Automated Identification of Causal Moderators in Time-Series Data

Min Zheng Mzheng3@stevens.edu
Department of Computer Science
Stevens Institute of Technology
Hoboken, NJ 07030, USA

Jan Claassen jc1439@columbia.edu
Department of Neurology
Columbia University
New York, NY 10027, USA

Samantha Kleinberg samantha.kleinberg@stevens.edu
Department of Computer Science
Stevens Institute of Technology
Hoboken, NJ 07030, USA

Editor: Thuc Duy Le, Kun Zhang, Emre Kıcıman, Aapo Hyvärinen, and Lin Liu

Abstract

Causal inference is often taken to mean finding links between individual variables. However in many real-world cases, such as in biological systems, relationships are more complex, with groups of factors needed to produce an effect, or some factors only modifying other relationships rather than producing outcomes alone. For instance, weight may alter the efficacy of a drug without causing side effects itself. Such moderating factors may change the timing, intensity, or probability of a causal relationship. Distinguishing moderators from genuine causes can lead to more effective medical interventions, and better strategies for bringing about a desired effect, since a moderator alone is ineffective. However, there have not yet been algorithms to automatically infer moderators in a large-scale automated way, and they cannot be easily read off from causal graphs. We introduce a set of temporal logic rules to automatically identify the asymmetric roles of causes and moderators in a computationally efficient manner. Experiments on simulated data demonstrate that even in challenging cases we can find moderators and avoid confounding, and on real neurological ICU data we show how the approach can find more descriptive and meaningful relationships than the state of the art.

Keywords: causality, time series, health informatics

1. Introduction

As causal inference methods are applied to increasingly large datasets with many variables, the resulting models can be prohibitively complex for people to reason with. Gene regulatory networks have thousands of densely connected genes, and models of disease risk contain many environmental and biological variables. Causal models are useful specifically because they can guide interventions, yet a graph or set of pairwise relationships does not capture the different roles of each variable. For example, race may moderate the effect of a particular

drug, and exercise in the past 24 hours can moderate the effect of insulin in people with diabetes, but these factors are not causal by themselves. On the other hand, genuine causes can be manipulated to bring about effects. To successfully intervene we need to know which factors are genuinely causal, and which intensify or weaken a relationship, but are not necessary to produce an outcome.

For these reasons, analysis of moderating factors is routine in medical research, psychology, statistics and other areas (Bauman et al., 2002; Kraemer et al., 2002). However, these works are generally hypothesis-driven, beginning with an effect of interest, such as the primary outcome of a randomized controlled trial or stress levels in a population, and a set of measured potential causes whose role is being studied. In contrast, computational inference focuses on finding relationships in a data-driven way, resulting in a model or set of causal relationships that describe all causal links between pairs of measured variables. However this does not distinguish between variables with different roles. While a cause is a useful target to intervene on to produce an effect, a moderator alone is not. On the other hand, applying existing methods for moderator analysis to all combinations of variables in a large dataset is computationally prohibitive.

We introduce a method to efficiently find causal moderators: factors that act like control knobs that strengthen or weaken a cause's impact or change when an effect occurs. The key is that causes and moderators have asymmetric roles in producing an effect: whether or not a moderator is present, a cause will have an impact on its effect, while a moderator in the absence of any causes will not. We show how criteria for finding moderators can be represented as logical formulas and can be efficiently tested, without increasing computational complexity or testing all combinations of variables. After showing the correctness of the approach on simulated data, we apply the method to neurological ICU data, demonstrating that it can identify a larger number of more descriptive relationships of the physiology of stroke recovery than existing methods for moderator analysis.

2. Related work

2.1 Moderator analysis

While a mediator can be viewed as a link in the chain between a cause and effect, moderation is where a third variable affects the direction or strength of a causal relationship (Baron and Kenny, 1986). Note that moderation (or an effect modifying variable) is not the same as interaction, where two or more causes have a non-additive effect when combined (VanderWeele, 2009). Conceptually, finding moderators is related to finding subgroups where effects differ, and understanding heterogeneity in causal effects (Wang and Ware, 2013). To assess causal heterogeneity of treatment effects, commonly used nonparametric methods include nearest-neighbor matching (Crump et al., 2008), kernel methods (Lee, 2009), and series estimation (Willke et al., 2012). However, these methods quickly break down when the number of covariates of data increases. To address these limitations, Wager and Athey (2017) proposed causal forests, which extend the random forest algorithm to estimate heterogeneous treatment effects. This estimator can construct valid asymptotic confidence intervals even as the number of covariates increases, but cannot handle latent variables. To address latent heterogeneity, Pearl (2015) showed that one can assess heterogeneous events without knowing the factors responsible for the heterogeneity. However, this relies on se-

lecting the correct measurements, and accurately estimating the difference between treated and untreated groups.

Most methods for moderator analysis assume that there is an effect of interest whose causes and their roles are to be found. Methods are primarily based on regression, such as Structural Nested Mean Models (SNMM) when treatments vary over time (Almirall et al., 2010; Robins, 1994) or marginal structural models in other cases (Robins, 2000). However, in data-driven causal inference, we aim to find all causes and effects, and need computationally efficient strategies for finding moderators. Besides, regression-based methods (e.g. Fairchild and MacKinnon, 2009) share the same problem that convergence cannot be guaranteed when sample size is small. Almirall et al. (2010) proposed a parametric two-stage regression estimator (2-SE here) built on Robins' structural nested models (SNMs) to assess time varying effect moderation. This method outperforms Robins' G-estimator (Robins, 1994) but the two-state estimator requires more information about covariates, and if this information is incorrect, results will be biased.

2.2 Data-driven causal inference

Causal inference on the other hand has focused on uncovering connections between variables in a dataset, where the output is either a set of relationships or graph with edges from causes to effects, such as a Bayesian network (BN) (Pearl, 2000; Spirtes et al., 2000) or its dynamical (Voortman et al., 2010) or temporal (Eichler and Didelez, 2007; Song et al., 2009) extensions. A BN along with a set of probability distributions encodes the probabilistic dependencies in a dataset, and with some assumptions (causal Markov condition, sufficiency, and faithfulness), is a causal model. These approaches have the advantage of not requiring temporal data, though they become increasingly complex to infer as the number of variables increases. One cannot easily read off complex moderating relationships from a BN and probability tables, but such an analysis is required to avoid ineffective interventions on a moderator alone. As shown in fig. 1c, if A causes E, with moderators B, C, and D, one compatible graphical model is where each of the variables is a cause of E. However, an intervention to produce E that doesn't involve A will be totally ineffective because moderators are not causes. This cannot be improved by adding more complex nodes such as AB, AC and AD, as complex causes formed by a conjunction of factors mean the effect only occurs in one state of the factor (e.g., A and B both true), whereas in moderating the effect simply differs.

To address this, VanderWeele and Robins (2007) gave criteria for classifying direct and indirect modification from a directed acyclic graph (DAG). However, identification of moderators is not guaranteed, as it assumes that specific structures will result from moderating, while multiple DAGs are compatible with the same relationships. Further, this does not identify the strength of modifiers. Given the relationship between moderation and finding subgroups, some approaches focus on more efficient subgrouping, though these generally do not distinguish the roles of individual variables (Athey and Imbens, 2015) or have not addressed complex time series data (Green and Kern, 2012). To distinguish the different roles of variables, Videla et al. (2015) proposed a method to identify whether a variable activates or inhibits another independently or if a third variable is required. However, this approach cannot be applied to time series data and has high computational

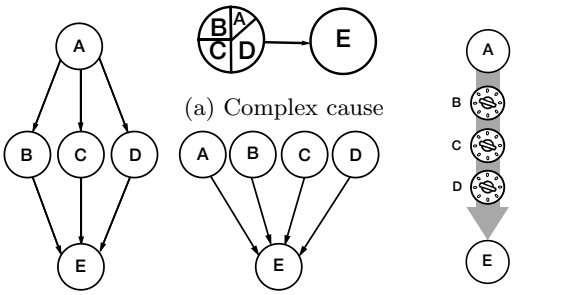


Figure 1: In each case A causes E, but how best to intervene to produce E differs.

complexity due to the exhaustive search. Su et al. (2012) introduced the facilitating score to handle confounding and interaction, using a recursive decomposition to create subgroups where the causal effect is homogeneous, and to estimate individual and average causal effects. That work does not distinguish between effect modification and interaction, even though this is important for interventions, and it assumes strong ignorability, meaning treatment assignment is independent of treatment response and unmeasured confounders.

Other works focus on finding causal structures from continuous variables when the observed variables are influenced by low dimensional latent variables (Silva et al., 2006; Kummerfeld and Ramsey, 2016; Monti and Hyvärinen, 2018), though these approaches have not been applied to time series data. There is a relationship to moderator analysis, in that these works look for factors other than observed causes that may affect the observed variables, however, they assume that each measured variable is conditionally dependent on a single latent variable and do not identify moderators specifically.

3. Methods

We propose that moderators can be more precisely represented as in fig. 1d, as control knobs along causal arrows. The key is that a moderator does not play the same role as a cause. We now show how the asymmetry of the relationship can be used to efficiently identify moderators.

Notation: Throughout the paper we use uppercase letters to denote variables or sets of variables (e.g., X, in eq. 3), and lowercase letters to denote specific values of variables (e.g., c in eq. 3).

3.1 Definitions

Before introducing our approach for efficiently identifying moderators, we begin by distinguishing between ways two or more causes may work together, and moderation.

- **Mediating** is when causal influence flows through a variable (e.g., causal chain). In total mediation, a mediator screens off cause and effect, so $P(\text{effect}|\text{mediator, cause}) = P(\text{effect}|\text{mediator})$.
- **Interaction** is when two or more causes have a different outcome together than individually. If two drugs have mild side effects when taken alone, but severe side

effects when taken together, then $P(e|d_1, d_2) \gg \max(P(e|d_1), P(e|d_2))$, but both drugs are individually causal.

- **Complex causes** are a group of factors that must be present for an effect to occur, where each factor alone is insufficient. Thus, $P(e|a, b) \gg 0$ while $P(e|a, \neg b) = P(e|b, \neg a) = 0$.
- **Moderators** attenuate or intensify a relationship, or change when an effect will occur, but are not themselves causes.

3.2 Key observation

The key difference between a moderator and a mediator (fig. 1b) or complex cause (fig. 1a) is that there is an asymmetry between cause and moderator, and we can use that to find these relationships. This asymmetry is not present in the other types of causal relationships defined above. Mediating relationships, for example, are akin to a causal chain, so if X causes Z via Y there are causal relationships $X \rightarrow Y$ and $Y \rightarrow Z$. Thus, X can be seen as an indirect cause of Z, while Y is a direct cause. Each can bring about Z, though interventions on Y may be more successful. In contrast, in a moderating relationship (e.g., Y moderates relationship $X \rightarrow Z$), X does not bring about the moderator Y, and there is a single causal relationship. Similarly, when X and Y are part of a complex cause that produces Z, then removing X or removing Y means Z will fail to occur (or have low probability). Here, neither X nor Y is a better strategy for bringing about Z, they are both equally necessary for the effect. To distinguish between moderators and causes, we simply need to test combinations of variables to find such asymmetries. That is, when X causes Z, and Y is a moderator of this relationship, the probability, timing, or value (in continuous cases) of Z will differ for $X \wedge Y$ and $X \wedge \neg Y$, and $\neg X \wedge Y$ will not be a significant cause of Z.

3.3 Background

Our approach depends on two key factors: 1) efficiently testing combinations of variables as causes, and 2) being able to directly compare the significance of causes. Thus, we build on the temporal-logic approach to causal inference of Kleinberg (2012), rather than on methods that find whole structures such as BNs or DBNs, as these methods have high computational complexity and cannot efficiently recalculate significance of a relationship after modifying a variable. We represent causal relationships as probabilistic computation tree logic (PCTL) formulas, or PCTLc for relationships involving both continuous and discrete variables (Kleinberg, 2011). This allows us to conveniently represent and test conjunctions. A causal relationship where c and m cause e in 10-20 minutes with probability 0.9 is represented as:

$$c \wedge m \leadsto^{\geq 10, \leq 20}_{\geq 0.9} e, \tag{1}$$

while raising the value of continuous variable e is:

$$c \wedge m \leadsto^{\geq 10, \leq 20}_{\geq 0.9} (e \geq E[e]). \tag{2}$$

Both cause and effect may be any PCTL or PCTLc formula. Then, causal significance is measured with the average difference a cause makes to the probability (or expected value)

of an effect, holding fixed each other possible cause of an effect. Formally, with a set of discrete variables and time series data, for potential cause c of effect e, where X is the set of possible causes of e (factors that raise the probability of e) causal significance is defined by:

$$\varepsilon_{avg}(c,e) = \frac{\sum_{x \in X} P(e|c \wedge x) - P(e|\neg c \wedge x)}{|X \backslash c|}. \tag{3}$$

For continuous variables, the probabilities above are replaced with conditional expectation (e.g., $E[e|c \wedge x]$), though the cause must be discretized. Holding fixed each $x \in X$ in turn, this finds the average difference the presence of the cause makes to the effect. Relationships where $\varepsilon_{avg}(c,e) > \varepsilon$ are considered ε-significant, and a threshold can be chosen using the p-value. For statistically significant causes to be provably causal, we must assume there are no hidden confounders, the data are faithful to the causal structure, and that relationships are stationary.

3.4 Formalizing and inferring moderators

Using our observations about the asymmetry of moderators and temporal logic formulas, we can formally define moderators and give algorithms for finding such relationships. Throughout the paper, we make the same key assumptions needed for learning causal structures: no hidden confounders, faithfulness (e.g., no canceling out of impact across multiple paths), and stationarity.

Definition 1 *A factor* m moderates *the causal relationship* $c \leadsto^{\geq r, \leq s}_{\geq p} e$ *when the following all hold:*

1. ***Effect is significantly modified*** *With* $c \wedge m \leadsto^{\geq r', \leq s'}_{\geq p'} e$ *and* $c \wedge \neg m \leadsto^{\geq r'', \leq s''}_{\geq p''} e$*, then at least one of the following must be true:*
   - $p' \neq p''$ *(probability modifier)*
   - $[r', s'] \neq [r'', s'']$ *(timing modifier)*
   - $E[e|c \wedge m] \neq E[e|c \wedge \neg m]$ *(intensity modifier).*
2. ***c is a cause with or without m*** $c \wedge m$*, and* $c \wedge \neg m$ *are significant causes of* e
3. ***m is not a cause alone*** $m \wedge \neg c$ *is an insignificant cause of* e

Rule 1 requires that there is a statistically significant difference – whether in timing (when the effect occurs), probability (how likely the effect will occur), or value (how a continuous effect changes) – between the outcome of c occurring in the presence and absence of moderator m. If there is no change regardless of the presence of m, then either it is not a moderator, or it is an improperly specified factor (e.g., the true moderator is a combination of necessary factors). When moderator m is binary, m and $\neg m$ are the occurrence and nonoccurrence of m respectively. When m has multiple states (e.g., glucose can be discretized as hypo-, hyper-, and euglycemia), m is a particular state (e.g., hypoglycemia) and $\neg m$ is all others (e.g., hyper- and euglycemia). Thus instead of a difference in outcome when m is present or absent, in this case we require a difference when changing from one of m's states to another. For probability and intensity differences, this means the conditional probability (or expected value) of e differs in the $c \wedge m$ and $c \wedge \neg m$ cases, even after holding

Algorithm 1 *discover_moderator*(T)

Input:
T, time series data for variables V
Output:
A set M storing all moderating relationships

```
Apply the causal inference method of Kleinberg (2012) to T to find significant relationships S
for each l (c ⇝^{≥r,≤s} e) in S do
  for each m ∈ V \ {c, e} do
    Test whether timing, intensity, or probability differ for c ∧ m ⇝ e and c ∧ ¬m ⇝ e
    If they differ significantly, calculate ε_avg for: c ∧ m, c ∧ ¬m, and m ∧ ¬c
    if only ε_avg(m ∧ ¬c, e) is insignificant, and the others are significant then
      add m moderates c ⇝^{≥r,≤s} e to M
return M
```

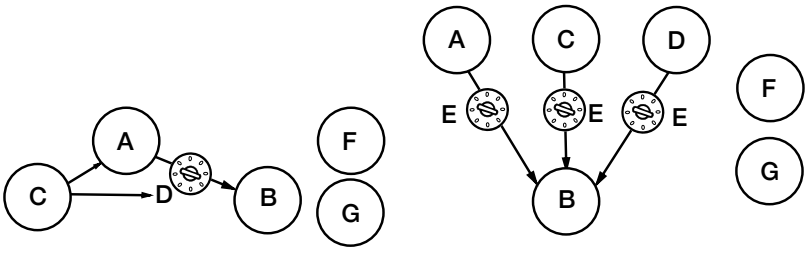
fixed all other factors as in eq. (3). For timing differences, the probabilities may be exactly the same, but $c \wedge m$ and $c \wedge \neg m$ are significant at different times.

Rule 2 requires that no matter the state of m, c is still a significant cause of e. This is to rule out complex causes, such as $c \wedge m$ being a cause of e (while $c \wedge \neg m$ is not a cause), or XOR type relationships. For example, in some biological cases there are backup mechanisms, such that $c \rightarrow e$, $c \rightarrow \neg d$, and $d \rightarrow e$. When c is not present, d will no longer be inhibited, and will cause e (rule 3 fails). Here d is not a moderator, and this rule ensures we will not identify it as one. We similarly rule out causal complexes, and these can be identified as cases where rule 1 and 3 hold but 2 does not.

Finally, we require that $m \wedge \neg c$ does not cause e. If m were an effective cause alone, then different behavior with c may be due to interaction rather than moderation. We do not require exact independence, only that $\varepsilon_{avg}(m \wedge \neg c, e)$ is less than a threshold (e.g., chosen by p-value), since exact independence is a very strict condition in realistic datasets.

Algorithm 1 highlights the key steps: 1) find significant causal relationships, and 2) refine relationships by finding moderators. We begin by identifying significant causal relationships ($c \rightsquigarrow^{\geq r, \leq s} e$, where $\varepsilon_{avg}(c, e) > \varepsilon$ for some significance threshold) from time series data using the method of Kleinberg (2012). For each significant relationship $c \rightsquigarrow^{\geq r, \leq s} e$, we test each variable $m \in V \backslash c$ as a possible moderator by evaluating $c \wedge m$, $c \wedge \neg m$, and $m \wedge \neg c$ (see Alg. 1). To compute the corresponding causal significance for each case, we now replace the proposition in eq. (3) with each of the conjunctions. The truth condition for $(c \wedge m)_t$ is logical AND: $c_t \wedge m_t$. For $\neg(c \wedge m)_t$, the truth condition is: $\neg c_t \vee \neg m_t$. Then, $c \wedge m$ is a significant cause for e if $\varepsilon_{avg}(c \wedge m, e) > \varepsilon$. Similarly, for testing other conjunctions $c \wedge \neg m$, and $m \wedge \neg c$, we use the same approach, replacing m with $\neg m$, and $\neg c$ with c to compute their corresponding causal significance ε_{avg}. Finally, we test each rule in def. 1 (change in effect, significance with and without moderator, insignificance of moderator alone). When all hold, the variable is a moderator.

See appendix A for proofs showing that all conditions hold when m is a moderator, and that the conditions do not hold in any other case.



(a) Common cause (b) Multiple moderation

Figure 2: Simulated difficult structures.

Table 1: Results of the proposed method on the common cause structure, with causal significance (Sig.) of cause (A) and moderator (D) before (top) and after (bottom) applying rules to discover moderators. * indicates statistical significance at all lags represented, and † significance of subset. Dashes indicate spurious time lags (no true results).

	Intensity		Probability		Timing	
Relationship	Sig. (true lags)	Sig. (other lags)	Sig. (true lags)	Sig. (other lags)	Sig. (true lags)	Sig. (other lags)
A → B	1.45*	.23	0.409*	-0.001	0.255*	0.000
D → B	–	.26†	–	0.047†	–	0.024
A∧D → B	2.22*	.04	0.510*	0.011	0.226*	0.001
A∧¬D → B	0.515*	-0.118	0.357*	-0.002	0.478*	0.000
¬A∧D → B	–	-.16	–	0.030	–	-0.005

3.5 Complexity

This procedure significantly reduces the computational complexity of finding moderators in a data-driven study. If we attempted to infer each as a complex cause (conjunction of factors), with only pairwise relationships between N variables there are N^3 possible cause-effect-moderator triples, and testing causal significance of all of these is $O(\text{causes}^2 \times \text{effects}) = O(N^4 \times N) = O(N^5)$.

Assuming that moderators will often be mistaken for causes, we only need to perform causal inference ($O(N^3T)$), then test our conditions for each pair of variables in the set of significant causes S of each effect, $O(S^2NT)$, where normally $S \ll N$. This assumption is reasonable because our primary motive is refining the set of relationships inferred to better understand complex causal structures and avoid confounding. When the assumption does not hold, though, complexity after the initial causal inference is $O(SN^2T)$, as we must test each of the N variables as a possible moderator of each of the significant causes S of each variable. However, as $S \ll N$, this is still feasible, and in the worst case, where every variable causes every other, is $O(N^3T)$, which is the same as the basic causal inference approach and still two orders of magnitude faster than testing all triples.

4. Experiments

We first evaluate our approach and compare it to regression (Baron and Kenny, 1986) and Structural Nested Mean Models (SNMM) (Almirall et al., 2010) on simulated data, demonstrating that ignoring moderation leads to confounding, and that we can more accurately

distinguish causes and moderators. Second, we apply our approach and SNMM to data from stroke patients, showing that we can identify more moderators and that these provide more information than a causal structure alone.

4.1 Simulated data

We begin with a description of the datasets created and our results, before examining results obtained with other approaches. We use two simulated structures (common cause, fig. 2a, and multiple moderators, fig. 2b), designed to test particularly difficult cases while allowing ground truth for evaluation. In both, we simulate 5000 timepoints, and use a threshold of $p < 0.01$ for determining causal significance. On average running time for our approach was 2.9min per experiment. The probability for all relationships is 0.9 unless otherwise specified. We simulate three types of moderators:

Intensity: These moderators change the expected value of the effect after the cause. For example, in both datasets, one true relationship is $A \leadsto^{2,3} B$. Without the occurrence of a moderator m, $E[B|A, \neg m] = 2$. When the cause and moderator occur, $E[B|A, m] = 5$. In the multiple moderation case, $m = E$, and in the common cause case $m = D$. All variables are discrete except for B, which is continuous. In the multiple moderation case, two additional relationships are moderated, $C \leadsto^{4,5} B$ and $D \leadsto^{5,6} B$, and their impact doubles from 3 to 6 and 4 to 8 respectively.

Probability: The relationships and time windows are as in the intensity case, but now the effect has probability 0.95 with the moderator, and 0.80 without.

Timing: Finally, these moderators change *when* the effect happens. For common cause, the time windows are: $A \wedge \neg D \leadsto^{1,2} B$, $C \leadsto^{3,4} A$ and $C \leadsto^{7,8} D$. When the cause and moderator occur: $A \wedge D \leadsto^{5,6} B$. For multiple moderators, windows are: $A \wedge \neg E \leadsto^{4,5} B$, $C \wedge \neg E \leadsto^{6,7} B$ and $D \wedge \neg E \leadsto^{7,8} B$, and $A \wedge E \leadsto^{1,2} B$, $C \wedge E \leadsto^{3,4} B$ and $D \wedge E \leadsto^{5,6} B$.

4.1.1 Common causes

First, we simulate a common cause of both the moderator and the effect (see fig. 2a). We add 4 other variables to the structure to act as noise. This is a challenging case, as the moderator's role may be confounded by the shared cause. Causes here have probability 0.2. The time window between cause and effect varied for each relationship and was designed to make the case more challenging (ensuring D happened temporally between A and B).

Table 1 has aggregated results. First, at the top is the result of applying the causal inference approach of Kleinberg (2012) without moderator analysis. A is correctly found significant at its actual time lags and insignificant at all others for all cases. For the intensity changing moderator, D is initially found significant at A's time lags. However, A's strength varies in the two scenarios (with and without D) (rule 1 holds in Def. 1) and both are significant (rule 2 holds in Def. 1), while D is never significant without A (rule 3 holds in Def. 1), allowing us to correctly classify A as a cause and D as a moderator of $A \leadsto B$. In the probability and timing change cases, results are similar, and we correctly find D makes A more likely to cause B and changes the time lags respectively. We can exclude D as a conjunct (part of a causal complex), since A is a significant cause without D and can exclude interacting relationships as D is insignificant when A is absent.

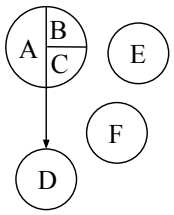


Figure 3: Causal complex

Table 2: Results of the proposed method on the multiple moderators structure, with significance shown for one relationship $A \rightsquigarrow B$ with moderator E. Markers as in table 1.

	Intensity		Probability		Timing	
Relationship	Sig. (true lags)	Sig. (other lags)	Sig. (true lags)	Sig. (other lags)	Sig. (true lags)	Sig. (other lags)
A → B	1.41*	0.09	0.289*	-0.007	0.201*	0.021
E → B	–	0.35†	–	0.019	–	0.036
A∧E → B	1.52*	.14	0.334*	-0.012	0.329*	0.028
A∧¬E → B	.93*	-0.01	0.239*	0.004	0.287*	0.026
¬A∧E → B	–	.30	–	0.025	–	0.008

After this result, which shows that we can correctly find that moderators are not simply causal complexes, we then simulated data that *did* have causal complexes (see fig. 3). Here the effect only occurs when all parts of the complex are present, as each alone is insufficient. In that data, none of the components of the causal complex were identified as moderators. Instead the only significant cause we identify ($p < .01$) is the combination of A, B, and C. This shows that in addition to successfully discovering moderators, we do not find additional moderators that we should not. This simulation tests the case when a moderator D and cause A share a common cause C (fig. 2a), we also test the case when a moderator and an effect share a common cause. See appendix B for details showing that we can still identify the moderator correctly.

4.1.2 Multiple moderation

The second case we simulate is when a single moderator moderates multiple causes of the same effect, as shown in fig. 2b. This type of situation is common in health related studies, where there may be multiple causes of a condition such as a heart attack, and each may be moderated by a demographic factor such as sex. This case is developed to challenge rule 3 of our definition, which is that the moderator has no effect alone. When it moderates multiple causes then even if one is absent, it may still appear significant. Once again we also include variables that act as noise, and are not involved in any causal relationships. The probability for noise variables is 0.1 and for causes is 0.15, and we test time lags [1,8].

Table 2 summarizes results for one of the causal relationships. As shown in the ground truth (fig. 2b), A causes B, moderated by E. With the moderator, the relationship is stronger. Before we apply the rules for finding moderators, we test pairwise relationships, and correctly find that A is a significant cause of E at all the true time lags, and not at any other lags. E at any time lag is a spurious cause of B, and overall we find it has low significance, though it was significant at one time lag, $t = 5$. Next, we test the three

combinations of A and E (with A and E alternately negated). Now we find that A is a significant cause of B in both scenarios (rule 2 holds in Def. 1), but that it is a much stronger cause with E than without E (rule 1 holds in Def. 1). This is the correct finding, and also provides more information on the relationship than the aggregated result found with A alone. Further, we now find that E alone is not a significant cause at any time lag (rule 3 holds in Def. 1) – correctly learning that its only role is moderating A's impact. For both the timing and probability change cases, the significance of the causes (A shown in the table) is correctly found to vary with the presence/absence of the moderator, while the moderator alone is insignificant.

4.1.3 Comparison with regression method

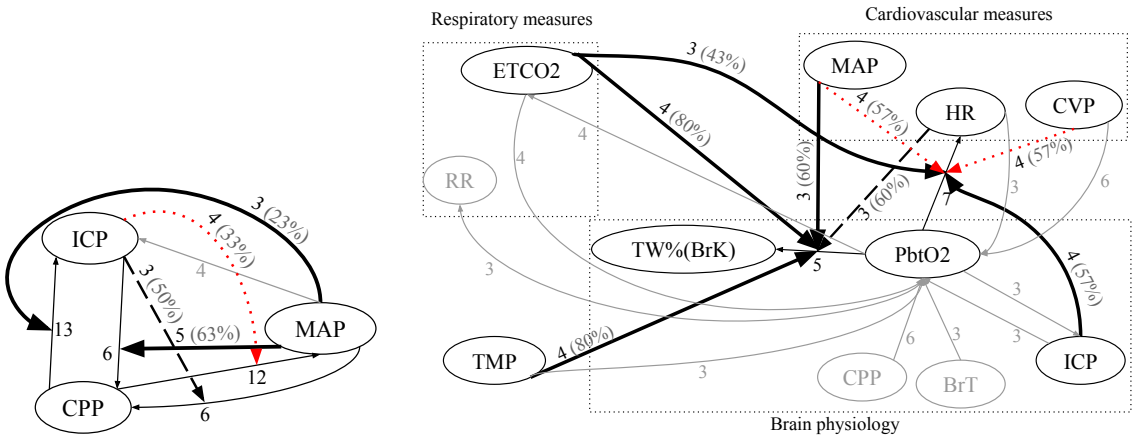
A common approach to moderator analysis is based on regression, testing the impact of different variables alone and together by determining whether their coefficients differ significantly from zero (Baron and Kenny, 1986). If X causes Y with moderator M, c in the following equation should be nonzero:

$$Y = i + aX + bM + cXM + \epsilon. \tag{4}$$

Using the data in the previous section, we test for moderation with: (1) linear regression, (2) logistic regression, and (3) vector auto regression (VAR), calling a variable a moderator when c differs significantly from zero. While c may be nonzero for other reasons such as interaction, we do not penalize algorithms for that. Even using a more generous threshold of $p < 0.05$, no value of c was significant for any case (common cause, multiple moderators) for any time lag or regression method. Looking at the individual values for c, not their significance, we see many false positives and negatives. For example, when moderator D changes the probability of $A \rightsquigarrow^{2,3} B$ of the common cause case in fig. 2a, the linear regression indicates $c > 0$ from time 2 to 7, and with VAR $c > 0$ from time lag 2 to 8. However, the true time window is $[2, 3]$. Logistic regression had no positive value for c at any time lag, for both datasets. Even though c is supposed to capture the change in effect when moderator is true, while a and b capture the individual impacts, we see that these are not being distinguished by the regression.

4.1.4 Comparison with Structural Nested Mean Models (SNMM)

Almirall et al. (2010) proposed a 2-Stage Regression Estimator of the Structural Nested Mean Model (denoted 2-SE here) to assess time-varying causal effect moderation. This method estimates the intermediate causal effect by conditioning on the tested moderators and estimating the difference in causal effect. We selected this approach for comparison as it outperforms Robins' G-estimator (Robins, 1994) when the parameters are correctly specified. When testing 2-SE, we use the true time lag of the simulated causal structure rather than requiring the approach to identify the lags. For the causal model in fig. 2a where D moderates relationship $A \rightarrow B$, we have datasets with three possible types of moderators (moderator D changes timing, probability, or value of relationship $A \rightarrow$ B). When D changes the probability of $A \rightarrow B$, the 2-SE finds D as a moderator with p-value 0.12, which is above our highest threshold of 0.05. Similarly, when D moderates the value of B, the 2-SE p-value for this relationship is 0.74. Finally, when D changes the timing



(a) Relationships among MAP, CPP, ICP (days 4-7)

(b) Relationships where PbtO2 is cause or effect (days 4-7)

Figure 4: Results on NICU data with moderators found in > 3 patients. Percent shows patients with the causal relationship that also had the moderator. Thick black edges are found only by our method, dashed by both, and dotted red only by 2-SE. Numbered edges without percentages are the number of patients in which that causal relationship was inferred.

of $A \rightarrow B$, the p-value according to 2-SE is 0.48. Thus in no case did 2-SE identify a moderator of the relationship with a p-value that would be accepted with even a more lenient threshold than we used for our own algorithm. For the structure in fig. 2b where E moderates multiple relationships, the lowest p-value output by 2-SE is 0.36 which is still above any commonly accepted significance threshold.

Thus, in comparison to 2-SE on simulated data, our proposed approach has much higher recall for identifying moderators, while 2-SE faces significant difficulties when a moderator moderates multiple relationships, as well as when it changes timing or intensity of a relationship. 2-SE did somewhat better with a probability changing moderator, but the significance score was still above the threshold for acceptance.

4.2 Neurological ICU data

Our main goal is to better understand complex medical data. To demonstrate how this approach can provide new insights into such data, we apply it to data from the Neurological Intensive Care Unit (NICU) at Columbia University. The data consists of 98 patients with subarachnoid hemorrhage (SAH) who underwent multimodality monitoring as part of routine care. Use of the data was IRB-approved. Prior work characterized the physiologic changes in these patients before and after seizures (Claassen et al., 2013) and the causal structure of their brain physiology (Claassen et al., 2016). Many expected relationships were inferred, but some were missing. In this work, we build on the causal structure inferred in that work, and aim to determine if we can identify clinically meaningful moderators of these relationships, and potentially get closer to the true causal structure.

Table 3: All NICU variables and their meaning.

NICU variables	Full name
TMP	body temperature
RR	respiratory rate
MV	minute ventilation
ETCO2 (CO2EX)	end tidal carbon dioxide
SPO2 (SPO2%)	oxygen saturation
HR	heart rate
MAP	mean arterial pressure
CVP	central venous pressure
CI	cardiac index
SVV	stroke volume variation
ELWI	extravascular lung water index
GEDI	global enddiastolic index
ICP	intracranial pressure
CPP	cerebral perfusion pressure
PbtO2	partial brain tissue oxygenation
rCBF	regional cerebral blood flow
TW% (BrK)	brain water content
BrT	brain temperature

4.2.1 Data and method

The data used in this study includes cardiovascular and respiratory parameters (e.g., heart rate), and brain physiology (e.g., brain oxygenation, blood flow in the brain, microdialysis measurement of brain metabolism). Table 3 shows the acronyms and full names of all variables. Not all patients have all variables measured, though, and monitors may be started at different times due to clinical practice. Further, data duration varies due to length of ICU stay (mean 12.3 days). Most variables are measured every 5 seconds, but some are roughly hourly (e.g., brain metabolism). We follow the approach of Claassen et al. (2016), where the data is broken into two clinically meaningful time periods (0-3, and 4-7 days post-SAH), is synchronized and minute-averaged, has missing values imputed (Rahman et al., 2015), and is discretized according to known physiologic ranges. We apply our approach for finding moderators to all causal relationships identified by Claassen et al. (2016), testing all variables as potential moderators of each. We applied each method to each patient's data individually, and depict only moderators found to be significant in at least 3 patients. For all methods, we use a p-value threshold of 0.05 for accepting a moderator. This threshold should be adjusted lower for applications where false positives have a higher cost, or adjusted upward for more exploratory analyses. To compare our approach fairly against 2-SE, we provide the method with the same inferred causal structures, and then test for moderators of the same set of relationships at the same time lags, [1,60]. In the figures, grey edges indicate relationships where no moderators were found. Thick black arrows into causal edges indicate moderators identified only by our approach, dashed black edges are moderators found by both approaches, and dotted red edges are moderators found only by 2-SE. Numbers on shared edges apply to our approach, and in general 2-SE either found the same number or fewer.

4.2.2 Results

We focus on further elucidating the key relationships that were identified by Claassen et al. (2016), testing the hypothesis that some of the expected causal relationships that were not identified may now appear as moderator effects. First, we examine whether we can now identify known relationships not found by Claassen et al. (2016). While CPP is defined as MAP−ICP, bidirectional causal relationships were not found between all pairs in that work due to the time lags. Results of our moderator analysis in figure 4a show that we now see moderating influences in some of these cases, such as MAP influencing ICP through CPP (as CPP depends on MAP). Note that our approach found a total of 3 moderators, while 2-SE found two. Next, the earlier work found that partial brain tissue oxygenation (PbtO2) depended mainly on cardiovascular parameters and measures of pressure in the brain. Interestingly, in addition to these plausible direct causal effects, we now find moderating effects including from ETCO2 (fig. 4b). ETCO2 (end tidal CO2) relates to cerebral blood flow by affecting vessel diameter and thereby plausibly affecting the relationship between heart rate and brain oxygenation. In contrast, 2-SE finds fewer moderators, and does not include the important moderators of PbtO2's effect on brain water content (TW%), a measure for brain swelling. Claassen et al. (2016) found TW% depended on fluid status during days 0-3 after stroke but during days 4-7 also depended on brain oxygenation. Unlike 2-SE we now find that during this later phase cardiovascular measures are important moderators of the effect of brain oxygenation on swelling. ETCO2, which may affect cerebral vascular diameter and thereby cerebral blood flow is in particular a biologically plausible moderator of brain swelling. Likewise, only our approach identified temperature (TMP), which affects both blood flow and brain metabolism as a moderator of the effect of PbtO2 on TW%. Figure 5 shows all relationships where TW% is either a cause or effect, along with moderators identified by our approach and by 2-SE.

Overall, for the 4-7 day time period, we identified 291 moderators, 125 of which were found by both methods, while 2-SE identified 197. The moderator effects identified are physiologically plausible, thereby supporting the hypothesis that identifying moderators further helps to unravel the complexity of physiologic interdependencies in a real world biological system. Further, our approach identified an overall larger number of moderators than 2-SE, as well as moderators that are more meaningful physiologically.

5. Conclusion

Causal relationships are often more complex than pairwise links between variables, with factors changing the intensity or timing of causes. With increasingly large datasets with many variables, providing insight into the exact nature of each variable's contribution to an effect is necessary for humans to make sense of the results. Methods that do not distinguish moderators from causes can lead to futile interventions on variables that are not causally efficacious. While moderators have been handled mainly via regression, there has not been a way to efficiently discover them in data-driven analyses. We have shown how rules for identifying moderators can be represented and efficiently tested via logical formulas – without increasing the complexity of causal inference or requiring tests of all combinations of variables. Our approach can be applied to many domains including health, finance, climate science, and biology to understand how causal relationships change in the presence

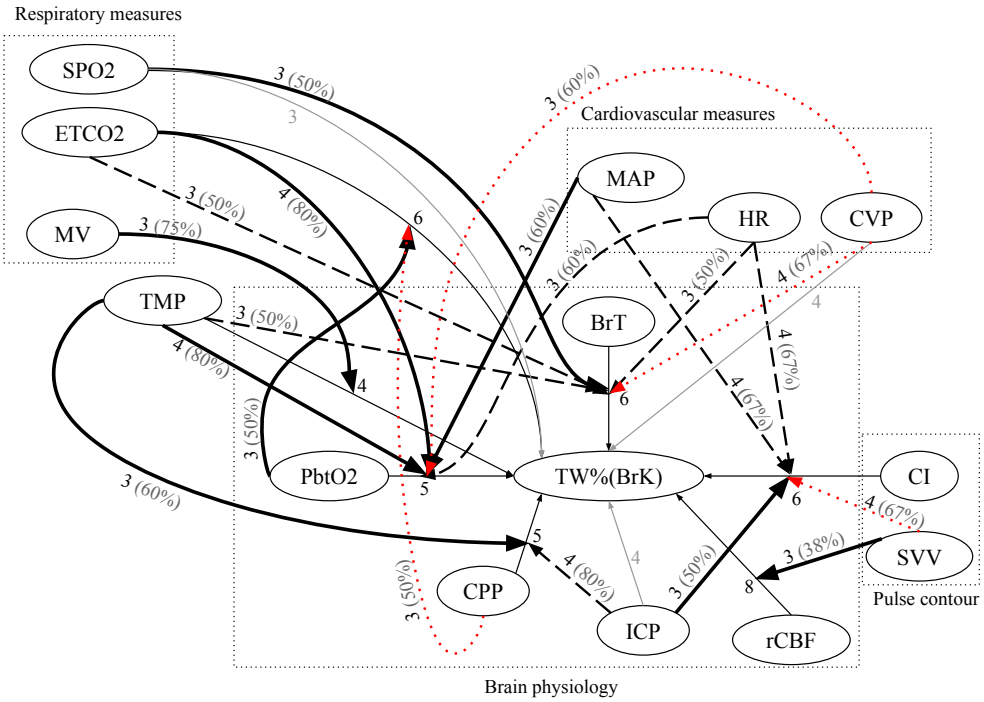


Figure 5: Relationships where TW%(BrK) is a cause or effect, for 4-7 day time period after stroke.

of other factors. In future work we aim to remove the assumption that there are no latent confounders, to enable wider applicability.

Acknowledgments

This work was supported in part by the James S. McDonnell Foundation, the NIH under award number R01LM011826, and NSF under award number 1347119.

Appendix A. Proof of correctness

Claim 1 *If m is a moderator for $c \rightsquigarrow^{\geq r,\leq s} e$, all rules in def. 1 will hold.*

Proof. For causal relationship $c \rightsquigarrow^{\geq r,\leq s} e$, based on equation 3, we have,

$$\varepsilon_{avg}(c \wedge m, e) = \frac{\sum_{x \in X} P(e|c \wedge m \wedge x) - P(e|\neg(c \wedge m) \wedge x)}{|X \backslash c|} \tag{5}$$

Similarly,

$$\varepsilon_{avg}(c \wedge \neg m, e) = \frac{\sum_{x \in X} P(e|c \wedge \neg m \wedge x) - P(e|\neg(c \wedge \neg m) \wedge x)}{|X \backslash c|} \quad (6)$$

For continuous data, replace the corresponding conditional probability in the above equations with conditional expectation (e.g., $E[e|c \wedge m \wedge x]$ and $E[e|c \wedge \neg m \wedge x]$).

Firstly, if m is a moderator, then the occurrence of m is not independent of c, which means $P(e|c \wedge m \wedge x) \neq P(e|c \wedge \neg m \wedge x)$ (or $E[e|c \wedge m \wedge x] \neq E[e|c \wedge \neg m \wedge x]$ for continuous valued data). Assuming that we only focus on variables that can strengthen or weaken a cause's impact or change when an effect occurs, then either $P(e|c \wedge m \wedge x)$ is significantly different with $P(e|c \wedge \neg m \wedge x)$ or the time window between $c \wedge m$ causing e and $c \wedge \neg m$ causing e is different (rule 1 holds).

Secondly, Given c is a significant cause of e across the time series ($\varepsilon_{avg}(c, e) > \varepsilon$ (statistical threshold), and $P(e|c \wedge x) \gg P(e|\neg c \wedge x)$), we focus on three types of moderators (moderating probability, timing or intensity):

1) If m changes the probability of $c \rightsquigarrow^{\geq r, \leq s} e$, then $P(e|c \wedge m \wedge x) \gg P(e|c \wedge x)$. Since $\varepsilon_{avg}(c, e) > \varepsilon$ (eq. 3), we can get $\varepsilon_{avg}(c \wedge m, e) > \varepsilon$ (eq. 5). Considering possible occurrence cases among x, m, c, we know that $\#(c \wedge x) = \#(c \wedge m \wedge x) + \#(c \wedge \neg m \wedge x)$ where $\#(c \wedge x)$ represents the number of cases when c and x occur together. Therefore, since c itself is a significant cause, unless m tends toward completely removing c's effect (approaching the causal complex case in section 3.1) $c \wedge \neg m$ will be a significant cause too ($\varepsilon_{avg}(c \wedge \neg m, e) > \varepsilon$). Thus, both $c \wedge m$, and $c \wedge \neg m$ are significant causes of e (rule 2 holds).

2) If m changes the timing of $c \rightsquigarrow^{\geq r, \leq s} e$, $c \wedge m$ and $c \wedge \neg m$ will cause effect e to occur at different time. Since c is a significant cause of e itself and m just changes the timing of the relationship, $P(e|c \wedge m \wedge x) \approx P(e|c \wedge \neg m \wedge x) \approx P(e|c \wedge x)$ at different timings. Because $\varepsilon_{avg}(c, e) > \varepsilon$, by comparing eq. 3, eq. 5 and eq. 6, we can get $\varepsilon_{avg}(c \wedge m, e) > \varepsilon$ and $\varepsilon_{avg}(c \wedge \neg m, e) > \varepsilon$. Therefore, rule 2 also holds.

3) If m changes the intensity of $c \rightsquigarrow^{\geq r, \leq s} e$, then $E(e|c \wedge m \wedge x) \gg E(e|c \wedge x)$. Given c is a significant cause of e, we have $E(e|c \wedge x) \gg E(e|\neg c \wedge x)$ (eq. 3 for continuous data), and $\varepsilon_{avg}(c, e) > \varepsilon$. Similar to case 1, we know that $\#(c \wedge x) = \#(c \wedge m \wedge x) + \#(c \wedge \neg m \wedge x)$. Thus, unless m tends toward completely removing c's effect (approaching the causal complex case in section 3.1) $c \wedge \neg m$ will be a significant cause too ($\varepsilon_{avg}(c \wedge \neg m, e) > \varepsilon$). Thus, both $c \wedge m$, and $c \wedge \neg m$ are significant causes of e (rule 2 holds).

Thirdly, if m is a moderator and not a cause, then $P(e|c \wedge m) \approx P(e|c)$ and $P(e|\neg c \wedge m) \approx P(e)$, so based on eq. (3), $\varepsilon_{avg}(\neg c \wedge m) < \varepsilon$. Thus, $\neg c \wedge m$ will be insignificant (rule 3 holds). Therefore, m is a moderator for $c \rightsquigarrow^{\geq r, \leq s} e$.

Claim 2 *If m is not a moderator for a true causal relationship $c \rightsquigarrow^{\geq r, \leq s} e$, but falls in one of the following cases, at least one rule in Def. 1 will fail:*
1. m ***is just noise****, rule 1 fails.*
2. m ***is a mediator****, rule 1 fails.*
3.m ***is a cause on its own****, rule 3 fails.*
4.m ***is part of a causal complex****, rule 2 fails.*

Proof. When m is unrelated to the effect (e.g., a noise variable), $P(e|c \wedge m \wedge x) \approx P(e|c \wedge \neg m \wedge x)$, according to eq. 5 and eq. 6, $\varepsilon_{avg}(c \wedge m, e) \approx \varepsilon_{avg}(c \wedge \neg m, e)$, therefore, e is not significantly changed when m occurs (rule 1 fails). If m is a mediator, similarly, $P(e|c \wedge m \wedge x) \approx P(e|c \wedge \neg m \wedge x)$ since the influence of c goes directly through m, thus rule 1 also fails. If m is simply another cause of e, then according to eq. 3, $P(e|\neg c \wedge m \wedge x) \gg P(e|\neg(\neg c \wedge m) \wedge x)$ and $\varepsilon_{avg}(\neg c \wedge m, e) > \varepsilon$, thus rule 3 fails. Finally, for causal complexes where $c \wedge m$ causes e but c and m are ineffective alone, then $P(e|c \wedge \neg m) \approx P(e)$ and $\varepsilon_{avg}(c \wedge \neg m, e) < \varepsilon$, thus, rule 2 fails.

Further, if m is a moderator that influences $c \rightsquigarrow^{\geq r, \leq s} e$ in different ways (e.g., changes both timing and probability), or a causal relationship has multiple moderators, we can still correctly find it. If m changes the timing and probability of $c \rightsquigarrow^{\geq r, \leq s} e$, by the definition of moderating, we will find the effect is significantly changed in both timing and probability (or intensity), as long as the timing case is tested first, and the new time windows are used in testing probability change. This is because if time windows change then the cause may seem insignificant in some cases using the original window. If a relationship has multiple moderators, based on Claim 1, we can find them by testing the timing case first.

Appendix B. Additional Simulated Experiments

Another challenging case is when there is a common cause of both the moderator and the effect. We tested this indirectly in the paper, but now show a more direct example (e.g. $C \rightarrow D$,$C \rightarrow B$, and D moderates $A \rightarrow B$). As shown in table 4 we once again find only the true causes, and the moderator is correctly identified.

Table 4: Common cause case, with significance of cause (A) and moderator (D) before (top) and after (bottom) applying rules to discover moderators. * indicates statistical significance at all lags represented, and † significance of subset. Dashes indicate spurious time lags (no true results).

	Intensity		Probability		Timing	
Relationship	Sig. (true lags)	Sig. (other lags)	Sig. (true lags)	Sig. (other lags)	Sig. (true lags)	Sig. (other lags)
A → B	1.301*	-0.027	0.387*	-0.006	0.349*	0.054
D → B	–	.095†	–	0.009	–	0.029†
A∧D → B	2.015*	-.0127	0.379*	-0.013	0.305*	-0.006
A∧¬D → B	0.771*	-0.059	0.341*	0.001	0.397*	0.009
¬A∧D → B	–	-.0136	–	-0.007	–	-0.001

References

D. Almirall, T. Ten Have, and S. A. Murphy. Structural Nested Mean Models for Assessing Time-Varying Effect Moderation. *Biometrics*, 66(1):131–139, 2010.

S. Athey and G. W. Imbens. Machine Learning Methods for Estimating Heterogeneous Causal Effects. *stat*, 1050:5, 2015.

R. M. Baron and D. A. Kenny. The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

A. E. Bauman, J. F. Sallis, D. A. Dzewaltowski, and N. Owen. Toward a Better Understanding of the Influences on Physical Activity: the Role of Determinants, Correlates, Causal Variables, Mediators, Moderators, and Confounders. *American journal of preventive medicine*, 23(2):5–14, 2002.

J. Claassen, A. Perotte, D. Albers, S. Kleinberg, J. M. Schmidt, B. Tu, N. Badjatia, H. Lantigua, L. J. Hirsch, S. A. Mayer, E. S. Connolly, and G. Hripcsak. Nonconvulsive seizures after subarachnoid hemorrhage: Multimodal detection and outcomes. *Annals of Neurology*, 74:53–64, 2013.

J. Claassen, S. A. Rahman, Y. Huang, H-P Frey, M. Schmidt, D. Albers, C. M. Falo, S. Park, S. Agarwal, E. S. Connolly, and S. Kleinberg. Causal structure of brain physiology after brain injury from subarachnoid hemorrhage. *PLoS ONE*, 11(4):1–18, 2016.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Nonparametric Tests for Treatment Effect Heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405, 2008.

M. Eichler and V. Didelez. Causal Reasoning in Graphical Time Series Models. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, 2007.

A. J. Fairchild and D. P. MacKinnon. A General Model for Testing Mediation and Moderation Effects. *Prevention Science*, 2009.

D. P. Green and H. L. Kern. Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees. *Public opinion quarterly*, 2012.

S. Kleinberg. A Logic for Causal Inference in Time Series with Discrete and Continuous Variables. In *IJCAI*, pages 943–950, 2011.

S. Kleinberg. *Causality, Probability, and Time*. Cambridge University Press, 2012.

H. C. Kraemer, G. T. Wilson, C. G. Fairburn, and W. S. Agras. Mediators and Moderators of Treatment Effects in Randomized Clinical Trials. *Archives of general psychiatry*, 59 (10):877–883, 2002.

E. Kummerfeld and J. Ramsey. Causal Clustering for 1-factor Measurement Models. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

M. J. Lee. Non-parametric Tests for Distributional Treatment Effect for Randomly Censored Responses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):243–264, 2009.

R. P. Monti and A. Hyvärinen. A Unified Probabilistic Model for Learning Latent Factors and Their Connectivities from High-Dimensional Data. *arXiv preprint arXiv:1805.09567*, 2018.

J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, 2000.

J. Pearl. Detecting Latent Heterogeneity. *Sociological Methods & Research*, 2015.

S. A. Rahman, Y. Huang, J. Claassen, N. Heintzman, and S. Kleinberg. Combining Fourier and Lagged k-Nearest Neighbor Imputation for Biomedical Time Series Data. *Journal of Biomedical Informatics*, 58:198–207, 2015.

J. M. Robins. Correcting for Non-Compliance in Randomized Trials Using Structural Nested Mean Models. *Communications in Statistics-Theory and methods*, 23(8):2379–2412, 1994.

J. M. Robins. Marginal Structural Models Versus Structural Nested Models as Tools for Causal Inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer, 2000.

R. Silva, R. Scheine, C. Glymour, and P. Spirtes. Learning the Structure of Linear Latent Variable Models. *Journal of Machine Learning Research*, 7:191–246, 2006. ISSN 1532-4435.

L. Song, M. Kolar, and E. P. Xing. Time-Varying Dynamic Bayesian Networks. In *NIPS*, pages 1732–1740, 2009.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

X. Su, J. Kang, J. Fan, R. A. Levine, and X. Yan. Facilitating Score and Causal Inference Trees for Large Observational Studies. *Journal of Machine Learning Research*, 13(Oct): 2955–2994, 2012.

T. J. VanderWeele. On the Distinction Between Interaction and Effect Modification. *Epidemiology*, 20(6):863–871, 2009.

T. J. VanderWeele and J. M. Robins. Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs. *Epidemiology*, 18(5):561–568, 2007.

S. Videla, C. Guziolowski, F. Eduati, S. Thiele, M. Gebser, J. Nicolas, J. Saez-Rodriguez, T. Schaub, and A. Siegel. Learning Boolean Logic Models of Signaling Networks with ASP. *Theoretical Computer Science*, 2015.

M. Voortman, D. Dash, and M. J. Druzdzel. Learning Why Things Change: The Difference-Based Causality Learner. In *UAI*, 2010.

S. Wager and S. Athey. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, 2017.

R. Wang and J. H. Ware. Detecting Moderator Effects Using Subgroup Analyses. *Prevention science*, 14(2):111–120, 2013.

R. J. Willke, Z. Zheng, P. Subedi, R. Althin, and C. D. Mullins. From Concepts, Theory, and Evidence of Heterogeneity of Treatment Effects to Methodological Approaches: a Primer. *BMC medical research methodology*, 12(1):185, 2012.