Mining Statistical Relations for Better Decision Making in Healthcare Processes

Jelmer J. Koorn
Utrecht University
Princetonplein 5, Utrecht
The Netherlands
j.j.koorn@uu.nl

Xixi Lu
Utrecht University
Princetonplein 5, Utrecht
The Netherlands
x.lu@uu.nl

Henrik Leopold
Kühne Logistics University
Großer Grasbrook 17, 20457 Hamburg
Germany
henrik.leopold@the-klu.org

Niels Martin
Hasselt University - Research Foundation Flanders (FWO)
Martelarenlaan 42, Hasselt - Egmontstraat 5, Brussels
Belgium
niels.martin@uhasselt.be

Sam Verboven
Vrije Universiteit Brussel
Pleinlaan 2 1050 Elsene
Belgium
sam.verboven@vub.be

Hajo A. Reijers
Utrecht University
Princetonplein 5, Utrecht
The Netherlands
h.a.reijers@uu.nl

***Abstract*—An important part of healthcare decision making is to understand how certain actions relate to desired and undesired outcomes. One key challenge is to deal with confounding variables, i.e., variables that influence the relation between actions and outcomes. Existing techniques aim to uncover the underlying statistical relations between actions and outcomes, but either do not account for confounding variables or only consider the process or case level instead of the event level. Therefore, this paper proposes a novel relation mining approach for healthcare processes that 1) explicitly accounts for confounding variables at the event level, and 2) transparently communicates the effect of the confounding variables to the user. We demonstrate the applicability and importance of our approach using two evaluation experiments. We use a real-world healthcare dataset to show that the identified relations indeed provide important input for decision making in healthcare processes. We use a synthetic dataset to illustrate the importance of our approach in the general setting of causal model estimation.**

***Index Terms*—process mining, statistical relations, confounding variables, healthcare**

I. Introduction

In many professional settings, taking the right actions often determines whether a desired outcome can be obtained or not. In a healthcare context, such desired outcomes include improving a patient's symptoms, 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 outcomes [1], [2]. The data that is collected by modern Health Information Systems (HISs) provides a valuable basis to study and understand this link. Among others, HISs record which actions have been taken, when these actions have been taken, and who was involved [3], [4]. Nonetheless, the relation between actions and outcomes is inherently complex and may depend on a large number of contextual factors [1].

In process mining literature, various techniques have been proposed that aim to discover such relations. They can be subdivided into three main groups: machine-learning-based approaches [5]–[7], statistics-based approaches [8]–[10], and context-based approaches [11], [12]. However, none of the existing techniques can deal with confounding variables at the event level. While techniques ignoring confounding variables can generate misleading or plainly wrong insights, techniques focusing on the process or case level can only predict or explain outcomes at those levels. Particularly in a healthcare context, the consideration of the event level is highly important since it allows to understand the outcome of individual actions, such as treatments.

Against the background of this research gap, we use this paper to propose a novel statistical relation mining technique for healthcare processes that 1) explicitly accounts for confounding variables at the event level and 2) transparently communicates the effect of the confounding variables to the user. We consider a scenario where a decision maker needs to choose from a set of responses given a particular action and we aim to understand whether these responses can be related to future actions. To develop and illustrate our conceptual ideas, we build on the problem of *aggressive behavior in residential care facilities*. In such a setting, we want to understand whether a certain response taken by a caretaker (e.g., isolating a client) to a patient's action (e.g., aggressive behavior towards people) is linked to a patient's actions in the future (e.g., no further aggressive behavior). At a technical level, we combine process mining techniques with statistical testing methods. As a result, our technique can identify complex hidden relations. By doing so, we pave the way for better understanding the complex relations in processes and improving decision-making in a data-driven way.

The rest of the paper is organized as follows. Section II elaborates on the problem and related work. Section III introduces the formal preliminaries. Section VI presents our approach for relation mining. Section V discuss the evaluation of our approach before Section VI concludes the paper.

II. Background

This section introduces the background of our research. Section II-A elaborates on the problem. Section II-B then discusses related work and highlights the research gap.

A. Problem

In many processes, it is desirable to understand the impact of executing certain actions, especially when there are several alternative options available. For example, 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, prevent instances of aggressive behavior [13]. If aggressive behavior occurs, care staff have a number of different options to respond to such an incident. Potential responses range from mild measures, such as warning the patient, to severe measures, such as secluding the patient. Whether the chosen response has been 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, the severity of the incident, or also other contextual factors.

To illustrate this, consider the scenario depicted in Figure 1. In the scenario at hand, we encountered an incident of *verbal aggression* and would like to understand to what extent *distracting* the client is a response that leads to deescalation, i.e., *no aggression*. Figure 1a shows a possible outcome for 100 cases if we do not consider any confounding variables. As indicated by the weights of the outgoing arcs, the results are highly inconclusive. It seems that distracting the client has no meaningful effect since all possible outcomes have the same probability (i.e., 25%). Figure 1b and Figure 1c, however, show that this is conclusion is incorrect. By accounting for the confounding variable *severity*, we realize that distracting the client can indeed lead to fewer cases of aggression, but only if the severity score is higher than 5 (see Figure 1c).

This example illustrates how complex understanding the impact of actions and responses can be in real-life situations. What is more, it shows that the consideration of confounding variables can be essential. Not accounting for confounding variables can lead to a flawed understanding of the underlying process and, therefore, to poor decision making [14]. In the next subsection, we briefly review existing literature to show that, currently, there is no approach available that addresses this problem.

B. Related Work

Understanding cause and effect relationships is an important ingredient for effective business process improvement [15]. Recognizing this, many researchers have developed approaches for analyzing and detecting such relationships based on process execution data. As a result, a variety of approaches exist that differ with respect to several dimensions. In general, we can subdivide existing approaches into three main categories based on the overall strategy they pursue to identify causal relationships: 1) machine-learning-based approaches, 2) statistics-based approaches, and 3) context-based approaches.

Machine-learning-based approaches build on traditional supervised learning techniques to both identify cause-effect relationships, and estimate their effects. The first approaches from this category mainly focus on understanding and explaining case-related phenomena [5], [15], [16]. What these approaches have in common is that they only identify *potential* causal relations. More recent approaches explicitly account for causation, for example, by combining action-rule mining with uplift trees [6] or using neural networks [7]. Both Verboven and Martin [7] and Bozorgi et al. [6] account for confounding as long as confounders are included in the data.

Statistics-based approaches build on established statistical tests and models to identify cause-effect relationships. Narendra et al. [9] propose an approach which uses structural causal models to encode and confirm existing assumptions about cause-effect relationships at the process level. The approach from Koorn et al. [10] builds on chi-square tests to identify relevant cause-effect relationships at the event level. However, it does not consider confounding variables. Qafari and van der Aalst [8] propose an approach based on structural equation models to, for instance, detect causal relationships between resources and process delays.

Context-based approaches build on techniques exploiting specific context dimensions, such as time or proximity, to identify cause-effect relationships. For example, Hompes et al. [12] use time series analysis to identify causal relations between business process characteristics and process performance indicators. Van Houdt et al. [17] leverage the time dimension by employing probabilistic temporal logic. Polyvyanyy et al. [11] present an approach that builds on the notion of proximity of events in terms of time, space, and semantics. In essence, they systematically develop domain-specific heuristics that allow them to identify relevant cause-effect relationships.

The review above illustrates that there are various approaches available for detecting cause-effect relationships based on process execution data. However, only a few existing approaches account for the problem of confounding variables, [6], [9], [17]. What is currently still missing is an approach that considers confounding variables at the event level. Such an approach would be able to make specific and valuable recommendations on what action to perform. We therefore use this paper to fill this gap.

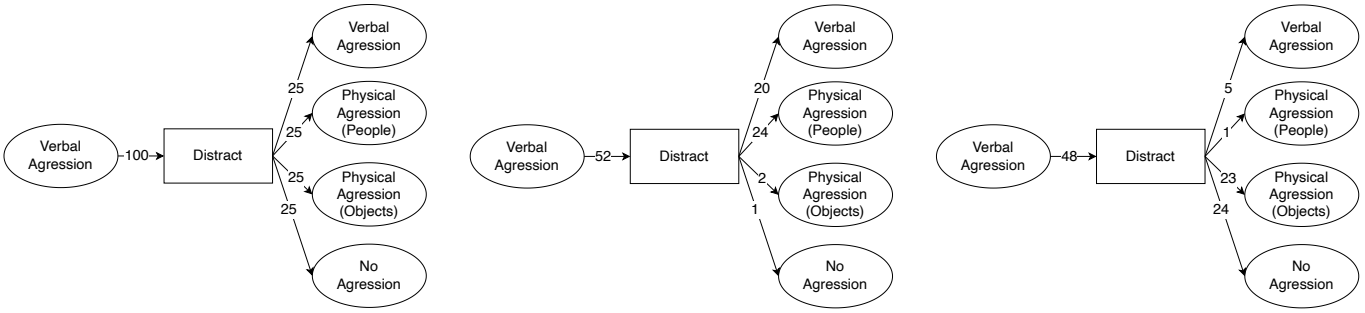
III. Preliminaries

In this section, we discuss the preliminaries of our work. Section III-A introduces the notion of action-response logs. Section III-B then explains how we can mine causal patterns from such logs.

A. Action-response logs

The *event logs* used by traditional process mining techniques are a collection of sequences of events. Each event records information such as the corresponding case, the task, the time of the execution, and the user (resource) who executed

Fig. 1: Impact of considering confounding variables



(a) No consideration (b) 'Severity' ≤ 5 (c) 'Severity' > 5

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

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

TABLE I: Excerpt of an action-response log

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 II: Excerpt of the tables for an individual response used to perform statistical tests; horizontal categories: effects on follow-up actions

the task. A sequence of events that correspond to a particular case is called a *trace*. In the context of this paper, we consider a specific type of event log: an *action-response log*. It differs from a traditional event log in the sense that such a log contains traces that alternate between actions and the responses towards these actions. We define an action-response log L as a specific type of log, where each event contains a case id (e.g., client id), an *action* (e.g., "Verbal aggression (VA)") and a *response* taken towards the action (e.g., "Distract Client").

To illustrate the notion of an action-response log, we use the example of aggressive behavior in residential care facilities. Table I illustrates action-response tuples, where each row records an occurred event. The column "Action" indicates the action of the event, and the column "Response(s)" lists the response(s) to the event. 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 log listed in Table I, for event e_1: $\pi_c(e_1) = 1$ is the case of event e_1, $\pi_a(e_1) =$ "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 action-response trace $\gamma(\sigma) = \langle (\pi_a(e_1), \pi_r(e_1)), ..., (\pi_a(e_n), \pi_r(e_n)) \rangle = \langle (a_1, r_1), ..., (a_n, r_n) \rangle$. An approach to convert a traditional event log into an action-response log has been described in Koorn et al. [10].

B. Mining causal patterns

Given an action-response log, it is possible to mine relevant causal patterns [10]. Such patterns reveal, given a particular action, which relations between responses and follow-up actions are statistically significant. As the existing notion of causal patterns introduced in Koorn et al. [10] does not take confounding variables into account, we set out to do so in this paper. Below, we explain how we can mine causal patterns from an action-response log using the chi-square-test.

Under the assumptions described in McHugh [18], the chi-square test is used to test the hypothesis whether a response has a significant effect on the follow-up actions, suggesting a potential *causal pattern*. To perform the test, we calculate the number of follow-up actions of different categories when there is a particular response and compare these numbers to when there is not such a response.

To test whether each response r_i has an influence on the follow-up action, we define $M_{a,r,A}$ as a $2\times |A|$ matrix:

$$M_{a,r,A} = \begin{pmatrix} f_{r,1} & f_{r,2} & \cdots & f_{r,n} \\ f_{\neg r,1} & f_{\neg r,2} & \cdots & f_{\neg r,n} \end{pmatrix} \quad (1)$$

where $f_{r,j} = |\{e_i \in L \mid \pi_a(e_i) = a \wedge \pi_r(e_i) = r \wedge \pi_a(e_{i+1}) = a_j\}|$ and $f_{\neg r,j}$ is the frequency distribution of effects of the responses other than r, i.e., $f_{\neg r,j} = |\{e \in L \mid \pi_a(e) = a \wedge r \notin \pi_r(e) \wedge \pi_a(e_{i+1}) = a_j\}|$. An example of $M_{a,r,A}$ where r is "Terminate contact" is listed in Table II. The chi-square

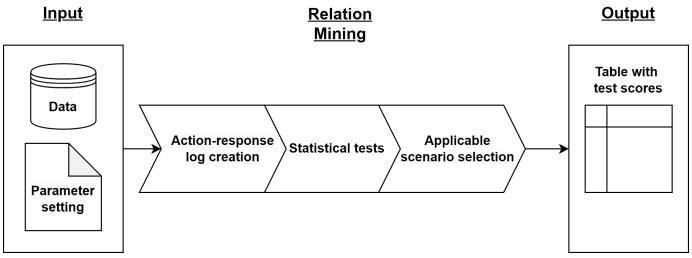


Fig. 2: A visual representation of the proposed technique

test calculates the expected frequencies and compares them to the observed frequencies M. If they differ significantly, then the null hypothesis is rejected, which means the responses have a statistically significant effect on the follow-up actions.

In the next section, we build on these concepts to define a relation mining approach that explicitly accounts for confounding variables.

IV. Approach

In this section, we present our approach for relation mining in healthcare processes. As shown in Figure 2, our technique consists of three main steps: 1) action-response log creation, 2) statistical tests, and 3) applicable scenario selection. As input, it expects data on potential confounding variables and two specific parameters. As output, it provides quantitative insights into the likelihood of confounding variables. Below, we elaborate on the input, the three main steps, and the output.

A. Input

The input for our technique consist of two main elements: 1) data on the candidate confounding variable and 2) parameters *k* and *M*. In general, the data that is used for the candidate confounding variable can be any attribute in the dataset. Note that an assumption of this technique is that all candidate confounding variables are included in the dataset. In Table I this is showcased by the last column *Severity*. The data can be of both categorical or continuous nature. In the proposed technique, we use two parameters to determine the strategy for detecting a possible confounding variable. First, we use *k* to denote the number of stratified datasets 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 the potential values are 'yes' or 'no'), we split the data into two datasets: one set only containing clients that have a history of aggressive behavior and another set only containing clients that were never aggressive before.

Second, we need a strategy to split the dataset 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 approach is recommended as 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. One option is 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.

B. Relation mining

The relation mining component is the core of our technique and consists of three specific steps: 1) action-response log creation, 2) statistical tests, and 3) applicable scenario selection.

Action-response log creation. As described above, we use an event log as starting point. We can convert an event log into an action-response log in two ways. If the response is a variable in the event log, we can use the conversion described in Section III. If the response variable is not present in the event log, the log can be converted to an action-response log following the approach presented in Koorn et al. [19].

Statistical tests. We perform statistical tests in two steps to identify confounding variables. The goal of each step is to: 1) determine an original test score for the action-response relation, and 2) test how the action-response relation holds when we consider a possible confounding variable. When looking for confounding variables, we first calculate a chi-square test score for the original dataset. The original dataset is the action-response log introduced in Section III. 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. If the test score is significant, this means there is at least one significant combination of action type and response type. As such, we can conclude that there is a relation between the action and response variable. If the test score is not significant, this means that there is no relation between the action and response variable. The exact mathematical approach of the chi-square in the action-response context is described in Section III.

Then, we introduce the possible confounding variable as an extra variable, which we will refer to as the *candidate variable* from here on. The candidate variable is used to stratify the original data into *k* smaller datasets called *subsets*. Following this, we perform the same test as the original 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 not significant, again indicating whether or not there is a relation between the action and the response variable.

Applicable scenario selection. At this point, there are three possible scenarios that can occur: 1) the original test score and subset test scores have the same results (indicating the absence of confounding variables), 2) the original test has the opposite results compared to each individual subset test scores (indicating the presence of a confounding variable), or 3) the original test score has the same result as at least one

Original data	Subsets	Confounding variable
Significant	All significant	No confounding variable
	Mixed results	Mediator variable
	All not significant	Confounding variable
Not significant	All significant	Confounding variable
	Mixed results	Mediator variable
	All not significant	No confounding variable

TABLE III: Scenarios for test results for confounding variable. The (not) significant values refer to the test scores

subset and opposes at least one subset (indicating the presence of a mediator variable). In Table III we present the different scenarios. Below, we outline the conceptual rationale for each scenario. In Section V, the scenarios are further exemplified using a real- world case study on aggressive behavior. As the technique is applicable beyond the setting of our case study, we use the statistically appropriate terms dependent, independent, and candidate variables below. In terms of the case of aggressive behavior, these terms respectively refer to response (independent), follow-up action (dependent), potential confounder (candidate).

Scenario 1: no confounding variable. The first scenario applies when all test results are either significant or not significant. 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 not significant, there is no relation between the independent and dependent variable, regardless of the candidate variable. In both situations, the candidate variable has *no effect* on the relationship between independent and dependent variable. Thus, we can conclude that the candidate variable is not a confounding variable.

Scenario 2: a confounding variable. The second scenario applies when the original test result is opposite (in terms of significance) to each of the subset test scores. In this scenario, there is a significant effect of the candidate variable on the relation between independent and dependent variable. When the original test score is significant and the subset test scores are not significant, the test results show that the hypothesized relation between independent and dependent variable disappears. In the opposite situation, where the original test score is not significant 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 variable is falsified. Thus, in both situations, we can conclude that the candidate variable is a *confounding variable* as the candidate variable presents an *alternative* explanation to the independent variable for explaining the dependent variable.

Scenario 3: a mediator variable. The third scenario applies when mixed results are obtained 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 score. However, the subset test scores show that this relation only holds under certain circumstances, captured in the candidate variable. In other words, the candidate variable mediates the relation between independent and dependent variable. 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 not significant, we hypothesize that there is no relation between the independent and dependent variables. However, the subset test scores show that under certain circumstances, captured in the candidate variable, the relation does exist. 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 mixed results are obtained 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*.

C. Output

The *output* of our technique is set of causal relations. When the technique detects a confounding variable, there are three scenarios, as described above: 1) no confounding variable, 2) confounding variable, and 3) mediator variable. Besides these scenario, it is also possible that there is not enough data. This occurs when, after data stratification, there is not enough data in each subset to perform the Chi-square tests. The technique produces test scores for the original set and each of the subsets per response type. For each test, it will indicate whether or not the score was significant or not significant. Based on this the appropriate scenario (1-4) is determined. To create a coherent overview, the output is presented in table format, see Table V as example.

V. Evaluation

This section presents a two-part evaluation of our approach. The first part (Section V-A) focuses on *applicability* and revisits the relationships discovered by Koorn et al. [10] in the context of a real-world case study. We show that considering potential confounding variables indeed leads to different results. The second part (Section V-B) uses a synthetic dataset to demonstrate the *importance* of our approach in the general setting of causal models. We show that our approach can help to ensure that no wrong assumptions about the existence of hidden confounders are made and how this impacts the results of causal models.

A. Real-world case

Dataset. For demonstration purposes, we use a real-world dataset of the care process at 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 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

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*
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 IV: Overview of the characteristics of the real-world dataset

response classes. We transformed this log into an action-response log by defining the next aggressive incident of a client as a follow-up action if it occurred within 9 days. Otherwise, the link is not considered and the follow-up action is defined as τ, i.e. none. This transformation procedure is in line with the approach followed by Koorn et al. [10]. As a result, we obtain a total of five different action classes. Table IV summarizes the characteristics of our dataset.

Candidate variable. To put the results produced by our approach into context, we compare our results to those from the ARE miner [10]. Since the ARE miner uses the same dataset as we do, we can revisit the relationships found and control for the effects of candidate variables on them. 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, as stated before, the ARE miner does not account for candidate variables that can influence the discovered relations. Based on insights from the healthcare organisation 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 well-being of both client and caretakers [20]. Research has also shown that support staff who experience high levels of stress due to aggressive behaviour of their clients are likely to respond in different ways [20]. Against this background, we hypothesize that the severity of an incident is a confounding variable for the relation between response of caretaker and future aggression of clients.

Results. The first step of the approach 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, parameter k equals the default value (k = 2). One subset is created for mild incidents, i.e., incidents with a score between 1 and 7. Another subset captures severe incidents, i.e., incidents with a score ≥ 7.

Recall that the relation miner produces results for action-response combinations. As such, a substantial amount of tables and graphical representations were produced when checking for the confounding variables. To illustrate the results, we focus on an exemplary case. Table V presents the results for the initial action self-injurious behavior where we check for the candidate variable severity of the incident. Below, we elaborate on the findings and show how each scenario described in Section IV-B is reflected in the case study results.

No confounding variable. In Table V, we see that a number of responses fall into the category of no confounding variable. For example, "Terminate contact" as a response has a significant effect on the future aggression after a self-injurious behavior incident.

When the data is then stratified and tested again, the chi-square results of the subset tests match the ones from the original test (see Table V). Thus, the pattern of response and follow-up action observed in the original data *remains the same* for each of the subsets. From the results in Koorn et al. [10], we know that this means that both for mild and severe incidents the response "Terminate contact" increases the chances of verbal aggression as future aggression. In addition, terminating contact reduces the chances of future self-injurious behavior. In other words, the severity score is *not* a confounding variable in this context.

A confounding variable. A confounding variable is identified when all the subsets return the opposite test scores compared to the test scores from the original dataset. Table V shows that the response "Send to other room" fits this description. In the original dataset, a significant relation is observed between this response and the follow-up actions. Koorn et al. [10] also found that the response "Send to other room" reduces the chance of a future repetition of the self-injurious behavior.

When the data is stratified and checked for the impact of sending a client to their room, we can see that there is no effect of this response on the future aggressive behavior of the client. This means that the original finding of sending a client to their 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 this relation should be disregarded from the original findings.

A mediator variable. A mediator variable is present when the tests results of the subsets are mixed. In our specific case, the response "Distract client" meets these criteria. The test score for the original data shows a significant relation. Koorn et al. [10] also found that distracting a client results in a higher chance of future violence against another person.

When stratifying the data based on the severity score, this relationship holds for mild incidents but not for severe incidents. The interpretation of the effect of distracting a client

Response type	Original set	Subset 1: severity <7	Subset 2: severity $\geq$ 7	Scenario
Terminate contact	Significant	Significant	Significant	1 (No confounding variable)
Send to other room	Significant	Insignificant	Insignificant	2 (Confounding variable)
Distract client	Significant	Significant	Insignificant	3 (Mediator variable)
Talk to client	Significant	Significant	Significant	1 (No confounding variable)
Seclusion	Insignificant	Not enough data	Not enough data	- (Not enough data)
Hold with force	Significant	Significant	Significant	1 (No confounding variable)
No measure	Significant	Significant	Insignificant	3 (Mediator variable)

TABLE V: 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)

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. We can therefore conclude that the severity score is a *mediator variable* as it influences the initially found relation between distracting a client and future aggression.

B. Synthetic experiment

Conditional Average Treatment Effect (CATE) estimation [21] has been the main driving force behind recent developments in personalized medicine, marketing, and policy research, based on observational data. In this section, we demonstrate that the approach proposed in this paper can help to ensure that the vital assumption of *No Hidden Confounders* in the context of CATE estimation is met.

Setting. CATE estimation involves the estimation of the causal effect of a response $W \in \{0, 1\}$ on a follow-up action $Y \in \mathbb{R}$ for an individual i characterized by features $X \in \mathcal{X} \subset \mathbb{R}^n$. As per the Rubin/Neyman Potential Outcomes framework [22], we assume that (in a binary setting) two potential follow-up actions (PFA), i.e., outcomes, exist for each individual, Y_0 and Y_1, associated with $W = 0$ and $W = 1$, respectively. For example, having an individual with "Verbal Agression", giving a response W being whether "Terminate contact" or not, we have the potential followup action $Y = 1$ if there is physcial aggression against people, or $Y = 0$ not. The CATE is then the expected difference between the PFAs of an individual, or:

$$\tau(x) := \mathbb{E}[Y_1 - Y_0 \mid X = x] = \mu_1(x) - \mu_0(x) \qquad (2)$$

where the expected potential follow-up action is denoted by $\mu_w(x) = \mathbb{E}[Y_w \mid x]$. Of these PFAs, we only ever observe the factual, i.e., $Y_f = Y(W) = WY_1 + (1 - W)Y_0$. As such, the CATE itself is unobservable for any i. For maximum clarity, we will consistently refer to $\tau(\mathbf{x})$ as CATE in the remainder of this paper.

While recent developments in CATE estimation has enabled granular causal effect estimation based on observational data, adherence to standard assumptions regarding *Stable Unit Treatment Value* (SUTVA), *Overlap*, and in particular *No Hidden Confounders* is required [23], [24]. No Hidden Confounders implies that confounders are observed, and included in the data set so that $(Y_0, Y_1) \perp\!\!\!\perp W|X = x$. The proposed approach helps to avoid violations of the no hidden confounders assumption through variable selection, i.e., that confounding variables are removed from the dataset since they are not considered relevant for the prediction task at hand. In the following paragraphs, we demonstrate how confounder-agnostic variable selection can harm the performance of causal models.

In particular, the impact of confounders on the potential follow-up actions may be offset by their selection effects. As such, using standard approaches to variable selection based on performance on Y_f can lead data scientists to omit confounders. Such *manual hiding of confounders* then leads to biased CATE estimates caused by the violation of the no hidden confounders assumption. To illustrate this paradox, and the importance of confounder detection for causal variable selection, we will simulate two scenarios in a synthetic example.

Dataset. To simulate the two scenarios based on the original data, but still have access to the ground truth, we extract 20 pre-response variables from the original dataset, including the confounder "severity". Furthermore, we one-hot encode all categorical variables, and standardize the continuous variables.

We adopt the same data generating process as [7], [25] to generate the PFA functions. This means that the variables and response assignment are the same as in the original data set, so the original response propensities ($\mathbb{P}(W_i = 1 \mid X_i = x)$) are retained and selection bias persists. Additionally, in the resulting PFA model the variables x_s are introduced and highly correlated to both PFA y_0 and y_1, but are unrelated to the resulting CATE. In Scenario 1 the practioner includes x_s but omits the confounding variables. In Scenario 2 the practitioner omits x_s, which highly correlates with both Y_0 and Y_1, but includes the confounders. Scenario 1 represents standard variable assignment, Scenario 2 represents taking into account confounder identification using our approach.

Results. We use cfrnet, a powerful neural network-based CATE estimator [26]. The performance on Y_f is evaluated using standard MSE. For the CATE estimation performance we use the Precision in Estimation of Heterogeneous Effects (PEHE) measure [27]:

$$PEHE = \frac{1}{N}\sum_{i=1}^{N}(CATE_i - \hat{CATE}_i)^2. \qquad (3)$$

Note that this metric is not observable in real data, i.e., when only Y_f is known.

As expected, in Scenario 1 cfrnet achieves better performance (24.53% lower MSE) on the observable outcome (Y_f) than with the confounders, but performs worse at CATE estimation than in Scenario 2, with a 47.15% higher PEHE.

We conclude that a trade-off can exist in causal estimation between accurate prediction of causal effects and PFA outcomes through confounding effects. As such, identification of confounders is paramount for variable selection in causal models, ensuring no confounders in the data are omitted. Finally, optimizing performance on Y_f can guide CATE estimation decisions only insofar confounding is corrected for.

VI. Conclusion

In this paper, we proposed a novel relation mining technique for healthcare processes that 1) explicitly accounts for confounding variables at the event level and 2) transparently communicates the effect of the confounding variables to the user. Using an evaluation based on a real-world case and a synthetic dataset, we demonstrated both the feasibility as well as the importance of our approach. We showed that ignoring the severity of incidents can lead to misleading conclusions and, hence, wrong recommendations. We further demonstrated that our approach can help ensure the no hidden confounder assumption in causal model estimation and how it impacts the causal model results.

Our approach is subject to two major limitations. First, we require a certain distribution of the data to be able to perform the statistical tests we employ. Second, we assume independence between the studied (candidate) variables. Our experiments, for example, assumed that the severity of an incident does not influence the duration of an incident. Such assumptions need to be manually checked prior to performing the statistical tests using correlation matrices.

In future work, we plan to validate and apply our approach on other healthcare cases. In addition, we will expand on this work by looking into automating the check for interaction effects among the confounding variables.

Acknowledgment. This research was supported by the NWO TACTICS project (628.011.004) and Lunet Zorg in the Netherlands. We would also like to thank the experts from the Lunet Zorg for their valuable assistance and feedback.

References

[1] X. Deng, J. Khuntia, and K. Ghosh, "Psychological empowerment of patients with chronic diseases: the role of digital integration," in *Proceedings of the 34th International Conference on Information Systems (ICIS'13)*, 2013, pp. 1–15.

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

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

[4] W. M. P. van der Aalst, "Data science in action," in *Process Mining*. Springer, 2016, pp. 3–23.

[5] S. Suriadi, C. Ouyang, W. M. P. van der Aalst, and A. H. ter Hofstede, "Root cause analysis with enriched process logs," in *International Conference on Business Process Management*. Springer, 2012, pp. 174–186.

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

[7] S. Verboven and N. Martin, "Combining the clinical and operational perspectives in heterogeneous treatment effect inference in healthcare processes," in *International Conference on Process Mining*. Springer, 2022, pp. 327–339.

[8] M. S. Qafari and W. M. P. van der Aalst, "Root cause analysis in process mining using structural equation models," in *International Conference on Business Process Management*. Springer, 2020, pp. 155–167.

[9] T. Narendra, P. Agarwal, M. Gupta, and S. Dechu, "Counterfactual reasoning for process optimization using structural causal models," in *International Conference on Business Process Management*. Springer, 2019, pp. 91–106.

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

[11] A. Polyvyanyy, A. Pika, M. T. Wynn, and A. H. Ter Hofstede, "A systematic approach for discovering causal dependencies between observations and incidents in the health and safety domain," *Safety science*, vol. 118, pp. 345–354, 2019.

[12] B. F. Hompes, A. Maaradji, M. La Rosa, M. Dumas, J. C. Buijs, and W. M. P. van der Aalst, "Discovering causal factors explaining business process performance variation," in *International Conference on Advanced Information Systems Engineering*. Springer, 2017, pp. 177–192.

[13] B. P. Lloyd and C. H. Kennedy, "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, 2014.

[14] M. Nørgaard, V. Ehrenstein, and J. P. Vandenbroucke, "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, 2017.

[15] T. Lehto, M. Hinkka, and J. Hollmén, "Focusing business improvements using process mining based influence analysis," in *International Conference on Business Process Management*. Springer, 2016, pp. 177–192.

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

[17] G. Van Houdt, B. Depaire, and N. Martin, "Root cause analysis in process mining with probabilistic temporal logic," in *International Conference on Process Mining*. Springer, 2022, pp. 73–84.

[18] M. L. McHugh, "The chi-square test of independence," *Biochemia medica: Biochemia medica*, vol. 23, no. 2, pp. 143–149, 2013.

[19] J. J. Koorn, X. Lu, F. Mannhardt, H. Leopold, and H. A. Reijers, "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*, 2022.

[20] R. P. Hastings, "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, 2002.

[21] I. Bica, A. M. Alaa, C. Lambert, and M. Van Der Schaar, "From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges," *Clinical Pharmacology & Therapeutics*, vol. 109, no. 1, pp. 87–100, 2021.

[22] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *J Am Stat Assoc*, vol. 100, no. 469, pp. 322–331, 2005.

[23] ——, "Randomization analysis of experimental data: The fisher randomization test comment," *Journal of the American statistical association*, vol. 75, no. 371, pp. 591–593, 1980.

[24] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[25] A. M. Alaa and M. van der Schaar, "Bayesian inference of individualized treatment effects using multi-task gaussian processes," *arXiv preprint arXiv:1704.02801*, 2017.

[26] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *International Conference on Machine Learning*, 2017, pp. 3076–3085.

[27] S. Athey and G. Imbens, "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7353–7360, 2016.