

# On the role of prognostic factors and effect modifiers in unraveling population heterogeneity

Author information scrubbed for double-blind reviewing

No Institute Given

**Abstract.** Causal effects vary across subgroups in a population. This is especially true for treatment effects in the medical setting where context information like family history, medical history and symptom presentation influence and explain differences between individuals. Estimating these local, causal treatment effects is not straightforward, especially since the covariate (context) information may i) interact with the outcome and/or treatment variable, ii) be measured at a lower hierarchical level than the level of the individual, and iii) contain high dimensional observations. In this paper, we perform two synthetic data experiments to further analyze and provide insights into the behaviour of two types of non-confounding covariates: *effect modifiers* influence the treatment effect and *prognostic factors* influence the outcome variable, but not the treatment effect. We temporarily disregard the fundamental problem of causal inference that factual and counterfactual outcomes cannot be observed together, and generate synthetic data where treatment assignment variable  $W$  is considered a time-variant variable: the outcome value is observed for every occasion (treatment) for every patient. Consequently, we obtain a hierarchical dataset where outcomes (and possibly, covariates) reside at a lower hierarchical level than the treatment effects, which reside at the level of the individual. Our findings from controlled experiments provide 1) a demonstration of existing knowledge that prognostic factors reduce left-over variance and improve precision, whereas effect modifiers determine and explain individual-specific deviations, and 2) an additional insight that in order to unravel population heterogeneity, aggregations of lower-level, high-dimensional observations should align with (possibly unknown) expectations about whether these observations display prognostic or effect modification behavior.

**Keywords:** Prognostic Factors · Effect Modifiers · Individual Treatment Effects · Multilevel Data · Synthetic Data · Simulation

## 1 Introduction

People differ, not only in physical appearance but in personality, cultural background, abilities and interests, as well as in cognitive, emotional and social behavior [8,14,2,1]. This is particularly apparent in the medical setting, where context information like family history, medical history and symptom presentation influence and explain differences between individuals, including differences in

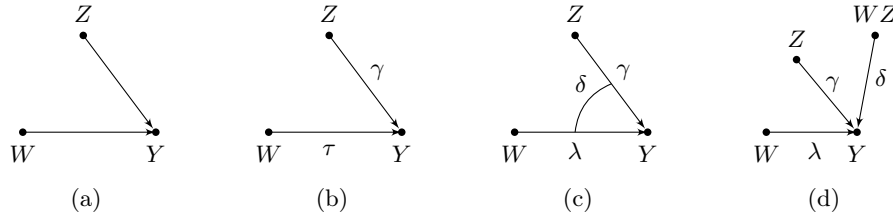


Fig. 1: Causal diagrams with three variables: outcome  $Y$ , treatment assignment  $W$  and covariate  $Z$  (a) no structural restrictions;  $Z$  can be a prognostic factor, effect modifier or both (b-d) the causal relations are assumed to be linear (b)  $Z$  is a prognostic factor (c)  $Z$  is an effect modifier, diagram cf. [37] (d)  $Z$  is an effect modifier, a variation of the diagram cf. [30].

causal treatment effects [41,10]. Considering such patient population heterogeneity, estimating Local or Conditional Average Treatment Effects (LATE, CATE) [41,48,22] for certain subgroups in the population is to be preferred over estimating an Average Treatment Effect (ATE) that aggregates over all individuals. Recent advances in Machine Learning (ML) move even further and estimate Individual Treatment Effect (ITE) [43,44,20,47,5,17].

A complicating factor in estimating treatment effects is the possibly nested, non-IID structure of observed covariate information. For instance, blood sampling tests may be repeated for each hospital visit, or blood glucose measurements are continuously taken using a FreeStyle Libre sensor, an intermittently continuous glucose monitoring (iCGM) sensor [12,7]. In this paper, we consider three particular aspects of observed covariate information that potentially influence the estimation of causal treatment effects: i) covariates may interact with the outcome variable and/or treatment assignment variable, ii) covariate values may be nested, that is, they may be measured at a lower hierarchical level than the level of the individual, and iii) covariate values may be observed with high dimensionality.

Learning causal representations without having access to ground-truth knowledge about causal relationships between variables, that is, without knowing the causal diagram, is non-trivial. Solving this problem requires, at the very least, a thorough understanding of how causal variables behave in a controlled setting. Therefore, this paper aims to analyze and provide insights into the behaviour of two types of non-confounding covariates.

On the one hand, *effect modifiers* influence the treatment effect (but not the treatment assignment) [51]. 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” [22, p.35]. On the other hand, we consider *prognostic factors*, which are variables that influence the outcome, but not the treatment effect [21] (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 [36,51,46] (Figure 1a).<sup>1</sup> 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. [37], and Figure 1d cf. [30]). Accordingly, the relation between treatment assignment variable  $W$ , covariate  $Z$  and outcome  $Y$  is modeled as

$$Y = \tau W + \gamma Z + \delta WZ, \quad (1)$$

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

The aim of this paper is to provide insights into the prognostic and effect modification behavior of covariate  $Z$ . We do this in a controlled setting, where we temporarily disregard the fundamental problem of causal inference that factual and counterfactual outcomes cannot be observed together [16].<sup>2</sup> and generate synthetic data where treatment assignment variable  $W$  is a time-variant variable: the outcome value is observed for every occasion (treatment) for every patient. Consequently, we obtain a nested, hierarchical data structure where outcomes (and covariates) reside at a lower hierarchical level than the level of the individual. We then study how population heterogeneity is influenced by whether covariate  $Z$  is a prognostic factor or an effect modifier.

The rest of this paper is structured as follows. In Section 2, we provide background information and notation for hierarchical data. We extensively introduce the techniques used in our experiments, that is, in Section 3 we assume covariate  $Z$  is an individual-level variable and we analyze three types of ITE distributions using within-subjects ANalysis Of VAriance (wANOVA), whereas in Section 4  $Z$  is a time-variant variable and we perform a local pattern mining method called Subgroup Discovery (SD) [15,9] to unravel patient population heterogeneity in an unsupervised setting. Section 5 concludes.

## 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  $\mathbf{x}_i = (w_i, z_i, y_i)$ .

<sup>1</sup> In Figure 1, we do not display the noise variables and assume they are jointly independent; the models satisfy the causal Markov condition [36].

<sup>2</sup> At the same time, we stay close to reality by making assumptions such as the Stable Unit Treatment Value Assumption (SUTVA) [42,19], no hidden confounders and overlap (i.e., positivity) [5].

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  $i \in \{1, 2, \dots, n\}$  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.

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  $(\mathbf{x}_i^1, \dots, \mathbf{x}_i^t, \dots, \mathbf{x}_i^T)$ . 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  $i, j \in \{1, 2, \dots, n\}$ ). 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) [18].<sup>3</sup>

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.<sup>4</sup> 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 [50]. 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.

<sup>3</sup> 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.

<sup>4</sup> We temporarily disregard the fundamental problem of causal inference [16], but still assume SUTVA [19]. We do not believe that time influences the treatment effect; we use time to model the treatment effect.

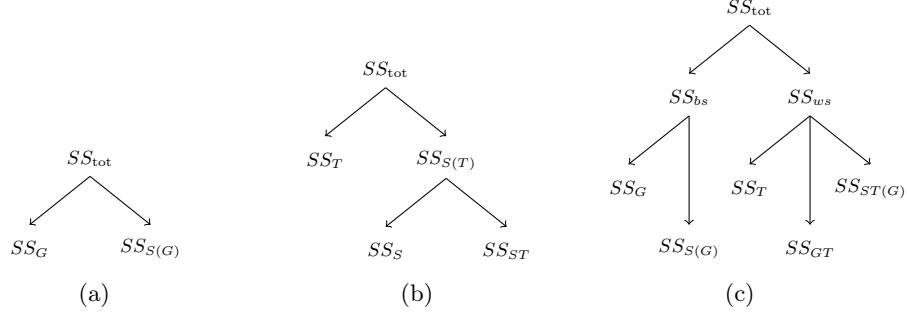


Fig. 2: Variance distribution in a numerical outcome variable for three ANOVA models. SS stands for Sum of Squares. (a) one-way between-subjects ANOVA (b) one-way within-subjects ANOVA (c) mixed between-within-subjects ANOVA.

To be specific, a between-subjects ANOVA (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 is therefore considered left-over or error variance (Figure 2a). The total variance in the outcome variable, expressed in terms of Sum of Squares (SS), is

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

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

$$SS_G = n_k \sum_{k \in \{0,1\}} (\bar{Y}_k - GM)^2, \quad (3)$$

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

$$SS_{S(G)} = \sum_{k \in \{0,1\}} \sum_{i=1}^{n_k} (Y_{ik} - \bar{Y}_k)^2. \quad (4)$$

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 2b 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  $t \in \{0, 1, \dots, T-1\}$ .

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 (3).

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

$$SS_S = \sum_{t=0}^{T-1} \sum_{i=1}^{n_t} (\bar{Y}_i - GM)^2 \quad (5)$$

where  $\bar{Y}_i = \frac{1}{T-1} \sum_{t=0}^{T-1} Y_{it}$  is a subject's average over the  $T-1$  measurement occasions. Any further unexplained variance can be calculated by subtracting Equation (5) from Equation (4), or by

$$SS_{ST} = \sum_{t=0}^{T-1} \sum_{i=1}^{n_t} (Y_{it} - \bar{Y}_t - \bar{Y}_i + GM)^2. \quad (6)$$

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 2c.

In this paper, we simulate the hypothetical situation that we know the factual and counterfactual outcomes. This allows us specify 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 [50].

### 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 [13,35]. 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* [23,49,24,25,3,9]. 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 (i.e., high treatment effects). We expect that prognostic factors are valuable for discovering the first type of subgroups, whereas effect modifiers are needed to discover the second type.

The next paragraphs provide more background information on two types of LPM frameworks: Subgroup Discovery (SD) [23,49,24,11,26] and Exceptional Model Mining (EMM) [28,9]. Both frameworks 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  $\geq 2$  target attributes.

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

$$r = (a_1, \dots, a_k, \ell_1, \dots, \ell_m), \quad (7)$$

where  $k$  and  $m$  are positive integers [9]. In EMM, we call  $a_1, \dots, a_k$  the *descriptive attributes* or *descriptors* of  $r$ , and  $\ell_1, \dots, \ell_m$  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  $D : \mathcal{A} \rightarrow \{0, 1\}$  which covers a record  $r^i$  if and only if  $D(a_1^i, \dots, a_k^i) = 1$ . 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 1 (Subgroup cf. [9]).** *A subgroup corresponding to description  $D$  is the bag of records  $G_D \subseteq \Omega$  that  $D$  covers:*

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

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 ( $w_1 \leq a_j \leq w_2$ ) [9,33,45].

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:

**Definition 2 (Quality Measure cf. [9]).** *A quality measure is a function  $\varphi : \mathcal{D} \rightarrow \mathbb{R}$  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 exceptionality: 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 proposed quality measure (equations) are given in Section 4.1.

The task of SD and EMM is to effectively search through the space of candidate subgroups to find the top- $q$  best-scoring subgroups [9]. Many search algorithms exist; some of them developed for particular kinds of exceptional behavior [29,6,4,31], others for particular data types [34,32,40,27]. 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 cf. Algorithm 1 in [9] as our search algorithm of choice, with the `lbca` discretization strategy [33] for numerical attributes with  $b = 4$  quantiles, search width  $w = 20$  and search depth  $d = 3$ .

### 3 Experiment 1: $Z$ is an individual-level covariate

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  $i \in \{1, 2, \dots, n\}$  and  $t \in \{0, 1\}$ . We then sample outcome values  $y_i^t = \pi_{0i} + \pi_{1i}w_i^t + e_i^t$ . 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 3. In Figure 3a, 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 3b). 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  $\bar{y}^0$  for the control group and  $\bar{y}^1$  for the treatment group. Thus, if we randomly sample  $y_i^{t=1}$  for individual  $i$ , then value  $y_i^0 \leftarrow \bar{y}^0$ , and similar for  $t = 0$ . The resulting ITE distribution is visualized in Figure 3c. 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  $\pi_{0i}$  and  $\pi_{1i}$  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:

$$y_i^t = \beta_{00} + \beta_{01}z_i + \beta_{10}w_i^t + \beta_{11}w_i^t z_i + \mu_{1i}w_i^t + \mu_{0i} + e_i^t. \quad (8)$$

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 4. Here, the left-most column shows the traditional bANOVA output and the three right columns the wANOVA results. We see that the total variance  $SS_{\text{tot}}$  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 4 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 ( $\bar{y}_i$ ) and the GM; no further variance is to be explained.

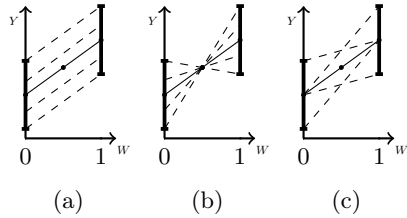


Fig. 3: Visualizations of three types of ITE distributions in a two-arm experiment with  $Y$  as the outcome variable and  $W$  the treatment assignment indicator. (a) all ITEs equal the ATE (b) all ITEs cross through GM (c) counterfactuals equal the group mean.

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

Fig. 4: 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 3.

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 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: $Z$ is a time-variant, lower-level covariate

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  $y_i^t = \mathcal{N}(10 + v_i^t, 0.1)$ , where we let  $v_i^t = \alpha w_i^t + \zeta z_i^t + \lambda w_i^t z_i^t$  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 and construct  $2 \times 2 \times 2$  scenarios: all combinations of  $Ztype \in \{\text{prognostic, effect modification}\}$ ,  $Aggfunc \in \{\text{imperfect, perfect}\}$  and  $Eval \in \{\text{average, increase}\}$ .

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 5a). There are main effects of

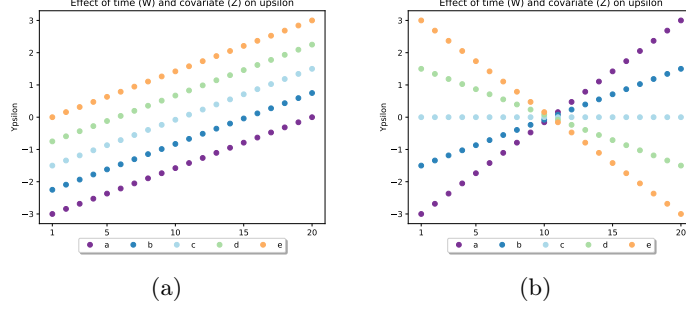


Fig. 5: Visualizations of the relation between the time indicator  $t$  (through  $w_i^t$ , x-axis), covariate values  $z$  (colors, from low (event type  $a$ , purple) to high (event type  $e$ , orange)) and  $v$ . Every individual  $i$  walks a path through these dots from left to right. (a)  $Z$  acts as a prognostic factor (b)  $Z$  acts as an effect modifier, there are no main effects.

time and variable  $Z$ , but no interaction effect. In contrast, Figure 5b 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  $v_i^t$  and average those using  $f_{\text{sum}}(v_i^1, \dots, v_i^T) = \sum_{t=1}^T v_i^t$  and  $f_{\text{incr}}(v_i^1, \dots, v_i^T) = v_i^T - v_i^1$ . 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 5 follows the same color) whereas the index location reflects 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  $f_{\text{freq}_h}$  and  $f_{\text{idx}_h}$  for  $h \in \mathcal{H}$ .

Third, we evaluate the exceptionality of individuals in two ways. If  $Eval = \text{average}$ , 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,  $\mu^{SG} = 1/n^{SG} \sum_{i \in SG} \sum_{t=1}^T y_i^t$  is the average outcome value of the individuals covered by the subgroup,  $\text{se}(\mu^{SG})$  is its standard error, and  $\mu^D$  is the average outcome in the entire dataset  $D$ . In contrast, if  $Eval = \text{increase}$ , 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  $Eval = \varphi_\mu$ ,  $Ztype = \text{prognostic}$ , and  $Aggfunc = \text{imperfect}$ , the subgroups covers individuals with low frequency

Table 1: Description of the most exceptional subgroup, discovered with  $\varphi_\mu$  (exceptionally high average outcome) and  $\varphi_\theta$  (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.

<i>Eval</i>	<i>Ztype</i>	<i>Aggfunc</i>	
		imperfect	perfect
$\varphi_\mu$	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	$f_{\text{idx}_a} \geq 11 \wedge f_{\text{idx}_e} \leq 10$	$f_{\text{sum}} \in [-134, -25]$
$\varphi_\theta$	prognostic	$f_{\text{idx}_e} \geq 10.5 \wedge f_{\text{idx}_a} \leq 9.5$	$f_{\text{incr}} = 3$
	effect modification	$f_{\text{freq}_e} \leq 3 \wedge f_{\text{freq}_a} \geq 4 \wedge f_{\text{idx}_e} \leq 14$	$f_{\text{incr}} \in [-38, -29]$

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 5a. They have the highest  $v$  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 5a).

The reversed line of reasoning holds if  $Ztype = \text{effect modification}$ , as depicted in Figure 5b. Then, discovering individuals with high outcome values ( $\varphi_\mu$ ) requires crossing through colors and hence, the usage of  $f_{\text{idx}}$  (second row in Table 1, whereas individuals with a high increase ( $\varphi_\theta$ ) 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  $v$ , we can construct an aggregation function  $f_{\text{sum}}$  that works for discovering individuals with high outcome values ( $\varphi_\mu$ ) independent of whether  $Z$  is a prognostic factor or effect modifier (right column in Table 1,  $f_{\text{sum}}$  is used in both rows corresponding to  $\varphi_\mu$ ). Similarly, we can construct  $f_{\text{incr}}$  in combination with ( $\varphi_\theta$ ), independent of the covariate role of  $Z$ .

## 5 Discussion and Conclusion

The aim of this paper is to unravel patient population heterogeneity by providing insights into the prognostic and effect modification behavior of covariate  $Z$  in the context of estimating causal treatment effects that vary across subgroups in the population. 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.

In Experiment 1 in Section 3, covariate  $Z$  is an individual-level covariate. Our findings demonstrate that variance distributions in a traditional two-arm experiment as measured by bANOVA (most-left column in Table 4) are the same as those in a wANOVA with covariate  $Z$  being an effect modifier (scenario (b) in Figure 3 and Table 4). 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  $\tau$  [19,50]: we would move from scenario (b) to scenario (c) to scenario (a) in Table 4.

Remark that in the real-world, it is non-trivial to distinguish prognostic factors from effect modifiers. Consequently, it 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 [38,39]. In fact, the scenario in Figure 3c represents a baseline double-model where the simplest counterfactual model is the group average ( $\bar{y}^k$ ). More advanced methods directly estimate the net difference between two treatment groups [20,43].

In Experiment 2 in Section 4,  $Z$  is a time-variant, lower-level covariate. Our findings demonstrate that the quality of our choice on how to create individual-level representations of lower-level measurements is determined by our assumption about the nature of the lower-level covariate  $Z$ . 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  $f_{\text{sum}}$  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.** Acknowledgments scrubbed for double-blind reviewing.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. op den Akker, H., Cabrita, M., op den Akker, R., Jones, V.M., Hermens, H.J.: Tailored motivational message generation: A model and practical framework for real-time physical activity coaching. *Journal of Biomedical Informatics* **55**, 104–115 (2015)
2. op den Akker, H., Jones, V.M., Hermens, H.J.: Tailoring real-time physical activity coaching systems: A literature survey and model. *User modeling and user-adapted interaction* **24**, 351–392 (2014)
3. Atzmueller, M., Mueller, J., Becker, M.: Exploratory subgroup analytics on ubiquitous data. In: *International Workshop on Mining Ubiquitous and Social Environments*. pp. 1–20 (2013)
4. Belfodil, A., Duivesteijn, W., Plantevit, M., Cazalens, S., Lamarre, P.: DEvIANT: Discovering significant exceptional (dis-) agreement within groups. In: *Proc. ECML PKDD*. pp. 3–20 (2020)
5. Bica, I., Alaa, A.M., Lambert, C., van der Schaar, M.: 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)
6. Bosc, G., Boulicaut, J.F., Raïssi, C., Kaytoue, M.: Anytime discovery of a diverse set of patterns with Monte Carlo Tree Search. *Data Mining and Knowledge Discovery* **32**(3), 604–650 (2018)
7. den Braber, N., Vollenborek-Hutten, M.M., Westerik, K.M., Bakker, S.J., Navis, G., van Beijnum, B.J.F., Laverman, G.D.: Glucose regulation beyond HbA1c in type 2 Diabetes treated with insulin: Real-world evidence from the DIALECT-2 cohort. *Diabetes Care* **44**(10), 2238–2244 (2021)
8. De Vries, H., Dijkstra, M., Kuhlman, P.: Self-efficacy: The third factor besides attitude and subjective norm as a predictor of behavioural intentions. *Health Education Research* **3**(3), 273–282 (1988)
9. Duivesteijn, W., Feelders, A.J., Knobbe, A.: Exceptional Model Mining. *Data Mining and Knowledge Discovery* **30**(1), 47–98 (2016)
10. Foster, J.C., Taylor, J.M., Ruberg, S.J.: Subgroup identification from randomized clinical trial data. *Statistics in Medicine* **30**(24), 2867–2880 (2011)
11. Gamberger, D., Lavrac, N.: Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research* **17**, 501–527 (2002)
12. Gant, C.M., Binnenmars, S.H., van den Berg, E., Bakker, S.J., Navis, G., Laverman, G.D.: Integrated assessment of pharmacological and nutritional cardiovascular risk management: Blood pressure control in the DIABetes and LiFEstyle Cohort Twente (DIALECT). *Nutrients* **9**(7), 709 (2017)
13. Hand, D.J.: Pattern detection and discovery. In: *Proc. Pattern Detection and Discovery: ESF Exploratory Workshop London*. pp. 1–12 (2002)
14. Hermens, H., op den Akker, H., Tabak, M., Wijsman, J., Vollenbroek, M.: Personalized coaching systems to support healthy behavior in people with chronic conditions. *Journal of Electromyography and Kinesiology* **24**(6), 815–826 (2014)
15. Herrera, F., Carmona, C.J., González, P., Del Jesus, M.J.: An overview on Subgroup Discovery: Foundations and applications. *Knowledge and Information Systems* **29**(3), 495–525 (2011)

16. Holland, P.W.: Statistics and causal inference. *Journal of the American Statistical Association* **81**(396), 945–960 (1986)
17. Hoogland, J., IntHout, J., Belias, M., Rovers, M.M., Riley, R.D., E. Harrell Jr, F., Moons, K.G., Debray, T.P., Reitsma, J.B.: A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. *Statistics in medicine* **40**(26), 5961–5981 (2021)
18. Hox, J., Moerbeek, M., van de Schoot, R.: *Multilevel analysis: Techniques and applications*. Routledge (2017)
19. Imbens, G.W., Rubin, D.B.: *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press (2015)
20. Jaskowski, M., Jaroszewicz, S.: Uplift modeling for clinical trial data. In: *Proc. of the ICML Workshop on Clinical Data Analysis*. vol. 46, pp. 79–95 (2012)
21. Kahan, B.C., Jairath, V., Doré, C.J., Morris, T.P.: The risks and rewards of covariate adjustment in randomized trials: An assessment of 12 outcomes from 8 studies. *Trials* **15**(1), 1–7 (2014)
22. Kent, D.M., Paulus, J.K., van Klaveren, D., D’Agostino, R., Goodman, S., Hayward, R., Ioannidis, J.P., Patrick-Lake, B., Morton, S., Pencina, M., et al.: The predictive approaches to treatment effect heterogeneity (PATH) statement. *Annals of Internal Medicine* **172**(1), 35–45 (2020)
23. Klösgen, W.: Explora: A multipattern and multistrategy discovery assistant. In: *Proc. PAKDD*. pp. 249–271 (1996)
24. Klösgen, W., May, M.: Census data mining — An application. In: *Proc. PKDD*. pp. 65–79 (2002)
25. Kralj Novak, P., Lavrač, N., Webb, G.I.: Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research* **10**(2) (2009)
26. Lavrač, N., Kavšek, B., Flach, P., Todorovski, L.: Subgroup Discovery with CN2-SD. *Journal of Machine Learning Research* **5**(2), 153–188 (2004)
27. Lavrac, N., Novak, P.K., Mozetic, I., Podpecan, V., Motaln, H., Petek, M., Gruden, K.: Semantic Subgroup Discovery: Using ontologies in microarray data analysis. In: *Proc. IEEE Engineering in Medicine and Biology Society*. pp. 5613–5616 (2009)
28. Leman, D., Feelders, A., Knobbe, A.: Exceptional Model Mining. In: *Proc. ECML PKDD*. pp. 1–16 (2008)
29. Lemmerich, F., Becker, M., Atzmueller, M.: Generic pattern trees for exhaustive Exceptional Model Mining. In: *Proc. ECML PKDD*. pp. 277–292 (2012)
30. Little, T.D., Card, N.A., Bovaird, J.A., Preacher, K.J., Crandall, C.S.: Structural Equation Modeling of mediation and moderation with contextual factors. *Modeling contextual effects in longitudinal studies* **1**, 207–230 (2007)
31. Mathonat, R., Nurbakova, D., Boulicaut, J.F., Kaytoue, M.: Anytime mining of sequential discriminative patterns in labeled sequences. *Knowledge and Information Systems* **63**(2), 439–476 (2021)
32. Mathonat, R., Nurbakova, D., Boulicaut, J.F., Kaytoue, M.: Anytime Subgroup Discovery in high dimensional numerical data. In: *Proc. DSAA*. pp. 1–10 (2021)
33. Meeng, M., Knobbe, A.J.: For real: A thorough look at numeric attributes in Subgroup Discovery. *Data Mining and Knowledge Discovery* **35**(1), 158–212 (2021)
34. Mollenhauer, D., Atzmueller, M.: Sequential exceptional pattern discovery using pattern-growth: An extensible framework for interpretable machine learning on sequential data. In: *Proc. XI-ML* (2020)
35. Morik, K., Boulicaut, J.F., Siebes, A.: *Local Pattern Detection: International Seminar Dagstuhl Castle, Germany, April 12-16, 2004, Revised Selected Papers*, vol. 3539. Springer (2005)

36. Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press, 2nd edn. (2009)
37. Pearl, J.: Detecting latent heterogeneity. *Sociological Methods & Research* **46**(3), 370–389 (2017)
38. Radcliffe, N.: Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal* pp. 14–21 (2007)
39. Radcliffe, N.J., Surry, P.D.: Real-world uplift modelling with significance-based uplift trees. White Paper TR-2011-1, Stochastic Solutions pp. 1–33 (2011)
40. Remil, Y., Bendimerad, A., Chambard, M., Mathonat, R., Plantevit, M., Kay-toue, M.: Mining java memory errors using subjective interesting subgroups with hierarchical targets. In: *Proc. ICDMW*. pp. 1221–1230 (2023)
41. Rolling, C.A.: Estimation of Conditional Average Treatment Effects. Ph.D. thesis, University of Minnesota (2014)
42. Rubin, D.B.: Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* **100**(469), 322–331 (2005)
43. Rudaś, K., Jaroszewicz, S.: Linear regression for uplift modeling. *Data Mining and Knowledge Discovery* **32**(5), 1275–1305 (2018)
44. Rudaś, K., Jaroszewicz, S.: Shrinkage estimators for uplift regression. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*. pp. 607–623. Springer (2019)
45. Schouten, R.M., Duivesteijn, W., Pechenizkiy, M.: Exceptional Model Mining for Repeated Cross-Sectional Data (EMM-RCS). In: *Proc. SDM*. pp. 585–593 (2022)
46. Sechidis, K., Papangelou, K., Metcalfe, P.D., Svensson, D., Weatherall, J., Brown, G.: Distinguishing prognostic and predictive biomarkers: An information theoretic approach. *Bioinformatics* **34**(19), 3365–3376 (2018)
47. Shalit, U., Johansson, F.D., Sontag, D.: Estimating individual treatment effect: generalization bounds and algorithms. In: *International conference on machine learning*. pp. 3076–3085. PMLR (2017)
48. Shpitser, I., Pearl, J.: Identification of conditional interventional distributions. *arXiv preprint arXiv:1206.6876* (2012)
49. Siebes, A.: Data Surveying: Foundations of an Inductive Query Language. In: *Proc. KDD*. pp. 269–274 (1995)
50. Tabachnick, B.G., Fidell, L.S.: Using multivariate statistics. Pearson/Allyn & Bacon, 5th edn. (2007)
51. VanderWeele, T.J., Robins, J.M.: Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology* **18**(5), 561–568 (2007)