence to the second aspect of *temporal order*. The temporal order describes the arrangement of events in time, i.e. does one activity follow another. Some theories require a more strict definition of temporal order. In that regard, another step would be to perform a/b testing, i.e. does smoking (A) lead to lung cancer (B) and does not smoking (not A) lead to the absence of lung cancer (not B)? This can be done using various theories that differ mainly in terms of strictness. Strictness here refers to how much counterevidence is needed to dispute the claim. To exemplify, does one counter-example of not a leading to b suffice to dispute the claim that a and b are not causally related? The theory of counterfactual conditional would argue so (Lewis, 1979). This theory supposes that smoking and lung cancer are not causally dependent, we can find examples of people that smoke but do not have lung cancer. Therefore, the theory of probabilistic causation adjusts for this by introducing the concept of likelihood rather than determinism (Pearl et al., 2000). Many theories address the concept of causality, each setting out its requirements for proof.

The final aspect of causality describes the possibilities of *alternative explanations*. An alternative explanation here means that a variable can explain the effect better than the cause variable. There are several ways to test for alternative explanations. One important approach is discussed in Chapter 6 where we try to identify and propose ways to deal with confounding variables. Confounding variables are variables that provide an alternative explanation that explains the discovered relation better. However, there is always a danger that the data for such a confounding variable is not captured in a data set. Especially in the settings of our case studies, this poses a real threat.

In conclusion, the techniques in this dissertation adhere to the three aspects of causality if interpreted in a not overly strict way. Once a stricter definition of the terms is adopted, the techniques have their limitations. As such, in this dissertation, we cannot prove causality in the ultimate sense with the proposed techniques. Thus, the term *potential causal relations* is used throughout the dissertation.

8.4 Future Work

The work in this dissertation can be built upon in two main ways: (1) technical work and (2) practical work. We see a lot of potential for pursuing the work in more *technical* terms. Based on the limitations addressed in the chapters, the main challenge, and the open issue discussed above, we can infer some of the directions for the future work of this dissertation. Specifically in terms of causality, generalizability, and understandability. Next to that, the work can be extended in a more *practical* way. Here, valorization is the main focus on how to develop the work of this dissertation further.

8.4.1 Technical Work

The most fruitful direction for future work in terms of the techniques is to pursue the line of *causality*. Ideally, future work pushes the boundaries of translating the findings of this technique to fit the more strict interpretation of causal relations. In an ideal scenario, we would also be able to indicate how certain it is that the relation discovered is causal. In terms of association, the common practice is to report significance values. An interesting idea would be to try and present an enhanced significance score that weighs in the other two aspects of temporal order and alternative explanations. Working on causality can be extremely valuable as it provides organizations with even more actionable insights to help improve their work processes. The enhanced significance score could help rank the improvement opportunities as it would indicate which improvements have the highest likelihood of having a real effect.

Next to that, building on the presented work in terms of *generalizability* is promising and exciting. Working outside the healthcare sector can be extremely interesting and present new and unexpected challenges. As we point out in Chapter 1, there are many more domains that could benefit from the techniques presented in this dissertation. We specifically see opportunities for applications in finance, customer service, and supply chain management. In these sectors, there is often a choice between activities and the impact is immediate. For example, in customer service, various strategies

can be used to handle incoming complaints. Understanding what the impact of each strategy is on the satisfaction of the customer is extremely valuable for organizations.

Future work can also focus on evaluating the proposed guidelines for qualitative evaluations. These have proven to work in other fields of study, mostly social science-related. However, the *understandability* of these guidelines in a process mining context is yet to be proven. Especially interesting is to focus on the degree to which these guidelines should be specified in detail or if a more open framework would aid researchers and practitioners as well. Next to that, identifying specific best practices from already conducted process mining projects might help make concrete what specific actions can be performed to better perform the evaluations.

8.4.2 Practical Work

The most promising direction for *valorization* would be to develop the techniques in such a way that they are proven to generate value for organizations. One way of doing this is by developing it in such a way that the technique becomes an extra tool in a process mining toolbox. In that way, process analysts can use it to further enhance their analysis when performing a process mining project within an organization. For this, it is important that the techniques are developed in a more user-friendly manner. As they are now, the techniques are accessible to a small, technical, audience. To make them more accessible, more intuitive interaction with the techniques needs to be facilitated. This can be done in several ways. One fruitful way of development is to enhance the output of the techniques by creating interactive graphs. Another, more generic, approach is to convert the techniques to a plug-in for existing process mining tools.

In this dissertation, we set out to help both organizations and process mining researchers. On the one hand, we hope to help organizations understand the impact of their activities to become more efficient and deal with the challenges they are faced with. On the other hand, we hope to aid process mining analysts in performing valid evaluations at the end of their projects. However, the research work is never finished, it will remain a *Work in Process*. In the meantime, it is my hope that this dissertation has a positive, however small, impact both on the research as well as the organizational community.

Appendices

APPENDIX A

Insights into Responses to Aggressive Behavior

Below, we present a detailed summary of the insights we have gathered throughout the various studies on aggressive behavior. We do this in three: Figure A.1, A.2, and A.3. We expand on the higher-level overview presented in Chapter 8 in two ways. First, we add the *specific observations* per caretaker response. In total, we have eleven observations for six out of the eight responses we studied. Second, in Chapter 6 we go into detail about *possible confounding variables*. The technique proposed in that work is also evaluated using the aggressive behavior data. For some of the relations that are defined in the three observations above we checked if they held if we differentiated between mild and severe incidents in terms of the perceived severity score of a caretaker. One can imagine that the perceived severity of an incident influences how effective the response of a caretaker is. The reasoning here is that the more severe an incident is, the less it might matter how a caretaker responds. In Figure 8.2 we mark the relations we checked with an asterisk (*). In the detailed explanation, behind the asterisk mark, we explain how the severity score impacts the relation.

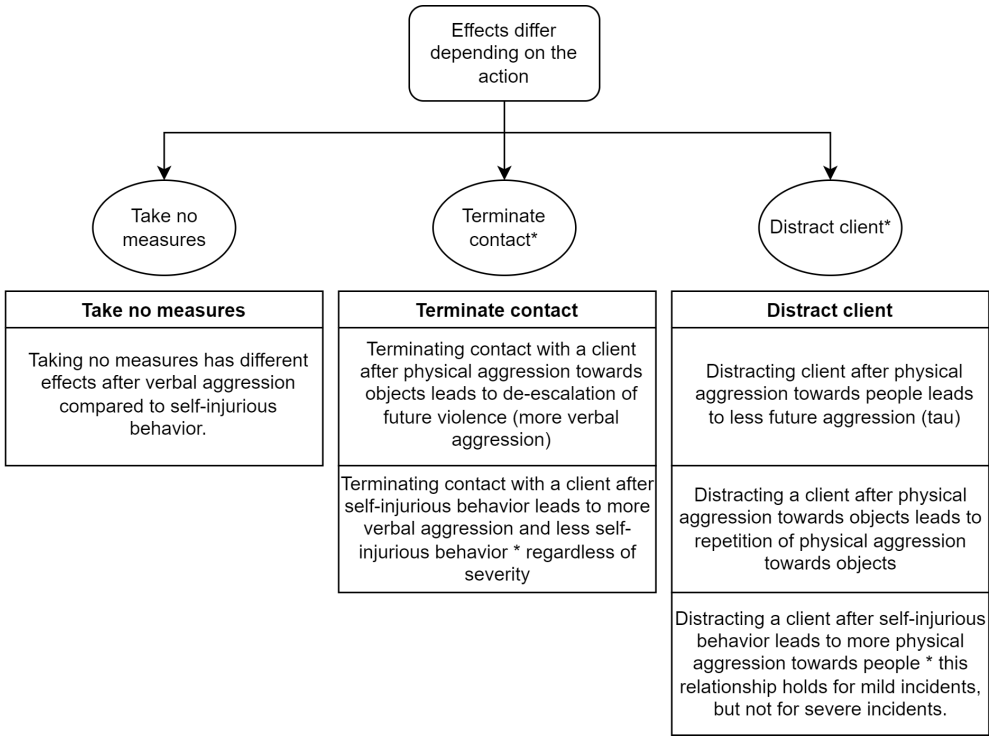


Figure A.1: Specific observations regarding the responses of caretaker that have a *different effect depending on the preceding action*

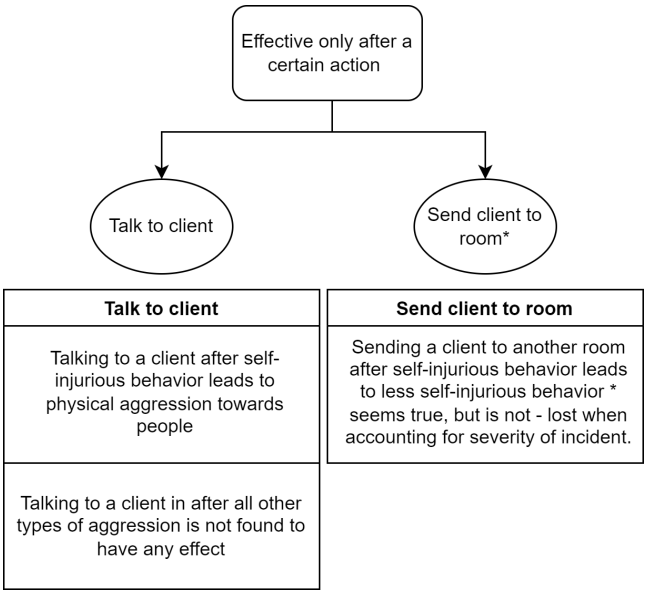


Figure A.2: Specific observations regarding the responses of caretaker that are *only effective after a certain action*

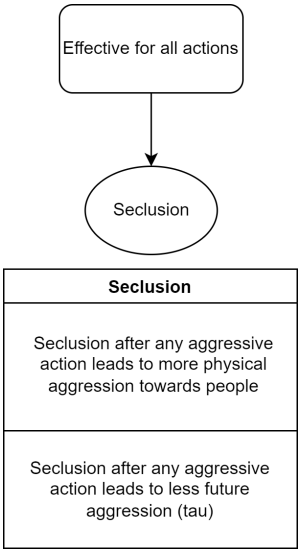


Figure A.3: Specific observations regarding the responses of caretaker that have the *same effect regardless of the preceding action*

Bibliography

Abo-Hamad, W. (2017), "Patient pathways discovery and analysis using process mining techniques: an emergency department case study", in: *International Conference on Health Care Systems Engineering*, Springer, pp. 209–219.

Adadi, A. & M. Berrada (2018), "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)", *IEEE Access*, vol. 6, pp. 52138–52160.

Agresti, A. (2003), *Categorical data analysis*, vol. 482, John Wiley & Sons.

Allen, M. (2017), *The SAGE encyclopedia of communication research methods*, SAGE Publications.

Alvarez, C., E. Rojas, M. Arias, J. Munoz-Gama, M. Sepúlveda, V. Herskovic & D. Capurro (2018), "Discovering role interaction models in the emergency room using process mining", *Journal of biomedical informatics*, vol. 78, pp. 60–77.

Andrews, R., C. G. van Dun, M. T. Wynn, W. Kratsch, M. Röglinger & A. H. ter Hofstede (2020a), "Quality-informed semi-automated event log generation for process mining", *Decision Support Systems*, vol. 132, p. 113265.

Andrews, R., M. T. Wynn, K. Vallmuur, A. H. Ter Hofstede & E. Bosley (2020b), "A comparative process mining analysis of road trauma patient pathways", *International journal of environmental research and public health*, vol. 17, no. 10, p. 3426.

Astromskis, S., A. Janes & M. Mairegger (2015), "A process mining approach to measure how users interact with software: an industrial case study", in: *Proceedings of the 2015 International Conference on Software and System Process*, pp. 137–141.

Augusto, A., R. Conforti, M. Dumas & M. La Rosa (2017), "Split miner: discovering accurate and simple business process models from event logs", in: *2017 IEEE International Conference on Data Mining (ICDM)*, IEEE, pp. 1–10.

Augusto, A., R. Conforti, M. Dumas, M. La Rosa, F. M. Maggi, A. Marrella, M. Mecella & A. Soo (2018), "Automated discovery of process models from event logs: review and benchmark", *IEEE TKDE*, vol. 31, no. 4, pp. 686–705.

Baier, T. & J. Mendling (2013), "Bridging abstraction layers in process mining: event to activity mapping", in: *Enterprise, Business-Process and Information Systems Modeling*, Springer, pp. 109–123.

Baier, T., J. Mendling & M. Weske (2014), "Bridging abstraction layers in process mining", *Information Systems*, vol. 46, pp. 123–139.

Bakhshandeh, M., D. M. Schunselaar, H. Leopold & H. A. Reijers (2017), "Predicting treatment repetitions in the implant denture therapy process", in: *2017 IEEE International Conference on Big Data (Big Data)*, IEEE, pp. 1259–1264.

Barnes, J. (1991), *Complete Works (Aristotle)*, Princeton University Press.

Beerepoot, I. M., N Martin & J. J. Koorn (Under review), "Evaluating process mining insights with healthcare professionals: the FEI funnel", *Business Process Management Journal*.

Beerepoot, I. M., J. J. Koorn, I. Van de Weerd, B. Van den Hooff, H. Leopold & H. A. Reijers (2019), "Working around health information systems: the role of power", in: *Proceedings of the 40th International Conference on Information Systems (ICIS)*, p. 2.

Benevento, E., P. M. Dixit, M. F. Sani, D. Aloini & W. M. Van der Aalst (2019), "Evaluating the effectiveness of interactive process discovery in healthcare: a case study", in: *BPM*, Springer, pp. 508–519.

Bernardi, M. L., M. Cimitile, C. Di Francescomarino & F. M. Maggi (2014), "Using discriminative rule mining to discover declarative process models with non-atomic activities", in: *International Symposium on Rules and Rule Markup Languages for the Semantic Web*, Springer, pp. 281–295.

Binder, M., W. Dorda, G. Duftschmid, R. Dunkl, K. A. Fröschl, W. Gall, W. Grossmann, K. Harmankaya, M. Hronsky, S. Rinderle-Ma, et al. (2012), "On analyzing process compliance in skin cancer treatment: an experience report from the evidence-based medical compliance cluster (ebmc 2)", in: *International Conference on Advanced Information Systems Engineering*, Springer, pp. 398–413.

Birt, L., S. Scott, D. Cavers, C. Campbell & F. Walter (2016), "Member checking: a tool to enhance trustworthiness or merely a nod to validation?", *Qualitative health research*, vol. 26, no. 13, pp. 1802–1811.

Bozorgi, Z. D., I. Teinemaa, M. Dumas, M. La Rosa & A. Polyvyanyy (2020), "Process mining meets causal machine learning: discovering causal rules from event logs", in: *2020 2nd International Conference on Process Mining (ICPM)*, IEEE, pp. 129–136.

Brookhart, M. A., T. Stürmer, R. J. Glynn, J. Rassen & S. Schneeweiss (2010), "Confounding control in healthcare database research: challenges and potential approaches", *Medical care*, vol. 48, no. 6 0, S114.

Brosnan, J. & O. Healy (2011), "A review of behavioral interventions for the treatment of aggression in individuals with developmental disabilities", *Research in developmental disabilities*, vol. 32, no. 2, pp. 437–446.

Broucke, S. K. vanden & J. De Weerdt (2017), "Fodina: a robust and flexible heuristic process discovery technique", *decision support systems*, vol. 100, pp. 109–118.

Brunk, J., M. Stierle, L. Papke, K. Revoredo, M. Matzner & J. Becker (2020), "Cause vs. effect in context-sensitive prediction of business process instances", *Information Systems*, vol. 95, p. 101635.

Buchanan, D. A. & A. A. Huczynski (2019), *Organizational behaviour*, Pearson UK.

Buijs, J. C., B. F. van Dongen & W. M. P. Van der Aalst (2012), "A genetic algorithm for discovering process trees", in: *2012 IEEE Congress on Evolutionary Computation*, IEEE, pp. 1–8.

Bygstad, B. & B. E. Munkvold (2011), "In search of mechanisms. conducting a critical realist data analysis", in: *Proceedings of the 32nd International Conference on Information Systems (ICIS'11)*, AIS.

Calvanese, D., T. E. Kalayci, M. Montali & S. Tinella (2017), "Ontology-based data access for extracting event logs from legacy data: the onprom tool and methodology", in: *International Conference on Business Information Systems*, Springer, pp. 220–236.

Cho, M., M. Song & S. Yoo (2014), "A systematic methodology for outpatient process analysis based on process mining", in: *Asia-Pacific Conference on Business Process Management*, Springer, pp. 31–42.

Cho, M., M. Song, S. Yoo & H. A. Reijers (2019), "An evidence-based decision support framework for clinician medical scheduling", *IEEE Access*, vol. 7, pp. 15239–15249.

Ciccio, C. D. & M. Mecella (2015), "On the discovery of declarative control flows for artful processes", *ACM Transactions on Management Information Systems (TMIS)*, vol. 5, no. 4, pp. 1–37.

Cochran, W. G. (1952), "The $\chi 2$ test of goodness of fit", *The Annals of Mathematical Statistics*, pp. 315–345.

Constantiou, I. D. & J. Kallinikos (2015), "New games, new rules: big data and the changing context of strategy", *Journal of Information Technology*, vol. 30, no. 1, pp. 44–57.

Cooper, S.-A., E Smiley, A Jackson, J Finlayson, L Allan, D Mantry & J Morrison (2009), "Adults with intellectual disabilities: prevalence, incidence and remission of aggressive behaviour and related factors", *Journal of Intellectual Disability Research*, vol. 53, no. 3, pp. 217–232.

Creswell, J. W. (2013), *Research design: Qualitative, quantitative, and mixed methods approaches*, SAGE Publications.

Creswell, J. W. & C. N. Poth (2016), *Qualitative inquiry and research design: Choosing among five approaches*, SAGE publications.

Crocker, A. G., C. Mercier, Y. Lachapelle, A. Brunet, D. Morin & M.-E. Roy (2006), "Prevalence and types of aggressive behaviour among adults with intellectual disabilities", *Journal of intellectual disability research*, vol. 50, no. 9, pp. 652–661.

Dani, V. S., H. Leopold, J. M. E. M. Van der Werf, X. Lu, I. M. Beerepoot, J. J. Koorn & H. A. Reijers (2021a), "Towards understanding the role of the human in event log extraction", in: *19th conference on Business Process Management*, *17th International Workshop on Business Process Intelligence (BPI'21)*.

Dani, V. S., E. Mahendrawathi, J. J. Koorn, J. M. E. M. Van der Werf, H. Leopold & H. A. Reijers (2021b), "Pair modeling: does one plus one add up?", in: *19th conference on Business Process Management*, *14th International Workshop on Social and Human Aspects of Business Process Management (BPMS2'21)*.

De Alvarenga, S. C., S. Barbon Jr, R. S. Miani, M. Cukier & B. B. Zarpelão (2018), "Process mining and hierarchical clustering to help intrusion alert visualization", *Computers & Security*, vol. 73, pp. 474–491.

De Leoni, M. & W. M. P. Van der Aalst (2013), "Data-aware process mining: discovering decisions in processes using alignments", in: *Proceedings of the 28th annual ACM symposium on applied computing*, pp. 1454–1461.

De Smedt, J., J. De Weerdt & J. Vanthienen (2015), "Fusion miner: process discovery for mixed-paradigm models", *Decision Support Systems*, vol. 77, pp. 123–136.

De Weerdt, J., A. Schupp, A. Vanderloock & B. Baesens (2013), "Process mining for the multi-faceted analysis of business processes—a case study in a financial services organization", *Computers in Industry*, vol. 64, no. 1, pp. 57–67.

Dees, M., M. de Leoni & F. Mannhardt (2017), “Enhancing process models to improve business performance: a methodology and case studies”, in: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, Springer, pp. 232–251.

Deng, X., J. Khuntia & K. Ghosh (2013), “Psychological empowerment of patients with chronic diseases: the role of digital integration”.

Denisov, V., D. Fahland & W. M. Van der Aalst (2018), “Unbiased, fine-grained description of processes performance from event data”, in: *BPM*, Springer, pp. 139–157.

Deokar, A. V. & J. Tao (2020), “Orgminer: a framework for discovering user-related process intelligence from event logs”, *Information Systems Frontiers*, pp. 1–20.

Dixit, P. M., H. Verbeek, J. C. Buijs & W. M. Van der Aalst (2018), “Interactive data-driven process model construction”, in: *International Conference on Conceptual Modeling*, Springer, pp. 251–265.

Du, M., N. Liu & X. Hu (2019), “Techniques for interpretable machine learning”, *Communications of the ACM*, vol. 63, no. 1, pp. 68–77.

Duma, D. & R. Aringhieri (2017), “Mining the patient flow through an emergency department to deal with overcrowding”, in: *International Conference on Health Care Systems Engineering*, Springer, pp. 49–59.

Dumas, M., M. La Rosa, J. Mendling, H. A. Reijers, et al. (2013), *Fundamentals of business process management*, vol. 1, Springer.

Eck, M. L. van, N. Sidorova & W. M. P. Van der Aalst (2016), “Discovering and exploring state-based models for multi-perspective processes”, in: *International Conference on Business Process Management*, Springer, pp. 142–157.

– (2017), “Guided interaction exploration in artifact-centric process models”, in: *2017 IEEE 19th Conference on Business Informatics (CBI)*, vol. 1, IEEE, pp. 109–118.

Elhadjamor, E. A. & S. A. Ghannouchi (2019), “Analyze in depth health care business process and key performance indicators using process mining”, *Procedia Computer Science*, vol. 164, pp. 610–617.

Emamjome, F., R. Andrews & A. H. ter Hofstede (2019), “A case study lens on process mining in practice”, in: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, Springer, pp. 127–145.

Emerson, R. W. (1860), *The Conduct of Life*, Ticknor and Fields.

Engel, R., W. M. Van der Aalst, M. Zapletal, C. Pichler & H. Werthner (2012), “Mining inter-organizational business process models from edi messages: a case study from the automotive sector”, in: *CAiSE*, Springer, pp. 222–237.

Epure, E. V., C. Hug, R. Deneckere & S. Brinkkemper (2014), “What shall i do next?”, in: *CAiSE*, Springer, pp. 473–487.

Epure, E. V., P. Martín-Rodilla, C. Hug, R. Deneckère & C. Salinesi (2015), “Automatic process model discovery from textual methodologies”, in: *2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS)*, IEEE, pp. 19–30.

Erdem, S. & O. Demirörs (2017), “An exploratory study on usage of process mining in agile software development”, in: *International Conference on Software Process Improvement and Capability Determination*, Springer, pp. 187–196.

Fei, H., N. Meskens, et al. (2010), "Discovering patient care process models from event logs", in: *Proc. 8th Int. Conf. Modeling Simulation (MOSIM)*, Citeseer, pp. 10–12.

Ferreira, D., M. Zacarias, M. Malheiros & P. Ferreira (2007), "Approaching process mining with sequence clustering: experiments and findings", in: *BPM*, Springer, pp. 360–374.

Fisher, R. A. & F. Yates (1938), *Statistical tables: For biological, agricultural and medical research*, Oliver and Boyd.

Gallupe, B., G. DeSanctis & G. W. Dickson (1986), "The impact of computer-based support on the process and outcomes of group decision making", in: ICIS 1986 Proceedings.

Gatta, R., M. Vallati, C. Fernandez-Llatas, A. Martinez-Millana, S. Orini, L. Sacchi, J. Lenkowicz, M. Marcos, J. Munoz-Gama, M. Cuendet, et al. (2019), "Clinical guidelines: a crossroad of many research areas. challenges and opportunities in process mining for healthcare", in: *International Conference on Business Process Management*, Springer, pp. 545–556.

Ghasemaghaei, M., S. Ebrahimi & K. Hassanein (2016), "Generating valuable insights through data analytics: a moderating effects model", in: *Proceedings of the 37th International Conference on Information Systems (ICIS'16)*, AIS.

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

Gorissen, P. (2017), "Business process reporting using process mining, analytic workflows and process cubes: a case study in education", in: *Data-Driven Process Discovery and Analysis: 5th IFIP WG 2.6 International Symposium, SIMPDA 2015, Vienna, Austria, December 9-11, 2015, Revised Selected Papers*, vol. 244, Springer, p. 28.

Greenland, S., J. Pearl & J. M. Robins (1999), "Confounding and collapsibility in causal inference", *Statistical science*, vol. 14, no. 1, pp. 29–46.

Gubela, R. M. & S. Lessmann (2020), "Uplift forest for multiple treatments and continuous outcomes", in: *Proceedings of the 41st International Conference on Information Systems (ICIS)*, AIS.

Gunnarsson, B. R., S. K. vanden Broucke & J. De Weerdt (2019), "Predictive process monitoring in operational logistics: a case study in aviation", in: *BPM*, Springer, pp. 250–262.

Günther, C. W. & W. M. P. Van der Aalst (2007), "Fuzzy mining–adaptive process simplification based on multi-perspective metrics", in: *International conference on business process management*, Springer, pp. 328–343.

Gupta, N., K. Anand & A. Sureka (2015), "Pariket: mining business process logs for root cause analysis of anomalous incidents", in: *International Workshop on Databases in Networked Information Systems*, Springer, pp. 244–263.

Hashem, I. A. T., I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani & S. U. Khan (2015), "The rise of "big data" on cloud computing: review and open research issues", *Information systems*, vol. 47, pp. 98–115.

Hastings, R. P. (2002), "Do challenging behaviors affect staff psychological well-being? issues of causality and mechanism", *American Journal on Mental Retardation*, vol. 107, no. 6, pp. 455–467.

Haupt, J., D. Jacob, R. M. Gubela & S. Lessmann (2019), "Affordable uplift: supervised randomization in controlled experiments", in: *Proceedings of the 40th International Conference on Information Systems (ICIS'19)*, AIS.

Haynes, W. (2013), "Bonferroni correction", in: *Encyclopedia of Systems Biology*, ed. by Dubitzky, W., Wolkenhauer, O., Cho, K.-H. & Yokota, H., New York, NY: Springer New York, pp. 154–154.

Hensel, J. M., Y. Lunsky & C. S. Dewa (2014), "Staff perception of aggressive behaviour in community services for adults with intellectual disabilities", *Community mental health journal*, vol. 50, no. 6, pp. 743–751.

Hensel, J., Y. Lunsky & C. S. Dewa (2012), "Exposure to client aggression and burnout among community staff who support adults with intellectual disabilities in ontario, canada", *Journal of Intellectual Disability Research*, vol. 56, no. 9, pp. 910–915.

Hessey, S. & W. Venters (2016), "Sensing distress–towards a blended method for detecting and responding to problematic customer experience events", in: *International Conference on HCI in Business, Government, and Organizations*, Springer, pp. 395–405.

Hevner, A. R., S. T. March, J. Park & S. Ram (2004), "Design science in information systems research", *MIS quarterly*, pp. 75–105.

Hompes, B. F., A. Maaradji, M. La Rosa, M. Dumas, J. C. Buijs & W. M. Van der Aalst (2017), "Discovering causal factors explaining business process performance variation", in: *CAiSE*, Springer, pp. 177–192.

Huckshorn, K. A. (2004), "Reducing seclusion & restraint use in mental health settings: core strategies for prevention", *Journal of psychosocial nursing and mental health services*, vol. 42, no. 9, pp. 22–33.

Jackson, K. & P. Bazeley (2019), *Qualitative data analysis with NVivo*, SAGE Publications.

Jalali, A. (2014), "Aspect mining in business process management", in: *International Conference on Business Informatics Research*, Springer, pp. 246–260.

Johnson, O. A., T. B. Dhafari, A. Kurniati, F. Fox & E. Rojas (2018), "The clearpath method for care pathway process mining and simulation", in: *BPM*, Springer, pp. 239–250.

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

Kitchenham, B. & S. Charters (2007), "Guidelines for performing systematic literature reviews in software engineering".

Knoll, D., G. Reinhart & M. Prüglmeier (2019), "Enabling value stream mapping for internal logistics using multidimensional process mining", *Expert Systems with Applications*, vol. 124, pp. 130–142.

Koorn, J. J., I. M. Beerepoot, V. S. Dani, X. Lu, I. Van de Weerd, H. Leopold & H. A. Reijers (2021), "Bringing rigor to the qualitative evaluation of process mining findings: an analysis and a proposal", in: *2021 3rd International Conference on Process Mining (ICPM)*, IEEE, pp. 120–127.

Koorn, J. J., H. Leopold & H. A. Reijers (2018a), "A task framework for predicting the effects of automation", in: *39th International Conference on Information Systems,*

ICIS 2018, 6th International Workshop on the Changing Nature of Work (CNoW): Bridging the Workplace of People, Data and Things.

Koorn, J. J., S. Zhang & J. V. Nickerson (2019a), "A fresh look at the impact of technology on occupations through software", in: *40th International Conference on Information Systems, ICIS 2019, 8th International Workshop on the Changing Nature of Work (CNoW): Digital Ecosystems and the Changing Nature of Work - Reconfiguring technology, practices, and organizations*.

Koorn, J. J., X. Lu, H. Leopold & H. A. Reijers (2019b), "Towards understanding aggressive behavior in residential care facilities using process mining", in: *International Conference on Conceptual Modeling*, Springer, pp. 135–145.

Koorn, J. J., X. Lu, H. Leopold & H. A. Reijers (2020), "Looking for meaning: discovering action-response-effect patterns in business processes", in: *International Conference on Business Process Management*, Springer, pp. 167–183.

Koorn, J. J., X. Lu, H. Leopold & H. A. Reijers (2022a), "From action to response to effect: mining statistical relations in work processes", *Information Systems*, DOI: `https://doi.org/10.1016/j.is.2022.102035`.

Koorn, J. J., X. Lu, H. Leopold & H. A. Reijers (2022b), "Mining statistical relations for better decision making in healthcare processes".

Koorn, J. J., X. Lu, F. Mannhardt, H. Leopold & H. A. Reijers (2022c), "Uncovering complex relations in patient pathways based on statistics: the impact of clinical actions", in: *Proceedings of the 55th Hawaii International Conference on System Sciences*.

Koorn, J. J., H. Leopold & H. A. Reijers (2018b), "A task framework for predicting the effects of automation", in: *Proceedings of the 26th European Conference on Information Systems (ECIS)*, p. 141.

Kurniati, A. P., C. McInerney, K. Zucker, G. Hall, D. Hogg & O. Johnson (2019), "A multi-level approach for identifying process change in cancer pathways", in: *BPM*, Springer, pp. 595–607.

Lee, C., K. L. Choy, G. T. Ho & C. H. Lam (2016), "A slippery genetic algorithm-based process mining system for achieving better quality assurance in the garment industry", *Expert systems with applications*, vol. 46, pp. 236–248.

Leemans, S. J., D. Fahland & W. M. P. Van der Aalst (2013), "Discovering block-structured process models from event logs-a constructive approach", in: *International conference on applications and theory of Petri nets and concurrency*, Springer, pp. 311–329.

Lehto, T., M. Hinkka & J. Hollmén (2016), "Focusing business improvements using process mining based influence analysis", in: *BPM*, Springer, pp. 177–192.

Lewis, D. (1979), "Counterfactual dependence and time's arrow", *Noûs*, vol. 13, no. 4, pp. 455–476, URL: `http://www.jstor.org/stable/2215339`.

Li, C., M. Reichert & A. Wombacher (2008), "Mining process variants: goals and issues", in: *2008 IEEE International Conference on Services Computing*, vol. 2, IEEE, pp. 573–576.

– (2011), "Mining business process variants: challenges, scenarios, algorithms", *Data & Knowledge Engineering*, vol. 70, no. 5, pp. 409–434.

Lincoln, Y., Y. Guba & E. Guba (1985), *Naturalistic Inquiry*, SAGE Publications, ISBN: 9780803924314.

Lloyd, B. P. & C. H. Kennedy (2014), “Assessment and treatment of challenging behaviour for individuals with intellectual disability: a research review”, *Journal of Applied Research in Intellectual Disabilities*, vol. 27, no. 3, pp. 187–199.

Lu, X., M. Nagelkerke, D. Van de Wiel & D. Fahland (2015), “Discovering interacting artifacts from erp systems”, *IEEE Transactions on Services Computing*, vol. 8, no. 6, pp. 861–873.

Ly, L. T., C. Indiono, J. Mangler & S. Rinderle-Ma (2012), “Data transformation and semantic log purging for process mining”, in: *CAiSE*, Springer, pp. 238–253.

Maggi, F. M., T. Slaats & H. A. Reijers (2014), “The automated discovery of hybrid processes”, in: *International Conference on Business Process Management*, Springer, pp. 392–399.

Mahendrawathi, E., H. M. Astuti & A. Nastiti (2015a), “Analysis of customer fulfilment with process mining: a case study in a telecommunication company”, *Procedia Computer Science*, vol. 72, pp. 588–596.

Mahendrawathi, E., H. M. Astuti & I. R. K. Wardhani (2015b), “Material movement analysis for warehouse business process improvement with process mining: a case study”, in: *Asia-Pacific Conference on Business Process Management*, Springer, pp. 115–127.

Mahendrawathi, E., S. O. Zayin & F. J. Pamungkas (2017), “Erp post implementation review with process mining: a case of procurement process”, *Procedia Computer Science*, vol. 124, pp. 216–223.

Mannhardt, F. & D. Blinde (2017), “Analyzing the trajectories of patients with sepsis using process mining.”, in: *RADAR+ EMISA@ CAiSE*, pp. 72–80.

Mannhardt, F., M. De Leoni, H. A. Reijers & W. M. P. Van der Aalst (2016), “Decision mining revisited-discovering overlapping rules”, in: *International Conference on Advanced Information Systems Engineering*, Springer, pp. 377–392.

Mans, R., H. A. Reijers, M. van Genuchten & D. Wismeijer (2012a), “Mining processes in dentistry”, in: *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, ACM, pp. 379–388.

Mans, R. S., H. A. Reijers, D. Wismeijer & M. van Genuchten (2013), “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.

Mans, R. S., M. H. Schonenberg, M. Song, W. M. P. Van der Aalst & P. J. M. Bakker (2008), “Application of process mining in healthcare–a case study in a dutch hospital”, in: *International joint conference on biomedical engineering systems and technologies*, Springer, pp. 425–438.

Mans, R. S., W. M. P. Van der Aalst & R. J. B. Vanwersch (2015), *Process mining in healthcare: evaluating and exploiting operational healthcare processes*, Springer.

Mans, R. S., W. M. Van der Aalst, R. J. Vanwersch & A. J. Moleman (2012b), “Process mining in healthcare: data challenges when answering frequently posed questions”, in: *Process Support and Knowledge Representation in Health Care*, Springer, pp. 140–153.

Marazza, F., F. A. Bukhsh, O. Vijlbrief, J. Geerdink, S. Pathak, M. van Keulen & C. Seifert (2019), “Comparing process models for patient populations: application in breast cancer care”, in: *BPM*, Springer, pp. 496–507.

Maris, A., R. Bijvank & P. Ravesteyn (2016), “The applicability of process mining to determine and align process model descriptions.”, in: *Bled eConference*, p. 43.

Martens, D., F. Provost, J. Clark & E. J. de Fortuny (2016), “Mining massive fine-grained behavior data to improve predictive analytics.”, *MIS Quarterly*, vol. 40, no. 4.

Martin, N., J. De Weerdt, C. Fernández-Llatas, A. Gal, R. Gatta, G. Ibáñez, O. Johnson, F. Mannhardt, L. Marco-Ruiz, S. Mertens, et al. (2020), “Recommendations for enhancing the usability and understandability of process mining in healthcare”, *Artificial Intelligence in Medicine*, vol. 109, p. 101962.

Martin, N., A. Martinez-Millana, B. Valdivieso & C. Fernández-Llatas (2019), “Interactive data cleaning for process mining: a case study of an outpatient clinic’s appointment system”, in: *BPM*, Springer, pp. 532–544.

McClintock, K, S Hall & C. Oliver (2003), “Risk markers associated with challenging behaviours in people with intellectual disabilities: a meta-analytic study”, *Journal of Intellectual Disability Research*, vol. 47, no. 6, pp. 405–416.

McHugh, M. L. (2013), “The chi-square test of independence”, *Biochemia medica: Biochemia medica*, vol. 23, no. 2, pp. 143–149.

McKenna, E. (2020), *Business psychology and organizational behaviour*, Routledge.

Mikalef, P., I. O. Pappas, J. Krogstie & M. Giannakos (2018), “Big data analytics capabilities: a systematic literature review and research agenda”, *Information Systems and e-Business Management*, vol. 16, no. 3, pp. 547–578.

Miles, M. B. & A. M. Huberman (1994), *Qualitative data analysis: An expanded sourcebook*, SAGE Publications.

Miller, R. G. J. (1981), *Simultaneous statistical inference*, Springer Series in Statistics.

Mills, S & J Rose (2011), “The relationship between challenging behaviour, burnout and cognitive variables in staff working with people who have intellectual disabilities”, *Journal of Intellectual Disability Research*, vol. 55, no. 9, pp. 844–857.

Montali, M., F. M. Maggi, F. Chesani, P. Mello & W. M. Van der Aalst (2014), “Monitoring business constraints with the event calculus”, *ACM transactions on intelligent systems and technology (TIST)*, vol. 5, no. 1, pp. 1–30.

Mueller-Wickop, N. & M. Schultz (2013), “Erp event log preprocessing: timestamps vs. accounting logic”, in: *International Conference on Design Science Research in Information Systems*, Springer, pp. 105–119.

Nguyen, H., M. Dumas, A. H. ter Hofstede, M. La Rosa & F. M. Maggi (2017), “Mining business process stages from event logs”, in: *International Conference on Advanced Information Systems Engineering*, Springer, pp. 577–594.

Nieuwenhuis, J. G., E. O. Noorthoorn, H. L. I. Nijman, P. Naarding & C. L. Mulder (2017), “A blind spot? screening for mild intellectual disability and borderline intellectual functioning in admitted psychiatric patients: prevalence and associations with coercive measures”, *PloS one*, vol. 12, no. 2, e0168847.

Nijman, H & T. Palmstierna (2002), “Measuring aggression with the staff observation aggression scale- revised”, *Acta Psychiatrica Scandinavica*, vol. 106, pp. 101–102.

Nijman, H. L., P. Muris, H. L. Merckelbach, T. Palmstierna, B. Wistedt, A. Vos, A. van Rixtel & W. Allertz (1999), “The staff observation aggression scale–revised (soas-r)”, *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, vol. 25, no. 3, pp. 197–209.

Noda, T, H Nijman, N Sugiyama, K Tsujiwaki, H Putkonen, E Sailas, R Kontio, H Ito & G Joffe (2012), "Factors affecting assessment of severity of aggressive incidents: using the staff observation aggression scale–revised (soas-r) in japan", *Journal of psychiatric and mental health nursing*, vol. 19, no. 9, pp. 770–775.

Nooijen, E. H., B. F. van Dongen & D. Fahland (2012), "Automatic discovery of data-centric and artifact-centric processes", in: *International Conference on Business Process Management*, Springer, pp. 316–327.

Nørgaard, M., V. Ehrenstein & J. P. Vandenbroucke (2017), "Confounding in observational studies based on large health care databases: problems and potential solutions–a primer for the clinician", *Clinical epidemiology*, vol. 9, p. 185.

Okhuysen, G. A. & B. A. Bechky (2009), "Coordination in organizations: an integrative perspective", *Academy of Management annals*, vol. 3, no. 1, pp. 463–502.

Park, M., M. Song, T. H. Baek, S. Son, S. J. Ha & S. W. Cho (2015), "Workload and delay analysis in manufacturing process using process mining", in: *Asia-Pacific Conference on Business Process Management*, Springer, pp. 138–151.

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

Patton, M. Q. (1999), "Enhancing the quality and credibility of qualitative analysis.", *Health services research*, vol. 34, no. 5 Pt 2, p. 1189.

Paulk, M. C., B. Curtis, M. B. Chrissis & C. V. Weber (1993), "Capability maturity model, version 1.1", *IEEE software*, vol. 10, no. 4, pp. 18–27.

Pearl, J. (2009), *Causality*, Cambridge university press.

Pearl, J. & D. Mackenzie (2018), *The book of why: the new science of cause and effect*, Basic books.

Pearl, J. et al. (2000), "Models, reasoning and inference", *Cambridge, UK: CambridgeUniversityPress*, vol. 19.

Pérez-Castillo, R., J. A. Cruz-Lemus, I. G.-R. de Guzmán & M. Piattini (2012), "A family of case studies on business process mining using marble", *Journal of Systems and Software*, vol. 85, no. 6, pp. 1370–1385.

Pérez-Castillo, R., M. Fernández-Ropero & M. Piattini (2019), "Business process model refactoring applying ibuprofen. an industrial evaluation", *Journal of Systems and Software*, vol. 147, pp. 86–103.

Pérez-Castillo, R., B. Weber, J. Pinggera, S. Zugal, I. G.-R. de Guzmán & M. Piattini (2011), "Generating event logs from non-process-aware systems enabling business process mining", *Enterprise Information Systems*, vol. 5, no. 3, pp. 301–335.

Pini, A., R. Brown & M. T. Wynn (2015), "Process visualization techniques for multi-perspective process comparisons", in: *Asia-Pacific Conference on Business Process Management*, Springer, pp. 183–197.

Popova, V., D. Fahland & M. Dumas (2015), "Artifact lifecycle discovery", *International Journal of Cooperative Information Systems*, vol. 24, no. 01, p. 1550001.

Prathama, F., B. N. Yahya, D. D. Harjono & E. Mahendrawathi (2019), "Trace clustering exploration for detecting sudden drift: a case study in logistic process", *Procedia Computer Science*, vol. 161, pp. 1122–1130.

Prodel, M., V. Augusto, X. Xie, B. Jouaneton & L. Lamarsalle (2015), "Discovery of patient pathways from a national hospital database using process mining and integer

linear programming", in: *2015 IEEE International Conference on Automation Science and Engineering (CASE)*, IEEE, pp. 1409–1414.

Provost, F. & R. Kohavi (1998), "Guest editors' introduction: on applied research in machine learning", *Machine learning*, vol. 30, no. 2, pp. 127–132.

Reijers, H. A., M. Song & B. Jeong (2007), "On the performance of workflow processes with distributed actors: does place matter?", in: *BPM*, Springer, pp. 32–47.

– (2009), "Analysis of a collaborative workflow process with distributed actors", *Information Systems Frontiers*, vol. 11, no. 3, pp. 307–322.

Rhyn, M. & I. Blohm (2019), "Patterns of data-driven decision-making: how decision-makers leverage crowdsourced data", in: *Proceedings of the 40th International Conference on Information Systems (ICIS'19)*, AIS.

Richards, J. C. & R. W. Schmidt (2013), *Longman dictionary of language teaching and applied linguistics*, Routledge.

Rinner, C., E. Helm, R. Dunkl, H. Kittler & S. Rinderle-Ma (2018), "Process mining and conformance checking of long running processes in the context of melanoma surveillance", *International Journal of Environmental Research and Public Health*, vol. 15, no. 12.

Rojas, E., A. Cifuentes, A. Burattin, J. Munoz-Gama, M. Sepúlveda & D. Capurro (2019), "Performance analysis of emergency room episodes through process mining", *International journal of environmental research and public health*, vol. 16, no. 7, p. 1274.

Rojas, E., J. Munoz-Gama, M. Sepúlveda & D. Capurro (2016), "Process mining in healthcare: a literature review", *Journal of biomedical informatics*, vol. 61, pp. 224–236.

Rollinson, D. (2008), *Organisational behaviour and analysis: An integrated approach*, Pearson Education.

Roubtsova, E. & N. Wiersma (2018), "Involvement of business roles in auditing with process mining", in: *International Conference on Evaluation of Novel Approaches to Software Engineering*, Springer, pp. 24–44.

Rovani, M., F. M. Maggi, M. De Leoni & W. M. P. Van der Aalst (2015), "Declarative process mining in healthcare", *Expert Systems with Applications*, vol. 42, no. 23, pp. 9236–9251.

Sagiroglu, S. & D. Sinanc (2013), "Big data: a review", in: *2013 international conference on collaboration technologies and systems (CTS)*, IEEE, pp. 42–47.

Samalikova, J., R. J. Kusters, J. J. Trienekens & A. Weijters (2014), "Process mining support for capability maturity model integration-based software process assessment, in principle and in practice", *Journal of Software: Evolution and Process*, vol. 26, no. 7, pp. 714–728.

Saunders, M., P. Lewis & A. Thornhill (2009), *Research methods for business students*, Pearson education.

Schönig, S., A. Rogge-Solti, C. Cabanillas, S. Jablonski & J. Mendling (2016), "Efficient and customisable declarative process mining with sql", in: *International Conference on Advanced Information Systems Engineering*, Springer, pp. 290–305.

Shortliffe, E. (2012), *Computer-based medical consultations: MYCIN*, vol. 2, Elsevier.

Sirgmets, M., F. Milani, A. Nolte & T. Pungas (2018), "Designing process diagrams–a framework for making design choices when visualizing process mining outputs",

in: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, Springer, pp. 463–480.

Song, M, C Günther & W. Van der Aalst (2009), *Trace Clustering in Process Mining. In (Ardagna, D., Mecella, M., Yang, J., Hrsg.): Business Process Management Workshops. BPM 2008. LNBIP 17*.

Song, W., H.-A. Jacobsen, C. Ye & X. Ma (2015), "Process discovery from dependence-complete event logs", *IEEE Transactions on Services Computing*, vol. 9, no. 5, pp. 714–727.

Stake, R. E. (1995), *The art of case study research*, SAGE Publications.

Stuit, M. & H. Wortmann (2012), "Discovery and analysis of e-mail-driven business processes", *Information Systems*, vol. 37, no. 2, pp. 142–168.

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

Suriadi, S., C. Ouyang, W. M. P. Van der Aalst & A. H. M. ter Hofstede (2012), "Root cause analysis with enriched process logs", in: *International Conference on Business Process Management*, Springer, pp. 174–186.

Swinnen, J., B. Depaire, M. J. Jans & K. Vanhoof (2011), "A process deviation analysis–a case study", in: *BPM*, Springer, pp. 87–98.

Syamsiyah, A., A. Bolt, L. Cheng, B. F. Hompes, R. J. C. Bose, B. F. Van Dongen & W. M. Van der Aalst (2017), "Business process comparison: a methodology and case study", in: *International Conference on Business Information Systems*, Springer, pp. 253–267.

Taylor, P., M. Leida & B. Majeed (2011), "Case study in process mining in a multinational enterprise", in: *International Symposium on Data-Driven Process Discovery and Analysis*, Springer, pp. 134–153.

Tenneij, N. & H. M. Koot (2008), "Incidence, types and characteristics of aggressive behaviour in treatment facilities for adults with mild intellectual disability and severe challenging behaviour", *Journal of Intellectual Disability Research*, vol. 52, no. 2, pp. 114–124.

Thabet, D., S. A. Ghannouchi & H. H. B. Ghezala (2018), "A process mining-based solution for business process model extension with cost perspective context-based cost data analysis and case study", in: *IFIP International Conference on Computer Information Systems and Industrial Management*, Springer, pp. 434–446.

Thiede, M., D. Fuerstenau & A. P. B. Barquet (2018), "How is process mining technology used by organizations? a systematic literature review of empirical studies", *Business Process Management Journal*.

Tyrer, F, C. McGrother, C. Thorp, M Donaldson, S Bhaumik, J. Watson & C Hollin (2006), "Physical aggression towards others in adults with learning disabilities: prevalence and associated factors", *Journal of intellectual disability research*, vol. 50, no. 4, pp. 295–304.

Van den Bogaard, K. J., H. L. Nijman, T. Palmstierna & P. J. Embregts (2018), "Characteristics of aggressive behavior in people with mild to borderline intellectual disability and co-occurring psychopathology", *Journal of Mental Health Research in Intellectual Disabilities*, vol. 11, no. 2, pp. 124–142.

Van den Ingh, L., R. Eshuis & S. Gelper (2021), "Assessing performance of mined business process variants", *Enterprise Information Systems*, vol. 15, no. 5, pp. 676–693.

Van der Aalst, W. M. P. (2019), "A practitioner's guide to process mining: limitations of the directly-follows graph", *Procedia Computer Science*, vol. 164, pp. 321 –328.

Van der Aalst, W. M. P. (2016), "Data science in action", in: *Process Mining*, Springer, pp. 3–23.

Van der Aalst, W. M. P., A. Adriansyah, A. K. A. De Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. Van den Brand, R. Brandtjen, J. Buijs, et al. (2011), "Process mining manifesto", in: *International Conference on Business Process Management*, Springer, pp. 169–194.

Van Eck, M. L., X. Lu, S. J. Leemans & W. M. Van der Aalst (2015), "Pm2: a process mining project methodology", in: *CAiSE*, Springer, pp. 297–313.

Van Eck, M. L., N. Sidorova & W. M. Van der Aalst (2016), "Enabling process mining on sensor data from smart products", in: *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*, IEEE, pp. 1–12.

Van Langerak, R., J. M. E. M. Van der Werf & S. Brinkkemper (2017), "Uncovering the runtime enterprise architecture of a large distributed organisation", in: *CAiSE*, Springer, pp. 247–263.

Verbeek, H., W. M. P. Van der Aalst & J. Munoz-Gama (2017), "Divide and conquer: a tool framework for supporting decomposed discovery in process mining", *The Computer Journal*, vol. 60, no. 11, pp. 1649–1674.

Villars, R. L., C. W. Olofson & M. Eastwood (2011), "Big data: what it is and why you should care", *White paper, IDC*, vol. 14, pp. 1–14.

Vogelgesang, T. & H.-J. Appelrath (2016), "Pmcube: a data-warehouse-based approach for multidimensional process mining", in: *BPM*, Springer, pp. 167–178.

Wang, Y., F. Caron, J. Vanthienen, L. Huang & Y. Guo (2014), "Acquiring logistics process intelligence: methodology and an application for a chinese bulk port", *Expert Systems with Applications*, vol. 41, no. 1, pp. 195–209.

Weijters, A. J. & W. M. P. Van der Aalst (2003), "Rediscovering workflow models from event-based data using little thumb", *Integrated Computer-Aided Engineering*, vol. 10, no. 2, pp. 151–162.

Yahya, B. N., M. Song, H. Bae, S.-o. Sul & J.-Z. Wu (2016), "Domain-driven actionable process model discovery", *Computers & Industrial Engineering*, vol. 99, pp. 382–400.

Yang, S., W. Ni, X. Dong, S. Chen, R. A. Farneth, A. Sarcevic, I. Marsic & R. S. Burd (2018), "Intention mining in medical process: a case study in trauma resuscitation", in: *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, pp. 36–43.

Zijlmans, L. J., P. J. Embregts, A. M. Bosman & A. P. Willems (2012), "The relationship among attributions, emotions, and interpersonal styles of staff working with clients with intellectual disabilities and challenging behavior", *Research in Developmental Disabilities*, vol. 33, no. 5, pp. 1484–1494.

Summary

Organizations strive to understand and optimize how they achieve their organizational goals. To do this, organizations increasingly record data on the activities that are performed. Taking a work process lens, one can study how organizational activities are organized in work processes. *Process mining* is a field of research that discovers, analyzes, and improves work processes. This dissertation contributes to the field of process mining in two ways: 1) by presenting new ways of discovering and analyzing work processes, and 2) by formulating guidelines to support qualitative evaluations.

First, traditional process mining techniques use a *control-flow* perspective, meaning they consider the order of activities. This helps inform organizations what activities are performed and when. In this dissertation, we propose several techniques that use a causal perspective. These techniques help understand the impact of activities on future work processes and ultimately can help inform organizations about how to achieve their goals.

Second, this dissertation adds value by providing support for qualitative evaluations of process mining projects. Organizational experts are often involved in process mining projects, especially during the evaluation of results. In this dissertation, we propose a set of guidelines that help improve the validity of qualitative evaluation results if domain experts are involved in the evaluation process.

Samenvatting

Organisaties streven ernaar om de manier waarop ze doelen bereiken te begrijpen en te optimaliseren. In dit kader wordt er steeds meer data verzameld over de activiteiten binnen organisaties. Door een werkproces lens kan men bestuderen hoe verschillende activiteiten in een organisatie zich samen vormen tot een werkproces. *Process Mining* is een onderzoeksveld dat zich richt op het ontdekken, analyseren en verbeteren van werkprocessen. Dit proefschrift draagt op twee manieren bij aan het veld van *process mining*: 1) door nieuwe technieken te presenteren om werkprocessen te ontdekken en analyseren en 2) door richtlijnen te presenteren om kwalitatieve evaluaties van *process mining* projecten te ondersteunen.

Ten eerste, traditionele *process mining* technieken gebruiken een *control-flow* perspectief. Hierin wordt gekeken naar de volgorde van de activiteiten. Dit perspectief helpt organisaties om te ontdekken welke activiteiten wanneer plaatsvinden. In dit proefschrift presenteren we een aantal technieken die vanuit een causaal perspectief kijken naar werkprocessen. Deze technieken geven inzicht in de impact die activiteiten hebben op toekomstige werkactiviteiten en -processen. Deze technieken informeren organisaties over hoe ze hun doelen beter kunnen bereiken.

Ten tweede, draagt dit proefschrift bij door het verschaffen van ondersteuning voor het uitvoeren van kwalitatieve evaluaties tijdens *process mining* projecten. Organisatie experts zijn vaak betrokken bij evaluaties in *process mining* projecten. In dit proefschrift presenteren we een reeks richtlijnen die bijdragen aan het verbeteren van de validiteit van kwalitatieve evaluaties in *process mining* projecten waarin een externe expert is betrokken.

Curriculum Vitae

A.1 Personal Statement

I am a PhD candidate at the Utrecht University specializing in the field of Process Mining. The interaction between organizations and their IT environments is what interests me. Especially, how IT can be utilized to fulfill the maximum potential of an organization. My main focus lies with the analysis of business processes and strategy decisions in IT. From a broader perspective, I am intrigued by the economic and technical perspectives of automation and its effects on society. My ambition is to work on finding ways to connect the organizational and IT side to perform in an optimal way. Next to my work as PhD candidate I have developed my own startup (ARE miner) during my PhD program and have been involved in a variety of board positions representing PhD and employee interests.

A.2 Education

- University Teaching Qualification (BKO) from the Vrije Universiteit Amsterdam and Learn! Academy. (2019)
- Various professional development courses on data analysis, academic development, and research communication, see below for details. (2017-present)
- Master of Science, University of Amsterdam (UvA), Information Studies: Business Information Systems. Electives in Business Process Analytics and The Social Web. (2016-2017)
  - Thesis with Prof. dr. ir. H. Reijers: How to predict the effects of automation on jobs? Grade: 9,0.
  - Cum laude distinction: 8,5
- Liberal Arts and Sciences Bachelor Program, Amsterdam University College (AUC). Economics and Health tracks. Courses completed in Advanced Statistics, Programming, International Relations, and Literature. (2012-2016)
  - Capstone with Prof. dr. E. C. Perotti on the effects of overconfidence in banking management on the performance of banks during the financial crisis of 2007-08. Grade: 8,7
  - GPA: 3,58 (cum laude)
- Pre-university education, subjects: Dutch, English, German, Chemistry, Biology,

Mathematics (statistics), NLT (Nature, Life and Technology), History, and Geography. (2006-2012)
  - PWS (concluding project): Reward and Punishment in the brain. Grade: 9,7

A.3 Academic Achievements

A.3.1 Teaching

- Lecturer:
  - Information systems – co-design of course - tutorials + lectures – Bachelor (UU, 2019-2020)
- Online:
  - MOOC: A Step-by-Step Introduction to Process Mining) – Teaching support (Hasso Plattner Institute, OpenHPI, 2021)
- Tutorial lecturer:
  - Wetenschappelijke onderzoeksmethoden - Bachelor (UU, 2020-2021)
  - Advanced Research Methods – Master (UU, 2019-2020)
  - Business Process Management – Master (VU, 2018-2019)
- Supervision:
  - Master student thesis - three projects (UU, 2019-2021)
  - Bachelor student thesis - one project (VU, 2018)

A.3.2 Grants

- Seed fund round 16: Institutes for Open Societies. Hajo A. Reijers, Eva Knies, Thomas Martens, Iris Beerepoot, Jelmer J. Koorn (2021). *"Investigating Teams of Healthcare Professionals: The Relation Between Organisational Processes and Employee Well-Being"* Future of Work hub Utrecht University.

A.3.3 Projects

- Reconnaissance project for future PhD project at the Dutch Ministry of Infrastructure and Waterworks (2021-2022)
- Validation Program for Science-Based Startups, 9-month program from Utrecht Inc. - the ARE miner (2021)

A.3.4 Conference Presentations

- Hawaiian International Conference on System Sciences – Research paper: "Uncovering Complex Relations in Patient Pathways based on Statistics: the Impact of Clinical Actions" – Maui, Hawaii (Online, 2022)
- International Conference on Process Mining (ICPM) – Research paper: "Bringing Rigor to the Qualitative Evaluation of Process Mining Findings: An Analysis and a Proposal" – Eindhoven, the Netherlands (2021)
- International Conference on Business Process Management (BPM) – Research paper: "Looking for Meaning: Discovering Action-Response-Effect Patterns in Business Processes" – Seville, Spain (Online, 2020)

- Congres Evaluatie Passend Onderwijs (Dutch) – Two workshops: Stelsel & Sturing and Pers & Politiek (Online, 2020)
- International Conference on Conceptual Modeling (ER) – Workshop paper: "Towards Understanding Aggressive Behavior in Residential Care Facilities Using Process Mining" – Salvador, Brazil (2019)
- International Conference on Information Systems (ICIS) – Research paper: "Working Around Health Information Systems: The Role of Power", Workshop paper: "A fresh look at the impact of technology on occupations through software." - Munich, Germany (2019)
- International Conference on Information Systems (ICIS) – Workshop paper: "A Task Framework for Predicting the Effects of Automation"– San Francisco, USA (2018)
- Academy of Management (ACM) – Workshop paper: "Proposing a Task Framework for Predicting the Effects of Automation on Jobs"- Chicago, USA (2018)
- European Conference on Information Systems (ECIS) – Research paper "A task framework for predicting the effects of automation" - Portsmouth, UK (2018)

A.3.5 Professional Training

- Selling your Science workshop, from Utrecht Inc. (2021)
- Workshop van aula naar binnenhof, from Samenweten (2021)
- Research Data Management course, from the UU library (2020)
- Academic writing course, 3-month program from Artesc (2020)
- Research Methods and Methodology for IKS, from Graduate School SIKS (2019)
- Foundations of Data Science: Data and Process Mining, from Graduate School SIKS (2019)
- Introduction to Multilevel Analysis, Summer School UU (2019)
- Scientific Integrity Course, VU (2019)
- Data Analysis in R, Winter School VU (2019)
- Process Mining Training, Fluxicon (2019)
- Process Mining Camp, from Fluxicon (2019)

A.4 Professional Development

A.4.1 Board Memberships

- Member of the assessment committee on research in computer science for the period 2015-2020 at the Technical University Delft (2021)
- Chair PROUT (PhD network Utrecht) (2020-2022)
- Representative PROUT in PNN (Promovendi Netwerk Nederland) (2020-2022)
- Member of the audit committee PNN (2020-2021)
- Treasurer and co-chair PROUT (2019-2020)
- Board member PROUT (2019 & 2022)
- Member of the ODC (representative advisory board for employees) at the Vrije Universiteit Amsterdam (2017-2019)

- General board member master committee under student association VIA (Information Sciences faculty, UvA) (2016-2017)
- Treasurer of foundation VER, organizing study trips for student association VIA (2016-2017)
- Member of the audit committee at AUC (2015)
- Treasurer of SOLACE, social event committee of AUC (2014-2015)
- Mentor introduction week AUC (2014)
- General board member of GIMP, student committee in high school – advocating a balanced diet and regular exercises. (2006-2012)

A.4.2 Experience Abroad

- Visiting researcher Stevens Institute of Technology - USA, New Jersey. (2019)
- Volunteer program SEED Madagascar, six week program combining: teaching English to children, building a primary school, conservation of flora and fauna, analyses for biological research. (2016)

SIKS Dissertation Series

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VU), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UVA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UVA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UVA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (UvT), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design

41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

---

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UVA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TUE), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UVA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VU), Logics for causal inference under uncertainty
23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VU), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
30 Wilma Latuny (UvT), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VU), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement

46 Jan Schneider (OU), Sensor-based Learning Support
47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
02 Felix Mannhardt (TUE), Multi-perspective Process Mining
03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
05 Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process
06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
11 Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks
12 Xixi Lu (TUE), Using behavioral context in process mining
13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14 Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters
15 Naser Davarzani (UM), Biomarker discovery in heart failure
16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17 Jianpeng Zhang (TUE), On Graph Sample Clustering
18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
20 Manxia Liu (RUN), Time and Bayesian Networks
21 Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games
22 Eric Fernandes de Mello Araujo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27 Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis
28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29 Yu Gu (UVT), Emotion Recognition from Mandarin Speech
30 Wouter Beek, The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

---

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
03 Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
05 Sebastiaan van Zelst (TUE), Process Mining with Streaming Data
06 Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
09 Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems
10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12 Jacqueline Heinerman (VU), Better Together
13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25 Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses

28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33 Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34 Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
35 Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming
36 Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills
37 Jian Fang (TUD), Database Acceleration on FPGAs
38 Akos Kadar (OUN), Learning visually grounded and multilingual representations

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
04 Maarten van Gompel (RUN), Context as Linguistic Bridges
05 Yulong Pei (TUE), On local and global structure mining
06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
07 Wim van der Vegt (OUN), Towards a software architecture for reusable game components
08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
10 Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining
11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
12 Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15 Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17 Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20 Albert Hankel (VU), Embedding Green ICT Maturity in Organisations
21 Karine da Silva Miras de Araujo (VU), Where is the robot?: Life as it could be
22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
24 Lenin da Nobrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
25 Xin Du (TUE), The Uncertainty in Exceptional Model Mining
26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
30 Bob Zadok Blok (UL), Creatief, Creatieve, Creatiefst
31 Gongjin Lan (VU), Learning better – From Baby to Better
32 Jason Rhuggenaath (TUE), Revenue management in online markets: pricing and online advertising
33 Rick Gilsing (TUE), Supporting service-dominant business model evaluation in the context of business model innovation
34 Anna Bon (MU), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development

2021 01 Francisco Xavier Dos Santos Fonseca (TUD), Location-based Games for Social Interaction in Public Space
02 Rijk Mercuur (TUD), Simulating Human Routines:Integrating Social Practice Theory in Agent-Based Models
03 Seyyed Hadi Hashemi (UVA), Modeling Users Interacting with Smart Devices
04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
07 Armel Lefebvre (UU), Research data management for open science
08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play

10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
16 Esam A. H. Ghaleb (UM), Bimodel Emotion Recognition from Audio-Visual Cues
17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
19 Roberto Verdecchia (VU), Architectural Technical Debt: Identification and Management
20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
22 Sihang Qiu (TUD), Conversational Crowdsourcing
23 Hugo Manuel Proença (LIACS), Robust rules for prediction and description
24 Kaijie Zhu (TUE), On Efficient Temporal Subgraph Query Processing
25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
26 Benno Kruit (CWI & VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2022 1 Judith van Stegeren (UT), Flavor text generation for role-playing video games
2 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
3 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
4 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
5 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
6 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
7 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
8 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
9 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
14 Michiel Overeem (UU), Evolution of Low-Code Platforms
15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining