On the role of prognostic factors and effect modifiers in structural causal models

Rianne M. Schouten*
Department of Mathematics & Computer Science, Data Mining Group
Eindhoven University of Technology, the Netherlands
r.m.schouten@tue.nl

Abstract

Causal effects vary within subgroups in the population. If causal effects are heterogeneous with respect to observed covariate information, it is possible to estimate these effects by conditioning on values of the covariates. We call these covariates *effect modifiers* and distinguish from *prognostic factors*, which influence the outcome but not the treatment effect. In this paper, we contribute to the understanding of structural causal relations by designing two controlled experiments. In both experiments, we temporarily disregard the fundamental problem of causal inference that factual and counterfactual outcomes cannot be observed together. We consider treatment assignment variable W as a time-variant variable where every possible treatment value is a measurement occasion and observe the outcome value for all possible treatments. This approach creates a nested, hierarchical data structure where we can study the relation between individual (lower-level) and average (higher-level) treatment effects. Specifically, we compare within-subjects variance for three hypothetical individual treatment effect distributions and demonstrate that a traditional two-arm trial without additional covariates assumes a worst-case scenario for underlying variance distributions. Second, we demonstrate how aggregation functions interfere with assumptions about the presence of prognostic factors and effect modifiers. Altogether, we believe that our findings provide valuable insights into the behavior of non-confounding covariates and contribute to a better understanding of structural causal relations.

1 Introduction

Learning causal representations without having access to ground-truth causal diagrams is non-trivial. Solving this problem requires, at the very least, a thorough understanding of how causal variables behave in a controlled setting. This paper contributes to such understanding by designing two controlled experiments. In both experiments, we temporarily disregard the fundamental problem of causal inference that factual and counterfactual outcomes cannot be observed together [8]. At the same time, we stay close to reality by making assumptions such as the Stable Unit Treatment Value Assumption (SUTVA) [34, 10], no hidden confounders and overlap (i.e., positivity) [3].

Causal effects vary within subgroups in the population. If causal effects are heterogeneous with respect to observed covariate information, it is possible to estimate these effects by conditioning on values of the covariates: we call them Conditional Average Treatment Effects (CATEs) [33]. In the medical domain, the same concept is referred to as Heterogeneity of Treatment Effect (HTE) and defined as "non-random variation in the magnitude or direction of a treatment effect across levels of a covariate against a clinical outcome" [14, p.35]. We follow [40] by referring to covariates that influence the treatment effect (but not the treatment assignment) as *effect modifiers*.

*https://rianneschouten.github.io

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

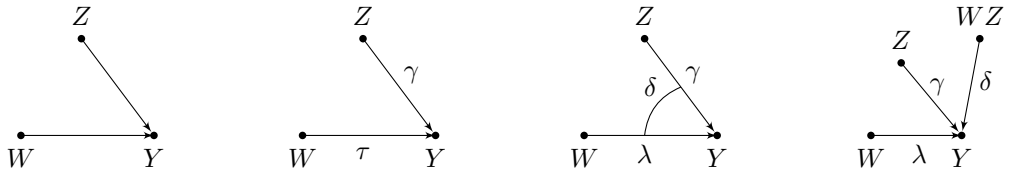


Figure 1: Causal diagrams with three variables: outcome Y, treatment assignment W and covariate Z (a) no structural restrictions; Z can be a prognostic factor, effect modifier or both (b-d) the causal relations are assumed to be linear (b) Z is a prognostic factor (c) Z is an effect modifier, diagram cf. [29] (d) Z is an effect modifier, a variation of the diagram cf. [22].

One can distinguish effect modifiers from another type of covariates: *prognostic factors*. These are variables that influence the outcome, but not the treatment effect [13] (nor the treatment assignment; effect modifiers and prognostic factors are not confounders). Both effect modifiers and prognostic factors are abundant in any observational dataset.

Even if we assume that we know the causal diagram, in the absence of restrictions on the structure of the causal relationships, we cannot distinguish effect modifiers from prognostic factors [28, 40] (Figure 1a).[2] Alternatively, in this paper, we assume that the causal relations are linearly structured. Then, the causal diagrams appear slightly different depending on whether Z is a prognostic factor (Figure 1b) or an effect modifier (Figure 1c cf. [29], and Figure 1d cf. [22]). Accordingly, the relation between treatment assignment variable W, covariate Z and outcome Y is modeled as

$$Y = \tau W + \gamma Z + \delta WZ, \tag{1}$$

where parameter τ represents the Average Treatment Effect (ATE) in the population, Z induces variation in the outcome variable Y through γ (prognostic effect) and δ quantifies the difference between CATE and τ for certain subgroups in the population (effect modification). If γ and δ are both significantly different from 0, covariate Z performs both roles simultaneously.

The aim of this paper is to give insights into the prognostic and effect modification behavior of covariate Z. We do this by considering W to be a time-variant variable where every possible treatment value is a measurement occasion. We then observe the outcome value for all possible treatments (i.e., for all possible time steps). This approach creates a nested, hierarchical data structure where variables are repeatedly measured per individual, which allows us to study the relation between individual (lower-level) and average (higher-level) treatment effects.

Specifically, we present results from two synthetic data experiments. In the first experiment, we evaluate within-subjects variance in outcome Y for three hypothetical Individual Treatment Effect (ITE) distributions. These distributions correspond to the various roles of covariate Z, which is time-invariant in this experiment. Second, we consider Z to be a time-varying covariate and demonstrate how aggregation functions interfere with assumptions about the presence of prognostic factors and effect modifiers. Altogether, we believe that our findings provide valuable insights into the behavior of non-confounding covariates and contribute to a better understanding of structural causal relations.

2 Background

2.1 Preliminaries

First, consider $D = \{\mathbf{x}_i\}_{i=1}^n$ to be a sample of n IID draws from $X = (W, Z, Y)$ with state space $\mathcal{X} = \mathcal{W} \times \mathcal{Z} \times \mathcal{Y}$ such that a sample is a tuple $\mathbf{x}_i = (w_i, z_i, y_i)$. Variable Y is a numerical outcome with $\mathcal{Y} \in \mathbb{R}$. The state spaces of treatment assignment variable W and covariate Z are provided in Sections 3.1 and 4.1, respectively. Note that in the traditional causal setting, every individual $i \in \{1, 2, ..., n\}$ is assigned to one out of k treatment groups. Often, we set $\mathcal{W} \in \{0, 1\}$ ($k = 2$); then $\{\mathbf{x}_i \in D \mid w_i = 1\}$ reflects the treatment group and $\{\mathbf{x}_i \in D \mid w_i = 0\}$ the control group.

[2] In Figure 1, we do not display the noise variables and assume they are jointly independent; the models satisfy the causal Markov condition [28].

Next, consider that for a individual i we draw repeated measurements from $\mathcal{X}$ (e.g., repeated blood tests or measuring heart rate over a period of time), denoted as a tuple $(\mathbf{x}_i^1, ..., \mathbf{x}_i^t, ..., \mathbf{x}_i^T)$. These measurements are non-IID if there exists some form of correlation between measurement $\mathbf{x}_i^t$ and $\mathbf{x}_i^{t+\lambda}$, or if the observed distribution $p(\mathbf{X}_i)$ differs from $p(\mathbf{X}_j)$ (for individuals $i, j \in \{1, 2, ..., n\}$). In the real world, such correlation structure naturally exists. Consequently, the resulting dataset has a hierarchical or multilevel structure where there exist multiple entity types where the measurements of the entities of one type (here, entity type *time* with T entities per individual) are nested in the entities of another type (here, entity type *individual* with n entities in the dataset) [9].[3]

In this paper, we create temporal structure by adopting a slightly unusual scenario where the treatment assignment value is set to be the time indicator: $w_i^t \leftarrow t$. Essentially, we create a hypothetical scenario where for all individuals, we observe both factual and counterfactual outcomes. In other words, we model the effect of treatment as time-varying effects.[4] This allows us to study the interaction between Z and W from a within-subjects perspective; an effect that takes place at a lower hierarchical level than the desired level of inference.

2.2 ANalysis Of VAriance

In Section 3, we use a within-subjects ANalysis Of VAriance (wANOVA) to investigate the variance distributions in outcome variable Y. An ANOVA is a statistical model that divides variance in an outcome variable over one or multiple components [39]. Depending on the data structure and research question, we distinguish a between-subjects ANOVA (bANOVA) from a wANOVA. A bANOVA divides the variance in the outcome variable between categories of a categorical variable G (an effect of the presence of groups) and a within-subjects variance that cannot be further explained and is therefore considered left-over or error variance. In a wANOVA, the grouping variable G is replaced with a time variable T. The variance in the outcome variable is then divided between an effect of time, an effect of systematic individual differences and a left-over variance. We provide more in-depth explanation, equations and a schematic illustration in Appendix A.

2.3 Local Pattern Mining

The concept of *global* models that explain most of the instances in the data is opposed with that of *local* models or patterns [7, 27]. Where global models tend to find the obvious patterns in the data, local patterns cover small parts of the data that deviate from the population distribution and display some internal structure; we call them *subgroups* [15, 38, 16, 17, 1, 5]. In Section 4, we use Local Pattern Mining (LPM) techniques to discover 1) subgroups of individuals with exceptionally high outcome values and 2) subgroups of individuals with an exceptional increase in outcome over time. Specifically, we use the beam search algorithm (a heuristic search algorithm, see Appendix B) and follow existing LPM literature in aggregating lower level measurements [12, 36] to discover subgroups of higher-level entities.

3 Experiment 1

In this experiment, we demonstrate how varying underlying ITE distributions affect the variance structure of an outcome variable Y. We demonstrate that averaging over individuals (as is done in reality, outside the hypothetical bubble of this experiment) is equivalent to assuming that none of the left-over variation can be explained by systematic differences between individuals.

3.1 Experimental setup

We generate synthetic data with a hierarchical structure as described in Section 2.1. We set $n = 100$, $T = 2$, and $w_i^t \leftarrow t$ for all $i \in \{1, 2, ..., n\}$ and $t \in \{0, 1\}$. We then sample outcome values

[3] The number of repeated measurements may vary between individuals. In addition, some variables may be measured only once per individual; others could be repeatedly sampled with varying counts and intervals. Depending on these data characteristics, the data may be formatted as a flat-table, a relational database or in any other format.

[4] We temporarily disregard the fundamental problem of causal inference [8], but still assume SUTVA [10]. We do not belief that time influences the treatment effect; we use time to model the treatment effect.

$y_i^t = \pi_{0i} + \pi_{1i} w_i^t + e_i^t$. Error $e_i^t \sim \mathcal{N}(0, \sigma_e^2)$ is normally distributed (and jointly independent). To be precise, we sample values for y_i^t from $\mathcal{N}(\mu^t, \sigma^t)$ with $\mu^0 = \pi_{0i} = 5$, $\mu^1 = \mu^0 + \pi_{1i} = 7.5$ and $\sigma^0 = \sigma^1 = 2$. Consequently, the ATE is fixed to 2.5.

Next, we re-order the values such that we create three possible ITE distributions as depicted in Figure 2. In Figure 2a, all patients have the same ITE and that ITE equals the ATE. Remark that there exists variation in Y, but that variation does not induce differences in treatment effects between individuals. In the second scenario, the ITEs all cross through the Grand Mean (GM) (Figure 2b). Here, an individual's ITE is as opposite to the ATE as possible. Third, for every individual we randomly sample one of the two outcome values as factual outcome, and we set the counterfactual to the group mean of the counterfactual group, denoted with $\overline{y}^0$ for the control group and $\overline{y}^1$ for the treatment group. Thus, if we randomly sample $y_i^{t=1}$ for individual i, then value $y_i^0 \leftarrow \overline{y}^0$, and similar for $t = 0$. The resulting ITE distribution is visualized in Figure 2c. Remark that in all three scenarios, the ATE, GM, and group means do not change; we only change the underlying ITE distributions.

We deliberately write the coefficients as π_{0i} and π_{1i} to indicate that the sampled values occur at the lowest hierarchical level (the time level); these effects can be further specified using second-level attributes, for instance a potential covariate Z. Then, $\pi_{0i} = \beta_{00} + \beta_{01} z_i + \mu_{0i}$ and $\pi_{1i} = \beta_{10} + \beta_{11} z_i + \mu_{1i}$ (error normally distributed and jointly independent). Integrating these higher level equations into the lowest level equation gives:

$$y_i^t = \beta_{00} + \beta_{01} z_i + \beta_{10} w_i^t + \beta_{11} w_i^t z_i + \mu_{1i} w_i^t + \mu_{0i} + e_i^t. \tag{2}$$

Consequently, scenario (a) equals the situation that $\beta_{01} \neq 0$ and $\beta_{11} = 0$: Z is a prognostic factor. In scenario (b), Z is an effect modifier and $\beta_{01} = 0$ and $\beta_{11} \neq 0$. In scenario (c), Z is an effect modifier with an additional prognostic component ($\beta_{01} \neq 0$ and $\beta_{11} \neq 0$).

3.2 Experimental results

Inspection with bANOVA and wANOVA gives results as presented in Table 3. Here, the left-most column shows the traditional bANOVA output and the three right columns the wANOVA results. We see that the total variance SS_{tot} in the outcome variable (1046) can be explained by an effect of group (219). Based on a statistical test, we would reject the null hypothesis that there is no effect of treatment with $F(1, 198) = 52.44, p < 0.001$.

Table 3 furthermore shows that in scenario (a), 814 of the 827 within-subjects variance can be explained by an effect of subject. This demonstrates the existence of systematic differences between subjects (individuals). In fact, in scenario (a), the only difference between subjects is given by the distance of their individual average outcome ($\overline{y}_i$) and the GM; no further variance is to be explained.

In contrast, in scenario (b), almost none of the within-subjects variance can be explained by a subject effect. Instead, a subject's average value does not say much about the outcome values, and there exists an interaction between treatment and subject that cannot be explained by observed information.

In scenario (c), the total variance decreases to 706 because 100 out of 200 outcome values are set to the group mean. These 100 group mean values all have a shorter distance to the GM than the original values and therefore, the total variance decreases. Yet, the ATE in scenario (c) is still the same, as

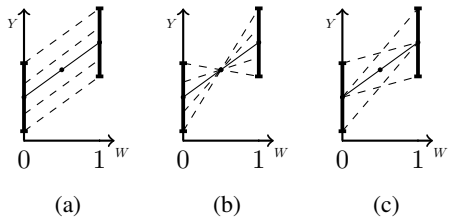


Figure 2: Visualizations of three types of ITE distributions in a two-arm experiment with Y as the outcome variable and W the treatment assignment indicator. (a) all ITEs equal the ATE (b) all ITEs cross through GM (c) counterfactuals equal the group mean.

	bANOVA	wANOVA		
		(a)	(b)	(c)
SS_{tot}	1046	1046	1046	706
SS_{group}	219	219	219	219
SS_{ind}	-	814	11	238
SS_{error}	827	13	816	238

Figure 3: Synthetic data results of Experiment 1 in Section 3. The table gives the Sum of Squares (SS) for a bANOVA and a wANOVA for three possible ITE distributions as visualized in Figure 2. More information on bANOVA and wANOVA can be found in Appendix A.

can be seen from the between-subjects variance of 219. Furthermore, the within-subjects variance is equally divided between an effect of subjects and an interaction effect. In other words, half of the variance can be explained by consistent differences between individuals, while the other half of the variance is non-systematic and therefore unexplained.

4 Experiment 2

In this experiment, we demonstrate that whether or not Z is a prognostic factor or effect modifier, high-quality individual-level representation of lower-level measurements can discover the variance distributions of these measurements. However, poorly chosen aggregation functions interfere with assumptions about the presence of prognostic factors and effect modifiers.

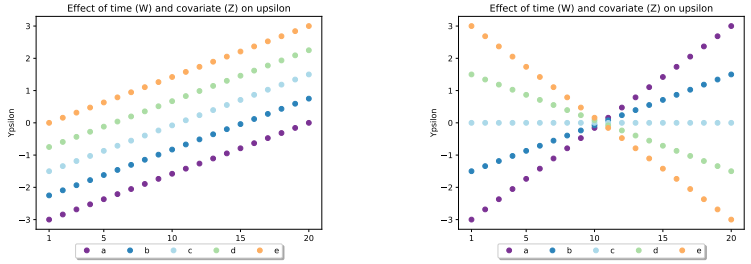
4.1 Experimental setup

We let covariate Z be time-varying, constructed as a random walk of $T = 20$ steps over nodes $h \in \mathcal{H} = \{a, b, c, d, e\}$, where $P(z_i^t = h \mid z_i^{t-1} = h') = 1/5$. In other words, per individual, covariate Z is modeled as an event-sequence of length T where the next event value is independent of the current event. We model the treatment assignment variable as a time-variant variable where every possible treatment value is a measurement occasion. Concretely, this means we set the treatment assignment value as the time indicator: $w_i^t \leftarrow t$.

Next, we generate $y_i^t = \mathcal{N}(10 + \upsilon_i^t, 0.1)$, where we let $\upsilon_i^t = \alpha w_i^t + \zeta z_i^t + \lambda w_i^t z_i^t$ be a linear function of the time indicator t (through w_i^t) and the integer position of covariate value z_i^t (i.e., $a = 1$, $b = 2$, etc; assumptions underlying linear regression are slightly violated, but do not influence the conclusions of our experiment). The aim is to discover exceptional subgroups of individuals using individual-level representations of the lower-level event sequences. We create various additional lower-level and individual-level noise variables. We use the beam search algorithm (see Appendix B) and construct $2 \times 2 \times 2$ scenarios: all combinations of $Ztype \in \{\text{prognostic}, \text{effect modification}\}$, $Aggfunc \in \{\text{imperfect}, \text{perfect}\}$ and $Eval \in \{\text{average}, \text{increase}\}$.

The simulation parameter $Ztype$ refers to whether variable Z acts as a prognostic factor or as an effect modifier. To let Z have a prognostic effect, $\alpha = \zeta = 1$ and $\lambda = 0$ (see visualization in Figure 4a). There are main effects of time and variable Z, but no interaction effect. In contrast, Figure 4b displays the scenario where Z is an effect modifier without main effects; $\alpha = \zeta = 0$ and $\lambda = 1$.

To represent lower-level measurements at the level of the individual, we compare two approaches for $Aggfunc$. In the *perfect* scenario, we know the ground truth values υ_i^t and average those using $f_{\text{sum}}(\upsilon_i^1, ..., \upsilon_i^T) = \sum_{t=1}^{T} \upsilon_i^t$ and $f_{\text{incr}}(\upsilon_i^1, ..., \upsilon_i^T) = \upsilon_i^T - \upsilon_i^1$. Alternatively, we construct an *imperfect* scenario that reflects handling event-sequences in practice: by counting the average frequency of each event type, and by determining the average index location. The idea is that high frequency values capture individuals with highly repetitive sequences (i.e., the random walk stays at the same node; the path through Figure 4 follows the same color) whereas the index location reflects



(a) (b)

Figure 4: Visualizations of the relation between the time indicator t (through w_i^t, x-axis), covariate values z (colors, from low (event type a, purple) to high (event type e, orange)) and υ. Every individual i walks a path through these dots from left to right. (a) Z acts as a prognostic factor (b) Z acts as an effect modifier, there are no main effects.

Table 1: Description of the most exceptional subgroup, discovered with φ_μ (exceptionally high average outcome) and φ_θ (exceptionally high increase in outcome). Aggregation of event-sequences to single values per individual is done with perfect, ground-truth knowledge and imperfect knowledge. Covariate Z acts as a prognostic factor or effect modifier.

		$Aggfunc$	
$Eval$	$Ztype$	imperfect	perfect
φ_μ	prognostic	$f_{\text{freq}_a} \leq 3 \wedge f_{\text{freq}_e} \geq 4 \wedge f_{\text{freq}_b} \leq 5$	$f_{\text{sum}} \in [5, 22]$
	effect modification	$f_{\text{idx}_a} \geq 11 \wedge f_{\text{idx}_e} \leq 10$	$f_{\text{sum}} \in [-134, -25]$
φ_θ	prognostic	$f_{\text{idx}_e} \geq 10.5 \wedge f_{\text{idx}_a} \leq 9.5$	$f_{\text{incr}} = 3$
	effect modification	$f_{\text{freq}_e} \leq 3 \wedge f_{\text{freq}_a} \geq 4 \wedge f_{\text{idx}_e} \leq 14$	$f_{\text{incr}} \in [-38, -29]$

where in the random walk, on average, a certain node is visited (e.g., more in the start or at the end of the sequence). The resulting 10 features are denoted as f_{freq_h} and f_{idx_h} for $h \in \mathcal{H}$.

Third, we evaluate the exceptionality of individuals in two ways. If $Eval = average$, discover subgroups of individuals with exceptionally high, average outcome values. We quantify exceptionality with quality measure $\varphi_\mu = (\mu^{SG} - \mu^D)/\text{se}(\mu^{SG})$. Here, $\mu^{SG} = 1/n^{SG} 1/T \sum_{i \in SG} \sum_{t=1} y_i^t$ is the average outcome value of the individuals covered by the subgroup, $se(\mu^{SG})$ is its standard error, and μ^D is the average outcome in the entire dataset D. In contrast, if $Eval = increase$, we discover subgroups of individuals with an exceptional increase in outcome between $t = 1$ and $t = T$. Here, $\varphi_\theta = (\theta^{SG} - \theta^\Omega)/\text{se}(\theta^{SG})$ with $\theta^{SG} = 1/n^{SG} \sum_{i \in SG} (y_i^T - y_i^1)$.

4.2 Experimental results

For all 8 simulation scenarios, Table 1 presents the description of the most exceptional subgroup. First, for the scenario where $Eval = \varphi_\mu$, $Ztype =$ prognostic, and $Aggfunc =$ imperfect, the subgroups covers individuals with low frequency of event type a, low frequency of event type b and high frequency of event type e. These are individuals with event sequences that stay close to the orange path in Figure 4a. They have the highest υ values and therefore, the highest outcome values.

In contrast, when $Eval = \varphi_{theta}$, individuals with an exceptional increase in outcome start their sequence with event type a ($f_{\text{idx}_a} \leq 9.5$) and end their sequence with event type e ($f_{\text{idx}_e} \geq 10.5$) (see Table 1, third row). In other words, they start at purple and cross through to end at orange (Figure 4a).

The reversed line of reasoning holds if $Ztype =$ effect modification, as depicted in Figure 4b. Then, discovering individuals with high outcome values (φ_μ) requires crossing through colors and hence, the usage of f_{idx} (second row in Table 1, whereas individuals with a high increase (φ_θ) need to stay close to the purple color (event type a) as long as possible ($f_{\text{freq}_a} \geq 4$).

Interestingly, if we have direct access to υ, we can construct an aggregation function f_{sum} that works for discovering individuals with high outcome values (φ_μ) independent of whether Z is a prognostic factor or effect modifier (right column in Table 1, f_{sum} is used in both rows corresponding to φ_μ). Similarly, we can construct f_{incr} in combination with (φ_θ), independent of the covariate role of Z.

5 Discussion and Conclusion

The aim of this paper is to give insights into the prognostic and effect modification behavior of covariate Z. We do this by designing two controlled experiments where we temporarily disregard the fundamental problem of causal inference that factual and counterfactual outcomes cannot be observed together. We consider linear relations between a treatment assignment variable W, a covariate Z and an outcome variable Y. Treatment assignment variable W is modeled as a time-variant variable where every possible treatment value is a measurement occasion. As such, we create a nested, hierarchical data structure which allows us to study the relation between individual (lower-level) and average (higher-level) treatment effects.

Our findings from Experiment 1 in Section 3 demonstrate that variance distributions in a traditional two-arm experiment as measured by bANOVA (most-left column in Table 3) are the same as those in a wANOVA with covariate Z being an effect modifier (scenario (b) in Figure 2 and Table 3). This

means that a bANOVA without additional covariates assumes a worst-case scenario for underlying ITE distributions. Indeed, in the real world, including covariates Z to control for prognostic and effect modification behavior reduces left-over variance and improves precision in estimating τ [10, 39]: we would move from scenario (b) to scenario (c) to scenario (a) in Table 3.

Remark that in the real-world, it is non-trivial to distinguish prognostic factors from effect modifiers. Consequently, is can be difficult to know how to include Z in the model. For instance, when assuming linear causal relations, Z should be included as a main effect to control for confounding effects and to control for prognostic effects, and Z should additionally be included in an interaction term to control for effect modification behavior. With many variables, the number of model parameters inflates quickly. In the domain of uplift modeling, the problem of causal inference is circumvented by fitting multiple models: one for each treatment group. An individual's ITE is then estimated by comparing the factual outcome with the predicted counterfactual [30, 31]. In fact, the scenario in Figure 2c represents a baseline double-model where the simplest counterfactual model is the group average ($\overline{y}^k$). More advanced methods directly estimate the net difference between two treatment groups [11, 35].

Our findings from Experiment 2 in Section 4 demonstrate that the quality of individual-level representations of lower-level measurements determines whether or not our assumptions about the nature of a time-varying covariate Z will influence higher-level inference making. With poorly chosen aggregation functions, assumptions regarding the effect of Z on ITEs influences results. For instance, in our experiment, aggregation functions based on frequency perform well when Z is a prognostic factor and the average effect is based on an estimate of the mean, or when Z is an effect modifier and the average effect is based on an estimate of the slope (first and last row in left column in Table 1). Aggregation functions based on location work well with reversed relations between Z and the average treatment effect. However, if the aggregation functions are close to the ground truth, the nature of Z does not matter for whether or not average treatment effects can be unraveled. For instance, aggregation function f_{sum} works well for discovering average effects based on estimates of the mean, whether or not Z is a prognostic factor or effect modifier (first two rows in right column in Table 1). Overall, the better the quality of individual-level representations, the less our results will interfere with assumptions about the role of covariate Z. We then do not need to rely on the reliability of those assumptions.

Acknowledgments and Disclosure of Funding

This work is part of the research program Data2People with project Exceptional and Deep Intelligent Coach (EDIC) and partly financed by the Dutch Research Council (NWO). We gratefully acknowledge the continuous support of Professor Mykola Pechenizkiy and dr. Wouter Duivesteijn. We thank Professor Cassio de Campos and dr. Devendra Singh Dhami for providing valuable feedback to earlier drafts of this paper.

References

[1] Martin Atzmueller, Juergen Mueller, and Martin Becker. Exploratory subgroup analytics on ubiquitous data. In *International Workshop on Mining Ubiquitous and Social Environments*, pages 1–20, 2013.

[2] Adnene Belfodil, Wouter Duivesteijn, Marc Plantevit, Sylvie Cazalens, and Philippe Lamarre. DEvIANT: Discovering significant exceptional (dis-) agreement within groups. In *Proc. ECML PKDD*, pages 3–20, 2020.

[3] Ioana Bica, Ahmed M. Alaa, Craig Lambert, and Mihaela 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*, 109(1): 87–100, 2021.

[4] Guillaume Bosc, Jean-François Boulicaut, Chedy Raïssi, and Mehdi Kaytoue. Anytime discovery of a diverse set of patterns with Monte Carlo Tree Search. *Data Mining and Knowledge Discovery*, 32(3):604–650, 2018.

[5] Wouter Duivesteijn, Ad J. Feelders, and Arno Knobbe. Exceptional Model Mining. *Data Mining and Knowledge Discovery*, 30(1):47–98, 2016.

[6] Dragan Gamberger and Nada Lavrac. Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research*, 17:501–527, 2002.

[7] David J Hand. Pattern detection and discovery. In *Proc. Pattern Detection and Discovery: ESF Exploratory Workshop London*, pages 1–12, 2002.

[8] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[9] Joop Hox, Mirjam Moerbeek, and Rens van de Schoot. *Multilevel analysis: Techniques and applications*. Routledge, 2017.

[10] Guido W. Imbens and Donald B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[11] Maciej Jaskowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. In *Proc. of the ICML Workshop on Clinical Data Analysis*, volume 46, pages 79–95, 2012.

[12] Nanlin Jin, Peter Flach, Tom Wilcox, Royston Sellman, Joshua Thumim, and Arno Knobbe. Subgroup Discovery in smart electricity meter data. *IEEE Transactions on Industrial Informatics*, 10(2):1327–1336, 2014.

[13] Brennan C. Kahan, Vipul Jairath, Caroline J. Doré, and Tim P. Morris. The risks and rewards of covariate adjustment in randomized trials: An assessment of 12 outcomes from 8 studies. *Trials*, 15(1):1–7, 2014.

[14] David M. Kent, Jessica K. Paulus, David van Klaveren, Ralph D'Agostino, Steve Goodman, Rodney Hayward, John P.A. Ioannidis, Bray Patrick-Lake, Sally Morton, Michael Pencina, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. *Annals of Internal Medicine*, 172(1):35–45, 2020.

[15] Willi Klösgen. Explora: A multipattern and multistrategy discovery assistant. In *Proc. PAKDD*, pages 249–271, 1996.

[16] Willi Klösgen and Michael May. Census data mining — An application. In *Proc. PKDD*, pages 65–79, 2002.

[17] Petra Kralj Novak, Nada Lavrač, and Geoffrey I Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10(2), 2009.

[18] Nada Lavrač, Branko Kavšek, Peter Flach, and Ljupčo Todorovski. Subgroup Discovery with CN2-SD. *Journal of Machine Learning Research*, 5(2):153–188, 2004.

[19] Nada Lavrac, Petra Kralj Novak, Igor Mozetic, Vid Podpecan, Helena Motaln, Marko Petek, and Kristina Gruden. Semantic Subgroup Discovery: Using ontologies in microarray data analysis. In *Proc. IEEE Engineering in Medicine and Biology Society*, pages 5613–5616, 2009.

[20] Dennis Leman, Ad Feelders, and Arno Knobbe. Exceptional Model Mining. In *Proc. ECML PKDD*, pages 1–16, 2008.

[21] Florian Lemmerich, Martin Becker, and Martin Atzmueller. Generic pattern trees for exhaustive Exceptional Model Mining. In *Proc. ECML PKDD*, pages 277–292, 2012.

[22] Todd D. Little, Noel A. Card, James A. Bovaird, Kristopher J. Preacher, and Christian S. Crandall. Structural Equation Modeling of mediation and moderation with contextual factors. *Modeling contextual effects in longitudinal studies*, 1:207–230, 2007.

[23] Romain Mathonat, Diana Nurbakova, Jean-François Boulicaut, and Mehdi Kaytoue. Anytime mining of sequential discriminative patterns in labeled sequences. *Knowledge and Information Systems*, 63(2):439–476, 2021.

[24] Romain Mathonat, Diana Nurbakova, Jean-François Boulicaut, and Mehdi Kaytoue. Anytime Subgroup Discovery in high dimensional numerical data. In *Proc. DSAA)*, pages 1–10, 2021.

[25] Marvin Meeng and Arno J. Knobbe. For real: A thorough look at numeric attributes in Subgroup Discovery. *Data Mining and Knowledge Discovery*, 35(1):158–212, 2021.

[26] Dennis Mollenhauer and Martin Atzmueller. Sequential exceptional pattern discovery using pattern-growth: An extensible framework for interpretable machine learning on sequential data. In *Proc. XI-ML*, 2020.

[27] Katharina Morik, Jean-François Boulicaut, and Arno Siebes. *Local Pattern Detection: International Seminar Dagstuhl Castle, Germany, April 12-16, 2004, Revised Selected Papers*, volume 3539. Springer, 2005.

[28] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[29] Judea Pearl. Detecting latent heterogeneity. *Sociological Methods & Research*, 46(3):370–389, 2017.

[30] Nicholas Radcliffe. Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal*, pages 14–21, 2007.

[31] Nicholas J Radcliffe and Patrick D Surry. Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions*, pages 1–33, 2011.

[32] Youcef Remil, Anes Bendimerad, Mathieu Chambard, Romain Mathonat, Marc Plantevit, and Mehdi Kaytoue. Mining java memory errors using subjective interesting subgroups with hierarchical targets. In *Proc. ICDMW*, pages 1221–1230, 2023.

[33] Craig Anthony Rolling. *Estimation of Conditional Average Treatment Effects*. PhD thesis, University of Minnesota, 2014.

[34] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[35] Krzysztof Rudaś and Szymon Jaroszewicz. Linear regression for uplift modeling. *Data Mining and Knowledge Discovery*, 32(5):1275–1305, 2018.

[36] Rianne M. Schouten, W. Duivesteijn, Pekka Räsänen, Jacob M. Paul, and Mykola Pechenizkiy. Exceptional Subitizing Patterns: Exploring Mathematical Abilities of Finnish Primary School Children with Piecewise Linear Regression. In *Proc. ECML PKDD*, 2024.

[37] Rianne Margaretha Schouten, Wouter Duivesteijn, and Mykola Pechenizkiy. Exceptional Model Mining for Repeated Cross-Sectional Data (EMM-RCS). In *Proc. SDM*, pages 585–593, 2022.

[38] Arno Siebes. Data Surveying: Foundations of an Inductive Query Language. In *Proc. KDD*, pages 269–274, 1995.

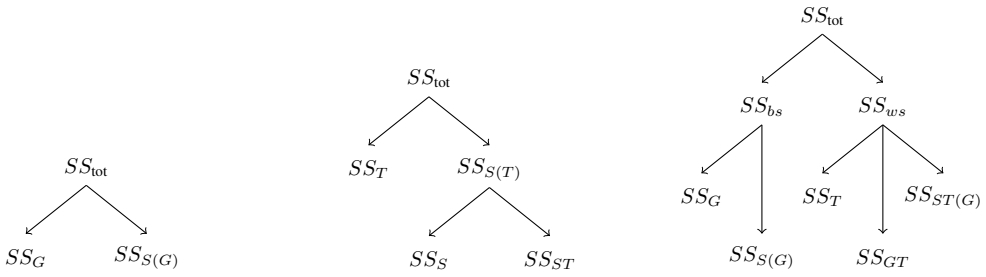
[39] Barbara G. Tabachnick and Linda S. Fidell. *Using multivariate statistics*. Pearson/Allyn & Bacon, 5th edition, 2007.

[40] Tyler J VanderWeele and James M Robins. Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology*, 18(5):561–568, 2007.

A More on bANOVA and wANOVA

An ANalysis Of VAriance (ANOVA) divides the variance in the outcome variable over one or multiple components [39]. An ANOVA is mathematically equivalent to linear regression in the case of binary-coded independent variables. In case of two treatment arms, a between-subjects ANOVA (bANOVA) therefore gives the same solution as regressing $Y \sim W$.

Specifically, a bANOVA divides the variance in the outcome variable between an effect of group effect of categorical variable G and a within-subjects variance that cannot be further explained and



(a) (b) (c)

Figure 5: Variance distribution in a numerical outcome variable for three ANOVA models. SS stands for Sum of Squares. (a) one-way between-subjects ANOVA (b) one-way within-subjects ANOVA (c) mixed between-within-subjects ANOVA.

is therefore considered left-over or error variance (Figure 5a). The total variance in the outcome variable, expressed in terms of Sum of Squares (SS), is

$$SS_{\text{tot}} = \sum_{k \in \{0,1\}} \sum_{i=1}^{n_k} (Y_{ik} - GM)^2, \tag{3}$$

where $GM = \frac{1}{N} \sum_{k \in \{0,1\}} \sum_{i=1}^{n_k} Y_{ik}$ is the Grand Mean. Any systematic difference between the groups is contained in the variance component

$$SS_G = n_k \sum_{k \in \{0,1\}} (\overline{Y}_k - GM)^2, \tag{4}$$

where $\overline{Y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} Y_{ik}$ is the mean of group k. Within a group, there are no variables that explain any variance and the left-over variance is

$$SS_{S(G)} = \sum_{k \in \{0,1\}} \sum_{i=1}^{n_k} (Y_{ik} - \overline{Y}_k)^2. \tag{5}$$

A bANOVA evaluates whether there is a difference in outcome Y between the groups, by comparing SS_G with $SS_{S(G)}$. While taking the degrees of freedom into account; if the former is substantially larger than the latter, differences in outcome between subjects can be explained by the group assignment. Then, the treatment has an effect.

Figure 5b displays the variance components in a within-subjects ANOVA (wANOVA). In a wANOVA, the grouping variable G is replaced with a time variable T, which indicates the time at which value Y_{it} was measured. Here, the idea is to follow subjects through time and Y_{it} is known for a subject i for all $t \in \{0, 1, .., T-1\}$.

In a wANOVA, the variance in the outcome variable is divided between an effect of time, an effect of systematic individual differences and a left-over variance. Compared to a bANOVA, the effect of time is calculated similar as the group effect in Equation (4).

In contrast to a bANOVA, the within-subjects variance in a wANOVA can be further explained by a subject effect,

$$SS_S = \sum_{t=0}^{T-1} \sum_{i=1}^{n_t} (\overline{Y}_i - GM)^2 \tag{6}$$

where $\overline{Y}_i = \frac{1}{T-1} \sum_{t=0}^{T-1} Y_{it}$ is a subject's average over the $T-1$ measurement occasions. Any further unexplained variance can be calculated by subtracting Equation (6) from Equation (5), or by

$$SS_{ST} = \sum_{t=0}^{T-1} \sum_{i=1}^{n_t} (Y_{it} - \overline{Y}_t - \overline{Y}_i + GM)^2. \tag{7}$$

Note that by writing ST in the subscript in SS_{ST}, we indicate that the left-over variance represents an interaction between subjects and time. In other words, any variance that cannot explained by a main effect of time, or a main effect of subject, is due to an interaction between these variables. In the context of linear relations, another word for interaction is effect modification.

It is important to realize that in a wANOVA, it is likely that the variance $SS_{ST} < SS_{S(T)}$. Consequently, it may be easier to find evidence for a main effect of time in a wANOVA than it is in a bANOVA, because SS_T is compared against a smaller remainder than SS_G. It does require some degrees of freedom though.

Furthermore, in a one-way wANOVA, there is no grouping variable other than time. Therefore, when one performs an RCT with repeatedly measured outcome values, the analysis to go to is a mixed between-within-subjects ANOVA (bwANOVA) rather than a wANOVA. A bwANOVA is shown in Figure 5c.

In this paper, we simulate the hypothetical situation that we know the factual and counterfactual outcomes. This allows us specifiy certain ITE distributions and to analyze variance components in a wANOVA. To translate our findings to a real RCT where covariate information is be used to explain variation in the data, one can use a two-way between-subjects ANOVA. There, a covariate is added as an extra grouping variable similarly like S in a wANOVA [39].

B More on Local Pattern Mining and Beam Search

Subgroup Discovery (SD) [15, 38, 16, 6, 18] and Exceptional Model Mining (EMM) [20, 5] aim to discover subgroups in the dataset that somehow behave exceptionally. Traditionally, SD focuses on exceptionality defined over 1 target attributes, whereas EMM evaluates a model fitted to ≥ 2 target attributes.

Traditionally, EMM assumes a dataset Ω to be a bag of n records $r \in \Omega$ of the form

$$r = (a_1, \ldots, a_k, \ell_1, \ldots, \ell_m), \tag{8}$$

where k and m are positive integers [5]. In EMM, we call $a_1, \ldots, a_k$ the *descriptive attributes* or *descriptors* of r, and $\ell_1, \ldots, \ell_m$ the *target attributes* or *targets* of r. For SD, $m = 1$, whereas for EMM, typically $m \geq 2$.

The descriptive attributes are used to describe and discover subgroups of cases. A subgroup is defined using descriptions; a description is a Boolean function $D : \mathcal{A} \to \{0, 1\}$ which covers a record r^i if and only if $D(a_1^i, \ldots, a_k^i) = 1$. Here, $\mathcal{A}$ is the collective domain from which the full set of descriptors is taken; a Cartesian product of the domains of each individual descriptor. Consequently, a subgroup is defined as follows:

Definition B.1 (Subgroup cf. [5]) *A subgroup corresponding to description D is the bag of records $G_D \subseteq \Omega$ that D covers:*

$$G_D = \{r^i \in \Omega \mid D(a_1^i, \ldots, a_k^i) = 1\}.$$

The complement contains all records that are not covered; $G^C = \Omega \setminus G_D$.

In EMM, the choice of *description language* $\mathcal{D}$ is free, though generally we let the description be a conjunction of selection conditions over the descriptors, where condition sel_j is a restriction on the domain $\mathcal{A}_j$ of the respective attribute a_j. For instance, for discrete variables the selector may be an attribute-value pair $(a_j = v)$; for continuous variables it could be a range of values $(w_1 \leq a_j \leq w_2)$ [5, 25, 37].

We aim to discover the descriptions for which the subgroups display exceptional behavior on a target model, fitted to a set of target attributes. Formally, we quantify exceptionality using a quality or interestingness measure. A quality measure quantifies the difference between behavior in the subgroup and some reference behavior, usually the the subgroup's complement:

Algorithm 1 Beam Search Algorithm cf. [5, Algorithm 1]

Input Dataset Ω, quality measure φ, refinement operator η, beam width w, beam depth d, result set size q, constraints $\mathcal{C}$

Output PriorityQueue resultSet

```
1: candidateQueue ← new Queue;
2: candidateQueue.enqueue({});
3: resultSet ← new PriorityQueue(q);
4: for (Integer level ← 1; level ≤ d; level++) do
5:     beam ← new PriorityQueue(w);
6:     while (candidateQueue ≠ ∅) do
7:         seed ← candidateQueue.dequeue();
8:         set ← η(seed);
9:         for all (desc ∈ set) do
10:            quality ← φ (desc);
11:            if (desc.SATISFIESALL(C)) then
12:                resultSet.insert_with_priority(desc,quality);
13:                beam.insert_with_priority(desc,quality);
14:    while (beam ≠ ∅) do
15:        candidateQueue.enqueue(beam.get_front_element());
16: return resultSet;
```

Definition B.2 (Quality Measure cf. [5]) *A quality measure is a function $\varphi : \mathcal{D} \to \mathbb{R}$ that assigns a numerical value to a description D.*

In this paper, we follow a SD scenario where outcome variable Y is our target attribute. We evaluate two types of exceptionalities: one where we aim to discover subgroups of individuals with high outcome values, and one where we aim to discover subgroups of individuals with a high increase in outcome values (from $t = 0$ to $t = (T - 1)$. The equations are given in the main text.

The task of SD and EMM is to effectively search through the space of candidate subgroups to find the top-q best-scoring subgroups [5]. Many search algorithms exist; some of them developed for particular kinds of exceptional behavior [e.g., 21, 4, 2, 23], others for particular data types [26, 24, 32, 19]. Nevertheless, most work on EMM considers the search space to be a general-to-specific search lattice and use a conjunction of selection conditions as description language. Then, the core difference between most search algorithms is the manner in which they traverse the search lattice; given a (candidate) subgroup description, the selection of subgroup members is comparable for many search algorithms.

In this paper, we choose *beam search* as our search algorithm of choice. We have two important reasons. First, beam search discovers exceptionally behaving subgroups using a heuristic search strategy. It is an intuitive method that can easily be understood at a conceptual, non-technical level. This makes beam search perfect when working together with experts from various domains (such as medical experts when studying the causal effects of treatments). Second, beam search is deterministic; given a dataset and fixed parameter settings, the algorithm will always return the same top-q subgroups.

The beam search algorithm is shown in Algorithm 1. In essence, beam search performs a level-wise search of d levels (line 4). At each level, w promising descriptions are selected into the beam (line 13); these descriptions are taken to the next level (line 15) and explored further. Exploration of candidate subgroups occurs by adding selection conditions to existing subgroup descriptions (line 8).

Following [5], the time complexity of the beam search algorithm is:

$$\mathcal{O}(dwkn(c + M(n, m) + \log(wq))). \tag{9}$$

Here, $M(n, m)$ is the time complexity of evaluating the quality of a target model on n records and m targets; c is the cost of comparing two models. After a pre-specified number of levels d, the top-q subgroups are returned.

In this paper, we perform beam search with the `lbca` discretization strategy [25] for numerical attributes with $b = 4$ quantiles, search width $w = 20$ and search depth $d = 3$.