The Annals of Statistics
2016, Vol. 44, No. 6, 2433–2466
DOI: 10.1214/15-AOS1411

CAUSAL INFERENCE WITH A GRAPHICAL HIERARCHY OF INTERVENTIONS

BY ILYA SHPITSER[1] AND ERIC TCHETGEN TCHETGEN[2]

Johns Hopkins University and Harvard University

Identifying causal parameters from observational data is fraught with subtleties due to the issues of selection bias and confounding. In addition, more complex questions of interest, such as effects of treatment on the treated and mediated effects may not always be identified even in data where treatment assignment is known and under investigator control, or may be identified under one causal model but not another.

Increasingly complex effects of interest, coupled with a diversity of causal models in use resulted in a fragmented view of identification. This fragmentation makes it unnecessarily difficult to determine if a given parameter is identified (and in what model), and what assumptions must hold for this to be the case. This, in turn, complicates the development of estimation theory and sensitivity analysis procedures.

In this paper, we give a unifying view of a large class of causal effects of interest, including novel effects not previously considered, in terms of a hierarchy of interventions, and show that identification theory for this large class reduces to an identification theory of random variables under interventions from this hierarchy. Moreover, we show that one type of intervention in the hierarchy is naturally associated with queries identified under the Finest Fully Randomized Causally Interpretable Structure Tree Graph (FFRCISTG) model of Robins (via the extended g-formula), and another is naturally associated with queries identified under the Non-Parametric Structural Equation Model with Independent Errors (NPSEM-IE) of Pearl, via a more general functional we call the edge g-formula.

Our results motivate the study of estimation theory for the edge g-formula, since we show it arises both in mediation analysis, and in settings where treatment assignment has unobserved causes, such as models associated with Pearl's front-door criterion.

1. Introduction. The goal of the empirical sciences is discerning cause-effect relationships by experimentation and analysis. This is made difficult by the ubiquity of hidden variables, and the difficulty of collecting data free from confounding and selection bias. Two useful frameworks for addressing these difficulties have been potential outcomes, introduced by Neyman [8], and expanded by Rubin [21],

Received November 2014; revised October 2015.
[1]Supported in part by NIH Grants R01 AI104459-01A1.
[2]Supported in part by NIH Grants ES020337 and AI104459.
MSC2010 subject classifications. 62H05, 62H99.
Key words and phrases. Causal inference, graphical models, mediation analysis, identification.

and causal graphical models, first used in linear models by Wright [35], and later expanded into a general framework (see, e.g., [30] and [11]). There exists a modern synthesis of these two frameworks, where causal models based on nonparametric structural equations are defined on potential outcome random variables, and assumptions defining these models can be represented by (absences) of arrows in a graph. See [11], Chapter 7, and [13] for a detailed treatment.

Potential outcome random variables represent outcomes under a hypothetical *intervention* operation, which corresponds to an idealized randomized control trial. Concepts such as the overall causal effect of a treatment can be represented as causal parameters on appropriate potential outcomes, and as statistical estimands if appropriate assumptions hold.

The synthesis of potential outcomes and graphs has been instrumental in much of the recent work on identification of various types of causal parameters such as total effects [14, 25–27, 33], and mediated effects [1, 10, 24].

Nevertheless, the existing literature suffers from three problems. First, a single graph may correspond to different causal models, which means a particular causal parameter may be identified under one causal model, but not under another, even though the models *share the same graph*. Second, different types of causal parameters seem to have different key issues underlying their identification, which makes it difficult to determine the specific assumptions that must hold for identification. For instance, certain types of unobserved confounding must be absent in order for overall effects to be identifiable, while even completely unconfounded mediated effects may be unidentified [1]. Finally, because of the complex nature of identification theory for causal parameters, existing conventional wisdom on what is identifiable is *too conservative*. For example, it is often assumed that a mediator and outcome must remain completely unconfounded in order to obtain identification of mediated causal effects. However, this is not true [24].

These issues make it difficult to determine *if* a particular causal parameter is identified, and *under what model*, what *assumptions* underlie this identification, and what the corresponding statistical parameter is. This complicates estimation theory, the development of parametric relaxations that permit identification and sensitivity analysis procedures.

1.1. *Outline of the paper.* The contents of the paper can be summarized by a picture in Figure 1. In Section 2, we introduce our notation, necessary graph theory, standard interventions (which we call node interventions in this manuscript) and potential outcomes, which are responses to node interventions. We also introduce the FFRCISTG model of Robins, which in this paper we call the "single world model (SWM)," and the NPSEM-IE of Pearl, which is a submodel of the FFRCISTG model, and which we call the "multiple worlds model (MWM)." The reasons for these names will become clear when these models are defined. The subset relationship of these two models is shown explicitly in Figure 1. Finally, we discuss targets of interest in causal inference known as total effects, which are

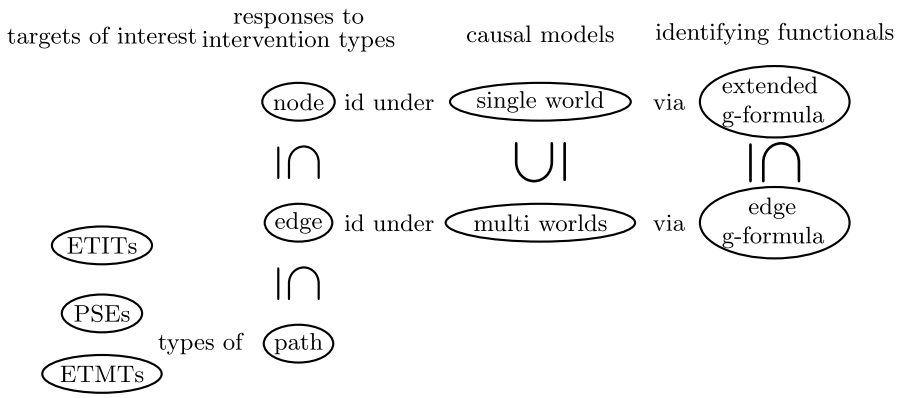

FIG. 1. *A hierarchy of responses to interventions defined with respect to features of a causal graph, the relationship of this hierarchy to targets of interest in causal inference, such as path-specific effects (PSEs), effects of treatments on the multiply treated (ETMTs), and new targets such as effects of treatments on the indirectly treated (ETITs) and identifiability under causal models defined in the literature.*

defined in terms of node interventions, and discuss identification theory for these targets under the SWM via the extended g-formula.

In Section 3, we define additional types of interventions, that we term edge and path interventions, and responses to these types of interventions via recursive substitution. Responses to node, edge and path interventions form an inclusion hierarchy in the sense that responses to node interventions are a special case of responses to edge interventions, which are in turn a special case of responses to path interventions. This inclusion is denoted by the subset relations in Figure 1. We also discuss how targets of inference in mediation analysis known as direct and indirect effects are defined in terms of edge interventions.

In Section 4, we show how we can express a wide variety of targets of interest in causal inference, such as path-specific effects (PSEs) or effects of treatment on the multiply treated (ETMTs) as responses to path interventions. In addition, we show that path interventions are general enough to accommodate novel targets which combine features of PSEs and ETMTs, which we call effects of treatment on the indirectly treated (ETITs). Our results then imply novel identification results for these targets, and others not previously considered in the literature, but expressible as path interventions.

In Section 5, we show that there is a natural correspondence between causal models and intervention types we discuss in the following sense. We show that responses to node interventions are identified under the SWM, and responses to edge interventions are identified under the MWM. Furthermore, we show that if a response to an edge intervention cannot be expressed as a node intervention, then it is *not* identified under the SWM, and if a response to a path intervention cannot be expressed as an edge intervention, then it is *not* identified under the MWM.

The identification of node interventions under the SWM is via the well-known *extended g-formula* [13, 18], which we give as equation (2). The identification of edge interventions under the MWM is via a generalization of (2), which we call the *edge g-formula*, and give as equation (5).

We also give examples of targets of interest in causal inference that do not correspond to responses to path interventions, as well as an example of a submodel of the MWM where even path interventions not ordinarily identified under the MWM are identified.

In Section 6, we briefly discuss the relationship of our results to Single World Intervention Graphs (SWIGs) [13].

Section 7 shows that a certain class of functionals that identify causal effects in latent variable causal models [25, 33] corresponds to functionals derived from the edge g-formula. This implies, in particular, that functionals that arise for treatment effects with unobserved causes of treatments, such as the front-door functional, also arise in mediation analysis.

In Section 8, we illustrate the connection of our work to existing estimation theory for causal parameters, and suggest avenues of future work, by giving a known example of an estimator for a parameter derived from a special case of the edge g-formula.

What the overall picture implies is that once we solve the identification problem for the responses to interventions in our hierarchy, as we do here, we immediately reduce the identification problem for a wide class of targets of interest to the much easier problem of translating those targets into responses to path interventions. Once that translation is complete, the question of what is identified under what model is immediately settled. In addition, our developments imply that estimation theory for functionals derived from the edge g-formula is relevant for a large class of inference targets identified under the MWM, including path-specific effects, effects of treatment on the multiply treated, and certain total causal effects with unobserved causes of treatments.

In the interests of space, the vast majority of arguments for our results appear in the Appendices in the supplementary materials [29]. In addition, the supplementary materials contains our rationale for the use of path interventions, rather than simpler or more algebraic representations of causal inference targets.

2. Notation and definitions.

We introduce graph theory terms, potential outcomes and statistical and causal graphical models.

2.1. *Graphs and random variables*.

We will associate random variables with vertices in graphs. We will denote *both* a single vertex and a single corresponding random variable as an uppercase Roman letter, for example, A. Sets of vertices (and corresponding random variables) will be denoted by uppercase bold letters, for example, $\mathbf{A}$.

For a random variable V, let $\mathfrak{X}_V$ be the state space of V. For example, if V is binary, then $\mathfrak{X}_V = \{0, 1\}$. We denote elements of a set $\mathfrak{X}_A$ (values of A) by lowercase Roman letters: $a \in \mathfrak{X}_A$. The state space of a set $\mathbf{V}$ of random variables is simply the Cartesian product of the individual state spaces: $\mathfrak{X}_{\mathbf{V}} = \times_{V \in \mathbf{V}}(\mathfrak{X}_V)$.

Sets of values corresponding to sets of random variables will be denoted by lowercase bold letters, for example, $\mathbf{a} \in \mathfrak{X}_{\mathbf{A}}$. Sometimes we will denote a restriction of a set of values by a set subscript. That is, if $\mathbf{v}$ is a set of values of $\mathbf{V}$, and $\mathbf{A} \subseteq \mathbf{V}$, then $\mathbf{v}_{\mathbf{A}}$ is a restriction of $\mathbf{v}$ to $\mathbf{A}$.

An edge in a graph is a vertex adjacency coupled with an orientation. A path in a directed graph is a (possibly empty) sequence of nodes of the form $(A_1 A_2 A_3 \cdots A_{k-1} A_k)$, where each node in the sequence occurs exactly once, and each A_i, A_{i+1} share an edge. The first vertex in a path sequence is called the source, and the last vertex is called the sink. A path with two vertices $(A_1 A_2)$ is just an edge.

A subpath of a path is a subsequence of edges in a path that themselves form a path. A suffix subpath of $(A_1 A_2 \cdots A_{m-1} A_m \cdots A_{k-1} A_k)$ is a subpath of the form $(A_{m-1} A_m \cdots A_{k-1} A_k)$, while a prefix subpath is a subpath of the form $(A_1 A_2 \cdots A_{m-1} A_m)$. A directed path from A_1 to A_k has edges for every i of the form $A_i \to A_{i+1}$. We will denote a directed path as $(A_1 A_2 \cdots A_k)_{\to}$, and also by Greek letters, for example, α, and sets of directed paths by bold Greek letters, for example, $\boldsymbol{\alpha}$. A source vertex of α will be written $\mathrm{so}_{\mathcal{G}}(\alpha)$, and the sink vertex will be written $\mathrm{sink}_{\mathcal{G}}(\alpha)$.

We say a directed cycle exists in a graph if it contains a path $(A_1 A_2 A_3 \cdots A_k)_{\to}$ and an edge $(A_k A_1)_{\to}$. A directed graph lacking directed cycles is called acyclic, abbreviated as DAG.

2.2. *Causal models of a DAG.* For a subset $\mathbf{A}$ of random variables $\mathbf{V}$, and a value assignment $\mathbf{a}$ to $\mathbf{A}$, we denote a forced assignment of $\mathbf{A}$ to an element of $\mathfrak{X}_{\mathbf{A}}$ as a *node intervention*. A node intervention which maps $\mathbf{A}$ to $\mathbf{a} \in \mathfrak{X}_{\mathbf{A}}$ will be denoted by $\nu_{\mathbf{a}}$. Pearl denoted node interventions $\nu_{\mathbf{a}}$ by $\mathrm{do}(\mathbf{a})$, and Robins by $g = \mathbf{a}$. We use alternative notation in this paper to avoid ambiguity, because we will consider other types of interventions. It is also possible to consider more complex types of interventions on nodes, known as *dynamic treatment regimes*, where assigned values to $\mathbf{A}$ are not constants, but functions of variables assigned and observed in the past [6, 7, 14]. Although generalizations of our results to this setting are possible, we do not pursue them in the interests of space.

For a random variable $Y \in \mathbf{V}$, and $\mathbf{a} \in \mathfrak{X}_{\mathbf{A}}$ for a set $\mathbf{A} \subseteq \mathbf{V}$, we denote a (random) response to a node intervention $\nu_{\mathbf{a}}$ as $Y(\mathbf{a})$. These random variables are also called *potential outcomes*, because Y is often an outcome of interest, and the intervention is often hypothetical, rather than actually occurring. Given a set $\mathbf{Y} = \{Y_1, \ldots, Y_k\}$ of random variables, we denote $\{Y_1(\mathbf{a}), \ldots, Y_k(\mathbf{a})\}$ by $\mathbf{Y}(\mathbf{a})$ or $\{\mathbf{Y}\}(\mathbf{a})$.

Let $\mathrm{pa}_{\mathcal{G}}(V)$ be the set of parents of V in $\mathcal{G}$, that is, the set $\{W | (WV)_{\to}$ is in $\mathcal{G}\}$. Following [13], given a DAG $\mathcal{G}$ with vertices $\mathbf{V}$, we will assume the existence of $V(\mathbf{v}_{\mathrm{pa}_{\mathcal{G}}(V)})$ for every $V \in \mathbf{V}$, and for all $\mathbf{v}_{\mathrm{pa}_{\mathcal{G}}(V)} \in \mathfrak{X}_{\mathrm{pa}_{\mathcal{G}}(V)}$, as well as a

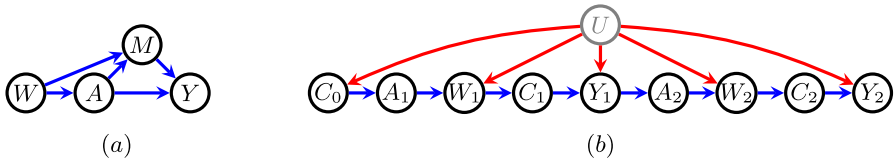

FIG. 2. (a) *A simple causal graph*. (b) *The transitive closure with respect to blue arrows of this graph is a causal graph representing two time slices of a longitudinal study in HIV research*.

well-defined joint distribution over these random variables, and use these potential outcomes, and the associated joint, to define others using recursive substitution.

In particular, for any $\mathbf{A} \subseteq \mathbf{V}$, and any $\mathbf{a} \in \mathfrak{X}_{\mathbf{A}}$, we define for every $V \in \mathbf{V}$

$$V(\mathbf{a}) \equiv V\big(\mathbf{a}_{\mathrm{pa}_{\mathcal{G}}(V)}, \{\mathrm{pa}_{\mathcal{G}}(V) \setminus \mathbf{A}\}(\mathbf{a})\big). \tag{1}$$

In words, this states that the response of V to $\nu_{\mathbf{a}}$ is defined as the potential outcome where all parents of V which are in $\mathbf{A}$ are assigned an appropriate value from $\mathbf{a}$, and all other parents are assigned whatever value they would have attained under a node intervention $\nu_{\mathbf{a}}$ (these are defined recursively, and the definition terminates because of the lack of directed cycles in $\mathcal{G}$). For example, in the graph in Figure 2(a), $Y(a) = Y(a, M(a))$.

It is possible to construct additional types of potential outcomes other than those that are responses to node interventions. We will discuss some such potential outcomes later. However, responses to node interventions are sufficient to define causal models. Just as a statistical model is a set of distributions over $\mathbf{V}$ defined by some restriction, we view a causal model as a set of distributions over $\{V(\mathbf{v}_{\mathrm{pa}_{\mathcal{G}}(V)}) | V \in \mathbf{V}\}$ defined by some restriction. We will call elements of a causal model *causal structures*, and denote them as $c(\mathbf{V}, \mathcal{G})$, by analogy with $p(\mathbf{V})$, but indexed by a graph. In this paper, we will consider two causal models.

We adopt the definitions presented in [13]. We define the *finest fully randomized causally interpretable structured tree graph* (*FFRCISTG*) *model* associated with a DAG $\mathcal{G}$ with vertices $\mathbf{V}$, as the set of all possible potential outcome responses subject to the restriction that the variables in the set

$$\{V(\mathbf{v}_{\mathrm{pa}_{\mathcal{G}}(V)}) | V \in \mathbf{V}\}$$

are mutually independent for every $\mathbf{v} \in \mathfrak{X}_{\mathbf{V}}$. We define the *nonparametric structural equation model with independent errors* (*NPSEM-IE*) associated with a DAG $\mathcal{G}$ with vertices $\mathbf{V}$, as the set of all possible potential outcome responses subject to the restriction that the sets of variables

$$\{\{V(\mathbf{a}_V) | \mathbf{a}_V \in \mathfrak{X}_{\mathrm{pa}_{\mathcal{G}}(V)}\} | V \in \mathbf{V}\}$$

are mutually independent. The NPSEM-IE associated with a particular graph is a submodel of the FFRCISTG model associated with the same graph, because it

always places at least as many restrictions on potential outcome responses, and in most cases many more.

For example, the binary FFRCISTG model associated with the DAG in Figure 2(a) asserts that variables W, $A(w)$, $M(a, w)$, $Y(a, m)$ are mutually independent for any $a, m, w \in \{0, 1\}$, while the binary NPSEM-IE model associated with the same DAG asserts that sets $\{W\}$, $\{A(w)|w \in \{0, 1\}\}$, $\{M(a, w)|a \in \{0, 1\}, w \in \{0, 1\}\}$, $\{Y(a, m)|a \in \{0, 1\}, m \in \{0, 1\}\}$ are mutually independent. The FFRCISTG model always imposes restrictions on a set of variables under a single set of interventions (a "single world"), while the NPSEM-IE may also impose restrictions on variables across multiple conflicting sets of interventions simultaneously. To emphasize this, we will refer to the FFRCISTG model as a "single world model" (SWM), and to the NPSEM-IE as a "multiple worlds model" (MWM) in the remainder of this paper.

A crucial difference between the SWM and the MWM is that the assumptions of the former are possible to test, at least in principle, by checking independences in a distribution of responses in an idealized randomized controlled trial. That is, if we wanted to check if W is independent of $A(w)$, we could check independence in a joint distribution obtained from recording, for a set of units, the values of W immediately before treatment w is assigned and the response values of A under that assignment. However, checking if $M(a)$ is independent of $Y(a', m)$ would entail somehow knowing how the response M of a unit behaves under assigned treatment a, and *simultaneously* how the response Y of the unit behaves under a conflicting treatment a' (and m). One may be able to argue for explicit construction of such joint responses in certain designs [5], or for certain types of units, for instance logic gates in a digital circuit. However, in general, assumptions defining the MWM are not experimentally testable.

2.3. *Identification of node interventions*. Responses to interventions of various types can be used to define targets of interest, discussed in more detail in Section 4. However, in order for these definitions to be useful, they must be linked to actually observed data. If such a link can be provided, that is, if a particular response can be expressed as a functional of the observed joint distribution $p(\mathbf{V})$ for any element of a causal model, we say that the response is *identified* under that causal model from $p(\mathbf{V})$.

In causal models, this link is typically provided via the *consistency assumption*, which is sometimes informally stated as "in the subpopulation where $\mathbf{A} = \mathbf{a}$, $\mathbf{Y}(\mathbf{a})$ behaves as $\mathbf{Y}$." Under the definition of the SWM (and the MWM), consistency is implied by (1); see [13], page 21. Thus, consistency is "folded in" to the model definition. Thus, we will describe identification in terms of a particular model, and not mention consistency itself. Note that (1) is an assumption defined using a particular graph. If we are mistaken about the true graph, for instance, due to the presence of unaccounted hidden variables, then some parts of (1), and thus some

parts of the consistency assumption, may not be justifiable under the true causal model.

Identification theory for node interventions in causal DAG models is well understood. Given a DAG $\mathcal{G}$ with vertices $\mathbf{V}$, and two arbitrary subsets $\mathbf{A}, \mathbf{Y}$ of $\mathbf{V}$ (not necessarily disjoint), the distribution $p(\mathbf{Y}(\mathbf{a}))$ for any value assignment $\mathbf{a} \in \mathfrak{X}_{\mathbf{A}}$ can be identified under the SWM as a functional of the observed distribution $p(\mathbf{V})$ using the *extended g-formula* [18], given by

$$p\big(\mathbf{Y}(\mathbf{a}) = \mathbf{v}_{\mathbf{Y}}\big) = \sum_{\mathbf{v}_{\mathbf{V}\setminus\mathbf{Y}}} \prod_{V \in \mathbf{V}} p\big(\mathbf{v}_V | \mathbf{a}_{\mathrm{pa}_{\mathcal{G}}(V)\cap\mathbf{A}}, \mathbf{v}_{\mathrm{pa}_{\mathcal{G}}(V)\setminus\mathbf{A}}\big), \tag{2}$$

where $\mathbf{v} \in \mathfrak{X}_{\mathbf{V}}$. A recent proof of this appears in [13]. Special cases of (2) where $\mathbf{A}$ and $\mathbf{Y}$ are disjoint are known as the *g-formula* [14], the *manipulated distribution* [30] or the *truncated factorization* [11]. Because the MWM is a causal submodel of the SWM, (2) also holds under the MWM.

2.4. *Total effects as responses to node interventions.* Node interventions are used to represent causal effects of treatments as a contrast of potential outcome responses to different treatment assignments. By considering an intervention, we remove the impact of confounding via assignment policy. For example, consider the simple causal graph shown in Figure 2(a), representing an observational study with a single application of one of two treatments m, m'. Variable M is assigned to either m or m' based on (observed) patient health status (A, W), and survival Y is measured. Doctors follow a known policy $p(M|A, W)$ in assigning M where sicker patients are more likely to get m. Note that $p(\text{alive}|m) < p(\text{alive}|m')$ may hold simply due to the assignment policy in the study which introduces confounding by health status, even if m is a better drug.

One appropriate contrast that adjusts for the influence of confounding by health status on the effect of interest can be expressed via node interventions, and is known as the average causal effect (ACE): $\mathbb{E}[Y(m)] - \mathbb{E}[Y(m')]$. This contrast can be computed from the distribution $p(Y(m))$ for all $m \in \mathfrak{X}_M$, which is equal, under (2), to

$$p\big(Y(m)\big) = \sum_{w,a,m'} p(Y|m, a, w)p\big(m'|a, w\big)p(a, w) = \sum_{w,a} p(Y|m, a, w)p(a, w).$$

This recovers the well-known back-door formula [11].

Consider now a more complex example corresponding to the following problem from HIV research. In a longitudinal study, HIV patients were put on an antiretroviral drug regimen, where the specific level of drug exposure over time was controlled by a known policy, which was based on covariates observed for each patient. However, the outcome of the study has been disappointing. The question is whether this was due to the drug itself performing poorly, or whether patient's adherence was poor. Consider a causal graph representing two time slices of this

longitudinal study. To avoid cluttering the figure with too many edges, we represent the causal graph schematically by its transitive reduction with respect to blue edges, shown in Figure 2(b). That is, the true graph $\mathcal{G}^*$ contains a blue arrow between any pair of nodes A, B connected by a blue directed path in Figure 2(b) (and inherits all red edges as well).

Here, C_0 is a vector of observed baseline confounders, A_1, A_2 are exposures over time, W_1, W_2 are drug toxicity levels at each exposure time, C_1, C_2 are adherence levels at each time, Y_1, Y_2 are outcomes, and U is an unobserved confounder. Both red and blue arrows represent direct causation. In general, a reasonable causal graph will contain unobserved common causes of most vertices, but in this example we assume adherence C_1, C_2, and treatments A_1, A_2 are only *directly* affected by the observed variables in the past, such as the toxicity level of the drug, and not by U. These assumptions are represented graphically by the absence of red edges from U to A_1, A_2, C_1, C_2.

We first consider the total effect of the two exposures on outcome Y_2, formalized as the two-exposure version of ACE. We consider more complex effects involving mediation by adherence in subsequent sections. The ACE contrast is defined with respect to active treatment levels, which we denote a_1, a_2, and baseline treatment levels, which we denote a_1', a_2'. In our case, the contrast is equal to $\text{ACE} \equiv \mathbb{E}[Y_2(a_1, a_2)] - \mathbb{E}[Y_2(a_1', a_2')]$. If we were able to randomize treatment assignment to A_1, A_2, we could evaluate the ACE directly from experimental data. However, our data comes from an observational longitudinal study and, therefore, we must properly adjust for observed confounders of the exposures. Robins [14] noted that in cases like these, assuming the underlying SWM represented by our graph is correct, we can get a bias-free estimand of the ACE from observational data using the *g-computation algorithm*, which in this case gives

$$\begin{aligned}\text{ACE} = &\sum_{y_1,c_1,w_1,c_0} \mathbb{E}[Y_2|a_2, y_1, c_1, w_1, a_1, c_0]p(y_1, c_1, w_1|a_1, c_0)p(c_0)\\ &- \sum_{y_1,c_1,w_1,c_0} \mathbb{E}\big[Y_2|a_2', y_1, c_1, w_1, a_1', c_0\big]p\big(y_1, c_1, w_1|a_1', c_0\big)p(c_0).\end{aligned}$$

This is, yet again, a special case of (2). This estimand can be estimated via either the parametric g-formula [15], inverse weighting methods [17] or doubly robust methods [20].

In the following section, we introduce intervention types that generalize node interventions, and consider other types of causal effects which may be represented as responses to such intervention types.

3. Edge and path interventions. We consider two additional types of interventions defined on graphical features, edge and path interventions, and define responses to these interventions using recursive substitution in a natural way. As we shall see, responses to path interventions include many targets of interest in causal inference, including effects of treatment on the treated, mediated effects, and even novel effects that combine features of both.

3.1. *Edge interventions.* For a set of edges $\boldsymbol{\alpha}$ in a DAG $\mathcal{G}$, define $\mathfrak{X}_{\boldsymbol{\alpha}} \equiv \mathfrak{X}_{\mathrm{so}_{\mathcal{G}}(\boldsymbol{\alpha})}$. In other words, $\mathfrak{X}_{\boldsymbol{\alpha}}$ is a Cartesian product of the state spaces of source variables of all directed edges in $\boldsymbol{\alpha}$.

The state space of a given vertex in $\mathcal{G}$ may occur *multiple times* in $\mathfrak{X}_{\boldsymbol{\alpha}}$ if multiple edges in $\boldsymbol{\alpha}$ share the same source vertex. We denote members of $\mathfrak{X}_{\boldsymbol{\alpha}}$ by lowercase Frankfurt font: $\mathfrak{a} \in \mathfrak{X}_{\boldsymbol{\alpha}}$. We do so to emphasize that elements of $\mathfrak{X}_{\boldsymbol{\alpha}}$ may contain multiple *conflicting* value assignments to the same random variable, unlike elements of $\mathfrak{X}_{\mathbf{A}}$. For example, consider the graph in Figure 2(a), where $\mathfrak{X}_A = \{0, 1\}$. Then if $\boldsymbol{\alpha} = \{(AM)_{\rightarrow}, (AY)_{\rightarrow}\}$, a valid element $\mathfrak{a}$ of $\mathfrak{X}_{\boldsymbol{\alpha}}$ associates 0 with the variable associated with the parent vertex A of $(AM)_{\rightarrow}$ and 1 with the variable associated with the parent vertex A of $(AY)_{\rightarrow}$. Unlike elements of $\mathfrak{X}_{\mathbf{A}}$, it is not immediately clear what set of edges $\mathfrak{a}$ is referring to, so we will subscript the set of edges, if necessary, like so: $\mathfrak{a}_{\boldsymbol{\alpha}}$.

We call a forced assignment of variables corresponding to source vertices of edges from $\boldsymbol{\alpha}$ to an element of $\mathfrak{X}_{\boldsymbol{\alpha}}$ an *edge intervention*. An edge intervention which assigns $\boldsymbol{\alpha}$ to an element $\mathfrak{a}_{\boldsymbol{\alpha}} \in \mathfrak{X}_{\boldsymbol{\alpha}}$ will be denoted by $\eta_{\mathfrak{a}_{\boldsymbol{\alpha}}}$. As with elements of $\mathfrak{X}_{\mathbf{A}}$, we denote a restriction of $\mathfrak{a}$ by a set subscript. That is, if $\mathfrak{a}_{\boldsymbol{\alpha}} \in \mathfrak{X}_{\boldsymbol{\alpha}}$, and $\boldsymbol{\beta} \subseteq \boldsymbol{\alpha}$, then $\mathfrak{a}_{\boldsymbol{\beta}}$ is a restriction of $\mathfrak{a}$ to variables corresponding to source vertices of $\boldsymbol{\beta}$.

We define responses of outcomes to edge interventions in the natural way using recursive substitution, the potential outcomes of the form $V(\mathbf{v}_{\mathrm{pa}_{\mathcal{G}}(V)})$, and a joint distribution over these potential outcomes. For every $V \in \mathbf{V}$, a set of edges $\boldsymbol{\alpha}$ in a DAG $\mathcal{G}$, and an element $\mathfrak{a}_{\boldsymbol{\alpha}} \in \mathfrak{X}_{\boldsymbol{\alpha}}$, we define the response of V to $\eta_{\mathfrak{a}_{\boldsymbol{\alpha}}}$ as

$$V(\mathfrak{a}_{\boldsymbol{\alpha}}) \equiv V\big(\mathfrak{a}_{\{(*V)_{\rightarrow} \in \boldsymbol{\alpha}\}}, \{\mathrm{pa}_{\mathcal{G}}^{\overline{\boldsymbol{\alpha}}}(V)\}(\mathfrak{a}_{\boldsymbol{\alpha}})\big), \tag{3}$$

where $\mathrm{pa}_{\mathcal{G}}^{\overline{\boldsymbol{\alpha}}}(V) \equiv \{A \in \mathrm{pa}_{\mathcal{G}}(V) | (AV)_{\rightarrow} \notin \boldsymbol{\alpha}\}$.

In words, this states that the response of V to $\eta_{\mathfrak{a}_{\boldsymbol{\alpha}}}$, where $\mathfrak{a}_{\boldsymbol{\alpha}} \in \mathfrak{X}_{\boldsymbol{\alpha}}$ is defined as the potential outcome where all parents of V along edges in $\boldsymbol{\alpha}$ are assigned an appropriate value from $\mathfrak{a}_{\boldsymbol{\alpha}}$, and all other parents are assigned whatever value they would have attained under an edge intervention $\eta_{\mathfrak{a}_{\boldsymbol{\alpha}}}$ (these are defined recursively, and the definition terminates because of the lack of directed cycles in $\mathcal{G}$).

As before, given a set $\mathbf{Y} = \{Y_1, \ldots, Y_k\}$ of random variables, we denote $\{Y_1(\mathfrak{a}_{\boldsymbol{\alpha}}), \ldots, Y_k(\mathfrak{a}_{\boldsymbol{\alpha}})\}$ by $\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}})$ or $\{\mathbf{Y}\}(\mathfrak{a}_{\boldsymbol{\alpha}})$.

3.2. *Direct and indirect effects as responses to edge interventions.* Just as responses to node interventions can be used to represent total causal effects, so can responses to edge interventions be used to represent direct and indirect effects. Consider again Figure 2(a), but now assume A is the treatment (one of two drugs a, a'), Y is the outcome (survival), and M is a dangerous side effect that mediates some of the effect of A on Y.

We may be interested in how much of the total effect, as formalized via the ACE contrast $\mathbb{E}[Y(a)] - \mathbb{E}[Y(a')]$, can be attributed to the direct effect of the drugs on Y,

and how much to the mediated effect via the side effect M. To formalize this, we want to consider how Y varies if we can set treatments separately for the purposes of the direct causal pathway represented by $(AY)_{\rightarrow}$ and the pathway mediated by M, represented by $(AM)_{\rightarrow}$. This is precisely what edge interventions allow us to do. Consider $\eta_{\mathfrak{a}}$ that sets $(AM)_{\rightarrow}$ to a and $(AY)_{\rightarrow}$ to a'. Then (3) implies $Y(\mathfrak{a}) = Y(a', M(a))$. We can use this type of response to define the direct effect as the contrast $\mathbb{E}(Y(a)) - \mathbb{E}[Y(a', M(a))]$, and the indirect effect as the contrast $\mathbb{E}[Y(a', M(a))] - \mathbb{E}[Y(a')]$. Note that the ACE is a sum of the direct and indirect effect contrasts above.

The idea of using nested responses like $Y(a', M(a))$ to represent direct and indirect effects for mediation analysis appears in [16], and is discussed in the context of graphical causal models in [10]. Our contribution is to aid interpretability of such nested responses by viewing them as responses to interventions associated with edges; graphical features intuitively associated with effects we are trying to formalize.

Just as it is good practice to only discuss node interventions in settings where it is possible, at least in principle, to assign treatment by fiat, so it is good practice to only discuss edge interventions in settings where it is possible, at least in principle, to conceive of assigning only those components of the overall treatment that influences a particular direct consequence. For instance, if smoking affects cardiovascular disease only by means of nicotine content, then we might simulate the absence of smoking, but only for the purposes of cardiovascular disease, by assigning the "treatment" of nicotine-free cigarettes. In this paper, we leave the issues of applicability of edge interventions and mediation analysis in particular settings aside [19], and consider, in subsequent sections, questions of identification and the form of resulting functionals.

3.3. *Path interventions*. We are going to define responses to path interventions, which associate a set of directed paths with values of sources of every path in the set. A response to a path intervention will behave as if the source of a path were set to a particular value, *but only for the purposes of a particular outgoing directed path*. This behavior generalizes the behavior of edge interventions, where vertices may behave differently with respect to different outgoing edges. Path interventions serve as a very general, graphical representation of counterfactual quantities associated with causal pathways that generalizes both edge and path interventions. The supplementary materials [29] contain our rationale for the use of path interventions versus simpler or more algebraic approaches to representing counterfactuals of interest.

To make sure we end up with well-defined responses, we insist on a property for sets of directed paths called *properness*. A set of directed paths $\boldsymbol{\alpha}$ in a DAG $\mathcal{G}$ is called *proper* if no path in $\boldsymbol{\alpha}$ is a prefix subpath of another path in $\boldsymbol{\alpha}$. A set consisting of a single path is always proper, as is a set of length 1 paths (e.g., a set

of edges). In the remainder of the paper, when we say "a set of paths $\boldsymbol{\alpha}$," we mean a proper set of directed paths.

For a set of paths $\boldsymbol{\alpha}$ in a DAG $\mathcal{G}$, define $\mathfrak{X}_{\boldsymbol{\alpha}} \equiv \mathfrak{X}_{\mathrm{so}_{\mathcal{G}}(\boldsymbol{\alpha})}$. In other words, $\mathfrak{X}_{\boldsymbol{\alpha}}$ is a Cartesian product of the state spaces of source variables of all directed paths in $\boldsymbol{\alpha}$. Since sets of paths clearly generalize sets of edges, the same issue occurs where a single vertex in $\mathcal{G}$ may occur multiple times in $\mathfrak{X}_{\boldsymbol{\alpha}}$. As before, to emphasize this, we will denote elements of $\mathfrak{X}_{\boldsymbol{\alpha}}$ by lowercase Frankfurt font: $\mathfrak{a}$, possibly indexed by a path set subscript: $\mathfrak{a}_{\boldsymbol{\alpha}}$.

We denote a forced assignment of variables corresponding to source vertices of paths from $\boldsymbol{\alpha}$ to an element of $\mathfrak{X}_{\boldsymbol{\alpha}}$ as a *path intervention*. A path intervention which assigns $\boldsymbol{\alpha}$ to an element $\mathfrak{a}_{\boldsymbol{\alpha}} \in \mathfrak{X}_{\boldsymbol{\alpha}}$ will denoted by $\pi_{\mathfrak{a}_{\boldsymbol{\alpha}}}$. As with elements of $\mathfrak{X}_{\mathbf{A}}$, we denote a restriction of $\mathfrak{a}$ by a set subscript. That is, if $\mathfrak{a}_{\boldsymbol{\alpha}} \in \mathfrak{X}_{\boldsymbol{\alpha}}$, and $\boldsymbol{\beta} \subseteq \boldsymbol{\alpha}$, then $\mathfrak{a}_{\boldsymbol{\beta}}$ is a restriction of $\mathfrak{a}$ to variables corresponding to source vertices of $\boldsymbol{\beta}$.

As was the case with node and edge interventions, our definition of path interventions will be inductive. To get the induction to work, we need to consider how treatments affect the response via pathways that end in a particular edge. We use the following definition to formalize this. Given a set of paths $\boldsymbol{\alpha}$ in a DAG $\mathcal{G}$, and an edge $(WY)_{\rightarrow}$, define a *funnel operator* $\lhd_{(WY)_{\rightarrow}}$ which maps from $\boldsymbol{\alpha}$ to the set of paths $\lhd_{(WY)_{\rightarrow}}(\boldsymbol{\alpha})$ obtained from $\boldsymbol{\alpha}$ by replacing any path of the form $(A, \ldots, W, Y)_{\rightarrow}$ by $(A, \ldots, W)_{\rightarrow}$, by removing all paths containing W but no suffix $(WY)_{\rightarrow}$, and keeping all other paths intact.

LEMMA 3.1. *If $\boldsymbol{\alpha}$ is proper, then for any edge $(WY)_{\rightarrow}$, so is $\lhd_{(WY)_{\rightarrow}}(\boldsymbol{\alpha})$.*

Given a path intervention π that assigns $\boldsymbol{\alpha}$ to $\mathfrak{a}_{\boldsymbol{\alpha}}$, and a funnel operator $\lhd_{(WY)_{\rightarrow}}$, we consider *funneled* path interventions on $\lhd_{(WY)_{\rightarrow}}(\boldsymbol{\alpha})$. For every α such that $\lhd_{(WY)_{\rightarrow}}(\alpha) = \alpha$, the funneled path intervention assigns α to $\mathfrak{a}_{\alpha}$, that is it keeps the same value assignment as the original path intervention. For the path $\alpha \equiv (A \cdots W, Y)_{\rightarrow}$ the funneled path intervention assigns $\lhd_{(WY)_{\rightarrow}}(\alpha)$ to $\mathfrak{a}_{(A \cdots WY)_{\rightarrow}}$, that is assigns the value given by the original intervention to $(A \cdots WY)_{\rightarrow}$. We denote such an assignment by $\mathfrak{a}_{\lhd_{(WY)_{\rightarrow}}(\boldsymbol{\alpha})}$.

Our insistence on $\boldsymbol{\alpha}$ being proper, together with Lemma 3.1, means that there is never any ambiguity in defining the funneled path intervention. That is, it is never the case that two distinct paths in $\boldsymbol{\alpha}$ are of the form $(A \cdots W)_{\rightarrow}$ and $(A \cdots WY)_{\rightarrow}$. If such a pair of paths were allowed, the difficulty would then be that these paths can both reasonably be claimed to represent an effect of setting A along the path $(A \cdots WY)_{\rightarrow}$, while potentially disagreeing on what that setting is.

We are now ready to define responses to path interventions. For every $V \in \mathbf{V}$, a set of paths $\boldsymbol{\alpha}$ in a DAG $\mathcal{G}$, and an element $\mathfrak{a}_{\boldsymbol{\alpha}} \in \mathfrak{X}_{\boldsymbol{\alpha}}$, we define the response of V to $\pi_{\mathfrak{a}_{\boldsymbol{\alpha}}}$ as

$$V(\mathfrak{a}_{\boldsymbol{\alpha}}) \equiv V\big(\mathfrak{a}_{(*V)_{\rightarrow} \in \boldsymbol{\alpha}}, \{W(\mathfrak{a}_{\lhd_{(WY)_{\rightarrow}}(\boldsymbol{\alpha})}) | W \in \mathrm{pa}_{\mathcal{G}}^{\overline{\boldsymbol{\alpha}}}(V)\}\big), \tag{4}$$

where $\mathrm{pa}_{\mathcal{G}}^{\overline{\boldsymbol{\alpha}}}(V) \equiv \{W \in \mathrm{pa}_{\mathcal{G}}(V) | (WV)_{\rightarrow} \notin \boldsymbol{\alpha}\}$.

In words, this states that the response of V to $\pi_{\mathfrak{a}_{\boldsymbol{\alpha}}}$, where $\mathfrak{a}_{\boldsymbol{\alpha}} \in \mathfrak{X}_{\boldsymbol{\alpha}}$ is defined as the potential outcome where all parents of V along edges which are (length 1) paths in $\boldsymbol{\alpha}$ are assigned an appropriate value from $\mathfrak{a}_{\boldsymbol{\alpha}}$, and all other parents W are assigned whatever value they would have attained under the funneled path intervention associated with a funnel operator for the edge between that parent W and V. Note that the definition is inductive for such parents, with the result of applying a funnel operator serving as the new set of paths. Lemma 3.1 ensures that properness propagates to this set, and thus the overall response is well defined.

For example, if $\pi_{\mathfrak{a}}$ assigns w to $(WAMY)_{\rightarrow}$ in Figure 2(a), then $Y(\mathfrak{a})$ is defined by (4) to equal $Y(M(A(w)), A)$. We will use a notational shorthand for responses to path interventions, where rather than listing nested responses in parentheses after the response, we list the paths with the source node replaced by the intervened on value. For example, we write $Y(\mathfrak{a}) = Y(M(A(w)), A)$ above as $Y((wAMY)_{\rightarrow})$. We use the same shorthand for responses to edge interventions.

As before, given a set $\mathbf{Y} = \{Y_1, \ldots, Y_k\}$ of random variables, we denote $\{Y_1(\mathfrak{a}_{\boldsymbol{\alpha}}), \ldots, Y_k(\mathfrak{a}_{\boldsymbol{\alpha}})\}$ by $\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}})$ or $\{\mathbf{Y}\}(\mathfrak{a}_{\boldsymbol{\alpha}})$.

3.4. *Responses to path interventions to natural values.* So far, we have defined path interventions as a mapping from a proper set of directed paths $\boldsymbol{\alpha}$ to values in $\mathfrak{X}_{\boldsymbol{\alpha}}$. However, we might be interested in considering responses to interventions that assign a variable not to a specific constant value, but to a value the variable would have attained under a no intervention regime. For instance, this might happen if the baseline exposure is one received by the general population, not a specific exposure level assigned by the experimenter, or if the effect of multiple treatments on the treated is of interest. In the context of node interventions, this situation was discussed in [4]. In order for responses to path interventions to include this case, we must extend the definition of path interventions to include intervening to *natural* values, that is values attained by variables under no interventions. Allowing arbitrary variables to be set to natural values may lead to identification difficulties even in very simple cases. Consider the following response to a node intervention in the MWM given by Figure 2(a), $\{A, Y\}(A, w)$. In words, this is the joint response of A and Y to an intervention where W is set to value w, and A is set to the natural value it attains under no interventions. The definition of responses to node interventions via recursive substitution shows that $\{A, Y\}(A, w) = Y(A), A(w)$. However, the distribution $p(A, A(w))$ is not identified under the MWM for the graph in Figure 2(a); see Lemma 5.8, and thus neither is the joint response in question.

To avoid this difficulty, we consider only a special subset of path interventions containing settings on natural values. This special subset can safely be rephrased in such a way that only interventions on constants remain explicit. To define this special subset, we need a few preliminary definitions.

For a node A, and a directed path (or an edge) α with source A, define the *extended state space* as follows $\mathfrak{X}_A^* \equiv \mathfrak{X}_A \cup \{A\}$, and $\mathfrak{X}_\alpha^* \equiv \mathfrak{X}_\alpha \cup \{A\}$. We define

the extended state space for sets of nodes, edges and paths via a Cartesian product as before. An intervention on an extended state space is allowed on either any constant value, or on the "natural value."

Given a set of paths $\boldsymbol{\alpha}$ and a response set $\mathbf{Y}$, we call a directed path α *relevant* for $\mathbf{Y}$ given $\boldsymbol{\alpha}$ if $\alpha = (A \cdots Y)_{\rightarrow}$, where $Y \in \mathbf{Y}$, and no path in $\boldsymbol{\alpha}$ is a subpath of α except possibly a prefix of α. We denote the set of all relevant paths for $\mathbf{Y}$ given $\boldsymbol{\alpha}$ in $\mathcal{G}$ by $\text{rel}_{\mathcal{G}}(\mathbf{Y}|\boldsymbol{\alpha})$.

Paths relevant for $\mathbf{Y}$ given $\boldsymbol{\alpha}$ are those paths consisting of vertices that follow a particular recursive sequence of invocations of definition (4). For example, assume we are interested in the singleton response set $\{Y\}$ and a singleton path set $\{(WAMY)_{\rightarrow}\}$ in Figure 2(a). Then defining $Y((wAMY)_{\rightarrow})$ for a particular w via (4) entails defining intermediate responses $M((wAM)_{\rightarrow})$ and $A((wA)_{\rightarrow})$. The sequence of vertices (A, M, Y) are all linked by directed edges by (4), and $(AMY)_{\rightarrow}$ is relevant for $\{Y\}$ given $\{(WAMY)_{\rightarrow}\}$. Similarly, $(WAMY)_{\rightarrow}$ and $(WAY)_{\rightarrow}$ are relevant for $\{Y\}$ given $\{(WAMY)_{\rightarrow}\}$.

We now give two useful results about relevant paths.

LEMMA 3.2. *If* $\alpha \in \text{rel}_{\mathcal{G}}(\mathbf{Y}|\boldsymbol{\alpha})$, *then* $\beta \in \text{rel}_{\mathcal{G}}(\mathbf{Y}|\boldsymbol{\alpha})$ *for any suffix subpath* β *of* α.

LEMMA 3.3. *If* $\boldsymbol{\beta} \subseteq \boldsymbol{\alpha}$, *then for any* $\mathbf{Y}$, $\text{rel}_{\mathcal{G}}(\mathbf{Y}|\boldsymbol{\alpha}) \subseteq \text{rel}_{\mathcal{G}}(\mathbf{Y}|\boldsymbol{\beta})$.

A set of interventions may not all have an effect on a response, due to constraints of the model. For instance, since $Y(a, m, w) \neq Y(a', m, w)$ but $Y(a, m, w) = Y(a, m, w')$ for any m, a, w, a', w' in Figure 2(a), A has an effect on Y, but W does not, given that we also intervene on A and M. We extend this notion to path interventions, and call those paths with sources that actually have an effect on the response, given interventions on other paths, *live*. More precisely, given a proper set of paths $\boldsymbol{\alpha}$ and a response set $\mathbf{Y}$, we call a path $\alpha \in \boldsymbol{\alpha}$ *live* for $\mathbf{Y}$ given $\boldsymbol{\alpha}$ if there is an element of $\text{rel}_{\mathcal{G}}(\mathbf{Y}|\boldsymbol{\alpha})$ containing α as a prefix.

Consider the maximal subset of $\boldsymbol{\alpha}$ consisting of paths in $\boldsymbol{\alpha}$ live for $\mathbf{Y}$ given $\boldsymbol{\alpha}$, or $\boldsymbol{\alpha}_{\mathbf{Y}} \equiv \{\alpha \in \boldsymbol{\alpha} | \alpha \text{ live for } \mathbf{Y} \text{ given } \boldsymbol{\alpha}\}$. We say a set of directed paths $\boldsymbol{\alpha}$ is live for $\mathbf{Y}$ if $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{\mathbf{Y}}$. When discussing path interventions, we can always restrict our attention to sets of paths live for $\mathbf{Y}$ without loss of generality, due to the following result.

LEMMA 3.4. *For any* $\mathbf{Y}$ *and* $\boldsymbol{\alpha}$ *proper for* $\mathbf{Y}$, $\text{rel}_{\mathcal{G}}(\mathbf{Y}|\boldsymbol{\alpha}) = \text{rel}_{\mathcal{G}}(\mathbf{Y}|\boldsymbol{\alpha}_{\mathbf{Y}})$, $(\boldsymbol{\alpha}_{\mathbf{Y}})_{\mathbf{Y}} = \boldsymbol{\alpha}_{\mathbf{Y}}$, *and in addition, for any* $\mathfrak{a}_{\boldsymbol{\alpha}}$, $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}})) = p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}_{\mathbf{Y}}}))$.

We now show that we can either ignore interventions to natural values in a response to a path intervention, or the response is not identified under the MWM. The set of paths for which the former is true for the response $\mathbf{Y}$ will be called *natural* for $\mathbf{Y}$. Due to this result, we do not need to consider interventions to natural values explicitly.

DEFINITION 1. Let $\boldsymbol{\alpha}$ be live for $\mathbf{Y}$. Let $\pi_{\mathfrak{a}_{\boldsymbol{\alpha}}}$ be a path intervention in $\mathcal{G}$ where a subset $\boldsymbol{\alpha}^* \subseteq \boldsymbol{\alpha}$ is assigned constant values, and $\boldsymbol{\alpha} \setminus \boldsymbol{\alpha}^*$ is assigned natural values. Then if no element of $\text{rel}_{\mathcal{G}}(\mathbf{Y}|\boldsymbol{\alpha}^*)$ with a prefix subpath in $\boldsymbol{\alpha}^*$ contains a subpath in $\boldsymbol{\alpha} \setminus \boldsymbol{\alpha}^*$, we say π is *natural* for $\mathbf{Y}$.

LEMMA 3.5. *Let $\pi_{\mathfrak{a}_{\boldsymbol{\alpha}}}$ be a path intervention natural for $\mathbf{Y}$, and $\boldsymbol{\alpha}^* \subseteq \boldsymbol{\alpha}$ is all paths assigned constant values by π. Then $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}})) = p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}^*}))$.*

LEMMA 3.6. *If $\pi_{\mathfrak{a}_{\boldsymbol{\alpha}}}$ is not natural for $\mathbf{Y}$ in $\mathcal{G}$, then $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}}))$ is not identified under the MWM for $\mathcal{G}$.*

Lemma 3.5 does not guarantee that a response to a natural path intervention is identifiable, merely that it can be expressed as a response to an intervention only setting to constant values.

4. Causal inference targets as responses to path interventions. In this section we consider how a number of targets of interest in causal inference, including novel targets not previously considered in the literature, may be expressed as responses to path interventions.

We use as our running example the two time point fragment of a longitudinal study in HIV research, described in Section 2.4. We consider path-specific effects that arise in mediation analysis, and effects of treatment on the multiply treated, which are of interest in tort cases (since these are effects of the exposure on those actually exposed), and in epidemiology if natural exposure levels carry information about the causal effect of the exposure. We also describe a novel inference target that combines features of mediated effects, and effects of treatment on the treated, that we call effect of treatment on the indirectly treated. It is not straightforward to see whether these types of effects are identifiable, and under what model, nor is it obvious whether there is a single unifying principle which governs identification for these effects.

By translating the effect types above into responses to path interventions, we show that such responses form a very general class of causal inference targets. Thus, the advantage of path interventions is that we can use them to give a single characterization for a wide variety of targets of interest at once. The close relationship between effects of treatment on the treated and mediated effects hinted by their common generalization as responses to path interventions is currently not widely known.

We will define a special set of directed paths important for our translation scheme. Given a treatment set $\mathbf{A}$ and an outcome set $\mathbf{Y}$ (that possibly intersect) in a DAG $\mathcal{G}$, define the set $\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}}$ to be the set of all directed paths with a source in $\mathbf{A}$, a sink in $\mathbf{A} \cup \mathbf{Y}$ and which do not intersect $\mathbf{A} \cup \mathbf{Y}$ except at the source and sink. Since $\mathbf{A}$ and $\mathbf{Y}$ are allowed to intersect, the names "treatment" and "outcome" are slightly misleading. We allow the intersection to admit cases such as effect of

treatment on the treated (ETT) where some treatments are also treated as outcomes for the purposes of certain paths.

LEMMA 4.1. *$\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}}$ is always proper.*

4.1. *Effects of treatment on the treated.* We consider an effect on the mean difference scale where we condition on the naturally observed treatment levels. This is known as the effect of treatment on the treated (ETT), and in our two time point HIV example, it is defined as follows:

$$\text{ETT} \equiv \mathbb{E}\big[Y(a_1, a_2)|a_1, a_2\big] - \mathbb{E}\big[Y\big(a_1', a_2'\big)|a_1, a_2\big].$$

This contrast is often of interest to epidemiologists. It also arises in cases where interventions are functions of the natural value of the exposure. For example, we may be interested in outcome for people who were encouraged to exercise for 30 more minutes than they normally would, which is a random variable of the form $Y(A+30) \equiv Y(a+30)|A=a$. These types of interventions are discussed in [36], in particular sufficient conditions for identification under the SWM, in terms of the extended g-formula (2) are given there and in [13].

Assume A_1 is a binary variable (only two treatment levels). If we consider, instead, the ETT with respect to only the exposure A_1, we obtain the following derivation for the second term in the contrast

$$p\big(Y_2\big(a_1'\big)|a_1\big) = \frac{p(Y_2(a_1'), a_1)}{p(a_1)} = \frac{p(Y_2(a_1')) - p(Y_2(a_1'), a_1')}{p(a_1)},$$

where the first identity is by definition, and the second by the binary treatment assumption. Since consistency implies $p(Y_2(a_1), a_1) = p(Y_2, a_1)$ for any value a_1, the ETT for a single binary exposure A_1 can be identified if $p(Y_2(a_1))$ is identified.

However, if the exposure is not binary, or if there are multiple exposures, as in our example, we cannot use the same algebraic trick to obtain identification, and we must proceed by exploiting additional assumptions in our causal model.

In our case, the first conditional mean in the contrast can be readily identified via consistency: $\mathbb{E}[Y(a_1, a_2)|a_1, a_2] = \mathbb{E}[Y|a_1, a_2]$. However, the second conditional mean presents a problem, because it contains a conflict between the naturally observed exposures, and the assigned exposures. Here, we show how to represent the underlying joint distribution over potential outcomes, $p(Y_2(a_1, a_2), A_1, A_2)$, in terms of path interventions, and then attack the identification problem for *all* responses to path interventions, which would then include the problematic second term of the ETT.

We consider all directed paths from A_2 to Y_2, which we assign a value a_2, all directed paths from A_1 to Y_2 not through A_2, which we assign a value a_1, and all directed paths from A_1 to A_2, which we assign the natural value of A_1. Note that this set of paths is simply $\boldsymbol{\alpha}_{\{A_1,A_2\},\{Y_2\},\mathcal{G}}$ for $\mathcal{G}$ that is the transitive closure with respect to blue edges of the graph in Figure 2(b), and thus is proper by Lemma 4.1.

We then consider the response of A_1, A_2, Y_2 to the path intervention so defined, or $\{A_1, A_2, Y_2\}(\mathfrak{a}_{\boldsymbol{\alpha}})$. By our definition, all paths set to a value ancestral for A_1, A_2 are set to natural values. Thus, $\{A_1, A_2\}(\mathfrak{a}_{\boldsymbol{\alpha}})$ is defined in terms of natural values of its direct causal parents, or as $A_1(C_0) = A_1$ and $A_2(Y_1, C_1, W_1, A_1, C_0) = A_2$.

Finally, we consider all paths ancestral for Y_2. Since A_1 and A_2 are parents of Y_2 in $\mathcal{G}^*$, the single edge paths $(A_1Y_2)_{\rightarrow}$ and $(A_2Y_2)_{\rightarrow}$ are in our set, thus we substitute a_1 and a_2 into the potential outcome answer. Furthermore, for other parents of Y_2, namely $C_0, U, W_1, C_1, Y_1, W_2$ and C_2, we consider an appropriate set derived from $\boldsymbol{\alpha}$. For example, for the node W_2, we replace the path $A_2 \rightarrow W_2 \rightarrow Y_2$ by a path $A_2 \rightarrow W_2$ (while keeping the assignment a_2). We proceed in this way recursively until we obtain the response for Y_2, which is

$$Y_2\big(a_1, a_2, U, C_0, W_1(a_1, \ldots), C_1(a_1, \ldots), Y_1(a_1, \ldots),$$
$$W_2(a_1, a_2, \ldots), C_2(a_1, a_2, \ldots)\big),$$

where $\ldots$ is a shorthand that means "include all earlier potential outcomes." For example, $C_1(a_1, \ldots)$ means $C_1(a_1, W_1(a_1, U, C_0), U, C_0)$. By definition of node intervention responses, this counterfactual is equal to $Y_2(a_1, a_2)$, and our overall joint distribution over the responses is $p(Y_2(a_1, a_2), A_1, A_2)$.

For arbitrary sets of treatments $\mathbf{A}$ and outcomes $\mathbf{Y}$, and active treatment values $\mathbf{a}$, we may still represent ETT as a single mean difference, for example, $\mathbb{E}[f(\mathbf{y})]_{[p(\mathbf{Y}(\mathbf{a})|\mathbf{a})]} - \mathbb{E}[f(\mathbf{y})]_{[p(\mathbf{Y}(\mathbf{a}')|\mathbf{a})]}$, for some function $f(\mathbf{y})$.

Note that though ETT resembles the total effect, it is in fact a more complex kind of counterfactual. This is because we are simultaneously interested in "outcome responses" $\mathbf{Y}$, and "treatment responses" $\mathbf{A}$. Defining these treatment responses may introduce conflicts among intermediate counterfactual responses, not well represented by node interventions, which is why we represent ETT as a response to a path intervention.

The *ETT path intervention* $\pi^{\mathbf{a}}_{\mathfrak{a}_{\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}}}}$ simply assigns all paths in $\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}}$ to the appropriate value. That is, paths from $\mathbf{A}$ to $\mathbf{A}$ are assigned the appropriate natural value, and paths from $\mathbf{A}$ to $\mathbf{Y}$ are assigned the appropriate value in $\mathbf{a}$. Given this definition, either the ETT is not identified, or the joint distribution from which ETT is obtained corresponds to the joint response of $\mathbf{Y} \cup \mathbf{A}$ to the ETT path intervention.

LEMMA 4.2. *If there exists* $A \in \mathbf{A}$ *such that* $A(\mathfrak{a}_{\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}}}) \neq A$, $p(\mathbf{Y}(\mathbf{a}), \mathbf{A})$ *is not identified under the MWM for* $\mathcal{G}$. *If there does not exist such an* A, $p(\mathbf{Y}(\mathbf{a}), \mathbf{A}) = p(\{\mathbf{Y} \cup \mathbf{A}\}(\mathfrak{a}_{\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}}}))$.

If $p(\mathbf{Y}(\mathbf{a}), \mathbf{A})$ is expressible as a response to a path intervention, it may still not be identifiable under the MWM.

Our subsequent results on identification of path interventions under the MWM complement identification results in [13, 36]. In particular, our results imply the distribution $p(Y(a, m)|A, M)$ is identified under the MWM for Figure 2(a), but not under the SWM for Figure 2(a).

4.2. *Path-specific effects*. Next, we consider the mediated effect of A_1, A_2 on Y_2 through C_1, C_2; in other words, the effect of exposures on outcome mediated by adherence. Originally these kinds of effects were considered in [3] in the context of linear models, and were generalized to a form not restricted by particular parametric models in [16]. We discuss a simple version of mediated effects in the graph in Figure 2(a), known as natural direct and indirect effects [10, 16] in Section 3.2, where we represented them as responses to edge interventions.

In our case, we are interested in a more complicated effect, but we can represent it using a similar idea using paths rather than edges—paths we are interested in are assigned active treatment values a_1, a_2, while paths we are not interested in are assigned baseline treatment values a'_1, a'_2. The paths we are interested in are all directed paths with the first edges are one of $\{(A_1C_1)_{\rightarrow}, (A_1C_2)_{\rightarrow}, (A_2C_2)_{\rightarrow}\}$, which end in Y_2, and which do not proceed through A_2 if started at A_1. The paths we are not interested in are all other paths which start with A_2 or A_1 (and do not proceed through A_2) and end in Y_2. Call this assignment $\mathfrak{a}_1$. Note that the assignment $\mathfrak{a}_1$ is on the set of paths that is precisely equal to $\boldsymbol{\alpha}_{\{A_1,A_2\},\{Y_2\},\mathcal{G}}$ for $\mathcal{G}$ that is the transitive closure with respect to blue edges of the graph in Figure 2(b), and thus is proper.

We apply our definition to obtain a response of Y_2 to this intervention. We must substitute a value for every parent of Y_2. The values for A_1, A_2 will be the baseline a'_1, a'_2, while the values for C_0, U will just be the natural values of those variables. Complications arise for other parents, due to the recursive nature of the definition. We proceed recursively:

$$
\begin{aligned}
Y_2(\mathfrak{a}_1) &= Y_2\big(a'_1, a'_2, \{C_2, W_2, Y_1, C_1, W_1, C_0\}(\mathfrak{a}_1), U\big),\\
C_2(\mathfrak{a}_1) &= C_2\big(a_1, a_2, \{W_2, Y_1, C_1, W_1, C_0\}(\mathfrak{a}_1), U\big),\\
W_2(\mathfrak{a}_1) &= W_2\big(a'_1, a'_2, \{Y_1, C_1, W_1, C_0\}(\mathfrak{a}_1), U\big),\\
Y_1(\mathfrak{a}_1) &= Y_1\big(a'_1, \{C_1, W_1, C_0\}(\mathfrak{a}_1), U\big),\\
C_1(\mathfrak{a}_1) &= C_1\big(a_1, \{W_1, C_0\}(\mathfrak{a}_1)\big),\\
W_1(\mathfrak{a}_1) &= W_1\big(a'_1, C_0(\mathfrak{a}_1), U\big),\\
C_0(\mathfrak{a}_1) &= C_0(U) = C_0.
\end{aligned}
$$

In the matter similar to direct and indirect effects, we can use this response along with the total effect responses to define "the effect along paths we want" as $\mathbb{E}[Y(\mathfrak{a}_1)] - \mathbb{E}[Y(a'_1, a'_2)]$, and "the effect along paths we do not want" as $\mathbb{E}[Y(a_1, a_2)] - \mathbb{E}[Y(\mathfrak{a}_1)]$. As before, the ACE additively decomposes into these two effect measures. This definition (without the use of path interventions) appears in [24].

We may also consider a response of Y_2 where the paths we are not interested in are assigned the natural values, as discussed in Section 3.4, rather than fixed baseline values. Such an effect is defined similarly.

Consider a set of active treatment values $\mathbf{a}$ of $\mathbf{A}$, a set of fixed baseline treatment values $\mathbf{a}'$, and a subset $\boldsymbol{\beta}$ of $\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}}$ (which contains "paths of interest"). Define the *fixed baseline PSE path intervention* $\pi^{\mathbf{a},\mathbf{a}',\boldsymbol{\beta}}_{\mathfrak{a}_{\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}}}}$ as a path intervention that assigns appropriate active values in $\mathbf{a}$ to sources in $\boldsymbol{\beta}$ and appropriate baseline values in $\mathbf{a}'$ to sources of all paths in $\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}} \setminus \boldsymbol{\beta}$.

Similarly, we call an intervention $\pi^{\mathbf{a},\boldsymbol{\beta}}_{\mathfrak{a}_{\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}}}}$ that assigns active values in $\mathbf{a}$ to sources of paths in $\boldsymbol{\beta}$ and appropriate *natural* values to sources of all paths in $\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}} \setminus \boldsymbol{\beta}$ the *average baseline PSE path intervention*.

Path specific effects along all paths in β (with a fixed baseline) can then be defined on the mean difference scale as $\mathbb{E}[\mathbf{Y}(\mathfrak{a}^{\mathbf{a},\mathbf{a}',\boldsymbol{\beta}}_{\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}}})] - \mathbb{E}[\mathbf{Y}(\mathbf{a}')]$, and along all paths not in β as $\mathbb{E}[\mathbf{Y}(\mathbf{a})] - \mathbb{E}[\mathbf{Y}(\mathfrak{a}^{\mathbf{a},\mathbf{a}',\boldsymbol{\beta}}_{\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}}})]$. Average baseline path specific effects on the difference scale are defined similarly.

4.3. *Effects of treatment on the indirectly treated.* In this section, we show that the language of path interventions is general enough to incorporate novel targets not currently considered in the literature. Our results immediately settle identification questions for any such target.

We consider a seemingly innocuous ETT with two treatments that in fact can only be represented by a path intervention, not an edge intervention, and variations of this target that are identified under the SWM and the MWM. Assume Figure 2(a) represents a simple two time point partially randomized observational study, where W and M are treatments at the first and second time points, respectively, A is an intermediate health measure, and Y is the outcome. We make very strong assumptions about this study. In particular, W is randomized, while M is only assigned based on A, W. Finally, no unobserved confounding exists anywhere, including between W, M and Y. We are interested in the effect of treatments W, M on the treated in this study. To obtain this contrast, we need to identify $p(Y(m, w)|W, M)$ which is identified if and only if $p(Y(m, w), W, M)$ is. It is not difficult to show that

$$\begin{aligned} p\big(Y(m, w), W, M\big) &= p\big(\{Y, M, W\}\big((wAY)_{\to}, (mY)_{\to}\big)\big) \\ &= p\big(Y\big(m, A(w)\big), M\big(A(W), W\big), W\big). \end{aligned}$$

As we will show in the next section, there is no way to express this response as a response to an edge intervention, and it is not identified under the MWM. This is the case despite the fact that there is no unobserved confounding in this study. The difficulty is that the response is defined in terms of $A(w)$ and A *jointly*, and the distribution $p(A(w), A)$ is not identified under the MWM without more assumptions.

To obtain a target that is identified under the SWM in this case, we may consider the response $Y(w, m)$ on the treated to the natural value W, and the value of M occurring under the intervention setting W to w. This results in

$p(Y(m,w),W,M(w)) = p(Y(m,A(M(w))),M(A(w)),W)$ which is then identified under the SWM. To obtain identification, we gave up on conditioning on the natural value of the second treatment M. This may not be "in the spirit" of the ETT target.

One compromise is to assume a stronger model, the MWM and allow the response M to be "as natural as possible" while still retaining identification. This would correspond to defining a contrast in terms of $p(\{Y,W,M\}((mY)_{\rightarrow},(wAY)_{\rightarrow},(wAM)_{\rightarrow}))$, which in turn is equivalent to $p(\{Y,W,M\}((mY)_{\rightarrow},(wA)_{\rightarrow}))$. A conditional distribution $p(Y((mY)_{\rightarrow},(wA)_{\rightarrow})|M((wA)_{\rightarrow},(w'M)_{\rightarrow}) = m', W = w')$ represents the response $Y(w,m)$ among those individuals for whom the value for W is naturally w' (untreated), and for whom the value for M would have been m' (untreated) under the situation where W acts as if set to treatment value w for paths shared by Y and M, and acts as if set to untreated value w' otherwise.

We can define a contrast based on this quantity as follows:

$$\begin{aligned}&\mathbb{E}\big[Y\big((mY)_{\rightarrow},(wA)_{\rightarrow}\big)|M\big((wA)_{\rightarrow},\big(w'M\big)_{\rightarrow}\big)=m',W=w'\big]\\&\quad-\mathbb{E}\big[Y\big(\big(m'Y\big)_{\rightarrow},\big(w'A\big)_{\rightarrow}\big)|M\big(\big(w'A\big)_{\rightarrow},\big(w'M\big)_{\rightarrow}\big)=m',W=w'\big],\end{aligned}$$

which we call "the effect of treatment on the indirectly treated (ETIT)." The name is due to the fact that we consider people whose natural treatment value W is untreated, and whose followup treatment M assumes the untreated value if viewed as a response to the indirect effect of the first treatment. Such a quantity would be difficult to conceive of without a direct representation of effects along pathways, something path interventions provide. Our results also directly imply that this quantity is identified under the MWM, but not SWM.

5. Identification of edge and path interventions. Having established a correspondence between responses to path interventions and a variety of targets of interest in causal inference, we now consider what assumptions are necessary to express path interventions as edge interventions, edge interventions as node interventions and edge and node interventions as functions of the observed data.

As we showed in Section 3.4, we can restrict our attention to path interventions that only assign paths to constant values, since paths that are assigned natural values either can be dropped from the intervention without affecting the response, or the overall response is not identified.

5.1. *Node and edge interventions as path interventions.* If node interventions are a special case of edge interventions, which are in turn a special case of path interventions, we ought to be able to give a path intervention the response of which is equal to the response to an arbitrary node or edge intervention. For any such response, there may be multiple path interventions the responses to which are identical. Here, we construct particular path interventions that work based on the set of paths $\alpha_{\mathbf{A},\mathbf{Y},\mathcal{G}}$ we defined earlier.

LEMMA 5.1. *Let* $\mathbf{A}, \mathbf{Y}$ *be disjoint vertex sets in a DAG* $\mathcal{G}$, *and* $\mathbf{a}$ *a value assignment to* $\mathbf{A}$. *Let* $\pi^{\mathbf{a}}_{\mathfrak{a}_{\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}}}}$ *assign each* $\alpha \in \boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}}$ *to* $\mathbf{a}_{\mathrm{so}_{\mathcal{G}}(\alpha)}$. *Then* $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}}})) = p(\mathbf{Y}(\mathbf{a}))$.

LEMMA 5.2. *Let* $\mathbf{Y}$ *be a vertex set in a DAG* $\mathcal{G}$, *and* $\boldsymbol{\alpha}$ *a set of edges, with* $\mathfrak{a}_{\boldsymbol{\alpha}}$ *an assignment to* $\boldsymbol{\alpha}$. *Let* $\mathbf{A} = \mathrm{so}_{\mathcal{G}}(\boldsymbol{\alpha})$, *and* $\boldsymbol{\alpha}_{\mathbf{Y},\mathcal{G}}$ *be a subset of* $\boldsymbol{\alpha}_{\mathbf{A},\mathbf{Y},\mathcal{G}}$ *consisting of paths with an edge prefix in* $\boldsymbol{\alpha}$. *Let* $\pi^{\boldsymbol{\alpha}}_{\mathfrak{a}_{\boldsymbol{\alpha}_{\mathbf{Y},\mathcal{G}}}}$ *assign each* $\alpha \in \boldsymbol{\alpha}_{\mathbf{Y},\mathcal{G}}$ *to the value assigned to the edge prefix of* α *by* $\mathfrak{a}_{\boldsymbol{\alpha}}$. *Then* $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}_{\mathbf{Y},\mathcal{G}}})) = p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}}))$.

5.2. *Identification of edge interventions.* The difficulty with edge interventions is that a single response to an edge intervention may involve other responses with conflicting treatment assignments. It is this feature of edge interventions which in general prevents their identification under the SWM, and which requires the stronger assumptions of the MWM. If such a conflicting assignment is absent, the edge intervention can be rephrased as a node intervention. We show this absence of conflict is characterized by a property we call node consistency.

A set of edges $\boldsymbol{\alpha}$ live for $\mathbf{Y}$ is called *consistent* for $\mathbf{Y}$ if for every node A, the set of prefix edges of the path set $\{\alpha \in \mathrm{rel}_{\mathcal{G}}(\mathbf{Y}|\boldsymbol{\alpha}) | \, \mathrm{so}_{\mathcal{G}}(\alpha) = A\}$ is either disjoint from $\boldsymbol{\alpha}$ or contained in $\boldsymbol{\alpha}$.

For a set of edges $\boldsymbol{\alpha}$ live and consistent for $\mathbf{Y}$, we call an edge intervention $\eta_{\mathfrak{a}_{\alpha}}$ *node consistent* for $\mathbf{Y}$ if for every node A, all edges in $\boldsymbol{\alpha}$ with A as the source node are assigned the same value (say a). Any edge intervention that is not node consistent we call node inconsistent, including any edge intervention on a set of edges not consistent for an outcome set of interest.

The edge set $\{(AY)_{\rightarrow}\}$ in Figure 2(a) is live but not consistent for $\{Y\}$, thus any edge intervention on this set (that sets to constant values) is inconsistent for $\{Y\}$. An edge intervention corresponding to $Y((aY)_{\rightarrow}, (aM)_{\rightarrow})$ is node consistent for Y, while an edge intervention corresponding to $Y((aY)_{\rightarrow}, (a'M)_{\rightarrow})$ is consistent, but not node consistent for Y.

For an edge intervention $\eta_{\mathfrak{a}_{\alpha}}$ node consistent for $\mathbf{Y}$, define the following set of value assignments to $\mathbf{A} = \mathrm{so}_{\mathcal{G}}(\boldsymbol{\alpha})$, $\mathbf{a}_{\boldsymbol{\alpha}} \equiv \{a | \eta \text{ assigns } a \text{ to } (AB)_{\rightarrow} \in \boldsymbol{\alpha}\}$. Let $\nu_{\mathbf{a}_{\alpha}}$ be the *induced node intervention* for $\eta_{\mathfrak{a}_{\alpha}}$.

LEMMA 5.3. *Given a DAG* $\mathcal{G}$ *with vertices* $\mathbf{V}$, *and an edge intervention* $\eta_{\mathfrak{a}_{\alpha}}$ *node consistent for* $\mathbf{Y} \subseteq \mathbf{V}$, $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}})) = p(\mathbf{Y}(\mathbf{a}_{\boldsymbol{\alpha}}))$.

PROOF. This follows by Lemmas 5.1 and 5.2. □

COROLLARY 5.1. *If* $\eta_{\mathfrak{a}_{\alpha}}$ *is node consistent for* $\mathbf{Y}$, *then* $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}}))$ *is identified as a functional of* $p(\mathbf{V})$ *under the SWM via the appropriate marginal of the extended g-formula* (2) *for the response to the corresponding induced node intervention.*

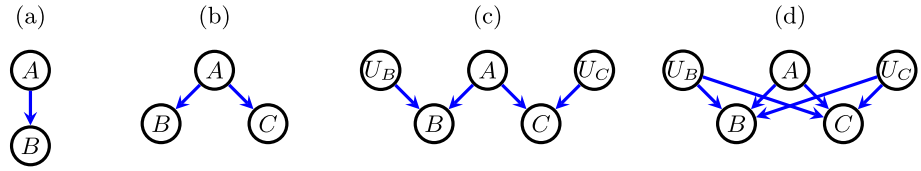

FIG. 3. (a) *A simple MWM where* $p(B(a), B(a'))$ *is not identified.* (b) *A simple SWM where* $p(B(a), C(a'))$ *is not identified.* (c), (d) *Two graphs corresponding to causal structures used in the proof of Lemma* 5.4.

We next show that if an edge intervention is not node consistent, then responses to this intervention are not identifiable from $p(\mathbf{V})$ under the SWM. By this we mean that the definition of identifiability given in Section 2.3 fails, and more specifically that we can find two elements of a causal model, in the sense of Section 2.2, that agree on $p(\mathbf{V})$ but disagree on the distribution of the response of interest. We start with a simple example of a nonidentified parameter in the SWM.

LEMMA 5.4. *Responses* $p(\{B, C\}((aB)_{\rightarrow}, (a'C)_{\rightarrow}))$, $p(\{B, C\}((aB)_{\rightarrow}))$, *and* $p(\{B, C\}((aC)_{\rightarrow}))$, *are not identifiable from* $p(A, B, C)$ *under the SWM for Figure* 3(b).

The proofs of this result, which appears in the Appendix, exhibits two causal structures $c_1(\{A, B, C\}, \mathcal{G})$ and $c_2(\{A, B, C\}, \mathcal{G})$ that agree on $p(A, B, C)$, but disagree on the above responses to (node inconsistent) edge interventions. These two structures corresponding to graphs in Figure 3(c), (d). In particular, c_2 is constructed in such a way that the confounding of B and C introduced by U_B and U_C is masked under any single node intervention, but manifests if we consider responses to multiple interventions simultaneously. This is similar in spirit to an example in [19]. We can extend this simple example to a general result, due to the following lemma (stated in a more general form in terms of path rather than edge interventions).

LEMMA 5.5. *Let* $\mathcal{G}$ *be a DAG,* $\mathbf{Y}$,$\mathbf{A}$ *disjoint sets of vertices in* $\mathcal{G}$, $\boldsymbol{\alpha}$ *a set of paths live for* $\mathbf{Y}$. *Let* $\mathcal{G}^*$ *be any edge supergraph of* $\mathcal{G}$, $\mathbf{Y}^*$ *any superset of* $\mathbf{Y}$ *in* $\mathcal{G}^*$, $\boldsymbol{\alpha}^*$ *a superset of* $\boldsymbol{\alpha}$ *in* $\mathcal{G}^*$ *live and proper for* $\mathbf{Y}^*$, *such that every path in* $\boldsymbol{\alpha}^* \setminus \boldsymbol{\alpha}$ *does not exist in* $\mathcal{G}$. *Finally, let* $\pi_{\mathfrak{a}_{\boldsymbol{\alpha}^*}}$ *be a path intervention. If* $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}}), \mathbf{A})$ *is not identified under the MWM (SWM) for* $\mathcal{G}$, *then* $p(\mathbf{Y}^*(\mathfrak{a}_{\boldsymbol{\alpha}^*}), \mathbf{A})$ *and* $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}^*}), \mathbf{A} \cup \mathbf{Y}^* \setminus \mathbf{Y})$ *are not identified under the MWM (SWM) for* $\mathcal{G}^*$.

THEOREM 5.1. *Consider a DAG* $\mathcal{G}$ *with vertices* $\mathbf{V}$, *and a set of edges* $\boldsymbol{\alpha}$ *live for* $\mathbf{Y}$. *Then* $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}}))$ *is identifiable from* $p(\mathbf{V})$ *under the SWM for* $\mathcal{G}$ *if and only if* $\eta_{\mathfrak{a}_{\boldsymbol{\alpha}}}$ *is node consistent. Moreover, if* $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}}))$ *is identifiable, it is equal to the appropriate marginal of the extended g-formula* (2) *for* $p(\mathbf{Y}(\mathbf{a}_{\boldsymbol{\alpha}}))$, *the response to the induced node intervention.*

What we have shown is that node consistent edge interventions are identifiable under the SWM, but an edge intervention that is node inconsistent is not, as long as this inconsistency is "causally relevant" for some response, in the sense of there existing causal pathways from the inconsistent edges to some responses that are not interrupted by other parts of the edge intervention. However, if we are willing to adopt stronger independence assumptions of the MWM, we obtain identification of any edge intervention via a modification of the g-formula, as the following result shows.

LEMMA 5.6 (Edge g-formula). *For a DAG $\mathcal{G}$ with vertices $\mathbf{V}$, and an edge intervention $\eta_{\mathfrak{a}_\alpha}$ on an edge set $\boldsymbol{\alpha}$, we have, under the MWM for $\mathcal{G}$,*

$$p\big(\mathbf{V}(\mathfrak{a}_{\boldsymbol{\alpha}}) = \mathbf{v}\big) = \prod_{V \in \mathbf{V}} p\big(V = \mathbf{v}_V | \mathbf{v}_{\mathrm{pa}_{\mathcal{G}}^{\overline{\boldsymbol{\alpha}}}(V)}, \mathfrak{a}_{\{(WV)_\to \in \boldsymbol{\alpha}\}}\big), \tag{5}$$

where $\mathrm{pa}_{\mathcal{G}}^{\overline{\boldsymbol{\alpha}}}(V) \equiv \{A \in \mathrm{pa}_{\mathcal{G}}(V) | (AV)_\to \notin \boldsymbol{\alpha}\}$.

For example, in the graph in Figure 2(a), we can express the distribution of the response of $Y((a'M)_\to, (aY)_\to)$ using (5) as follows:

$$\begin{aligned} p\big(Y\big(a, M\big(a'\big)\big) = y\big) &= \sum_{w,a'',m} p(y|m,a)p\big(m|a',w\big)p\big(a''|w\big)p(w) \\ &= \sum_{m,w} p(y|m,a)p\big(m|a',w\big)p(w). \end{aligned}$$

If we are interested in a mean difference parameter, for example, $\mathbb{E}[Y(a, M(a'))] - \mathbb{E}[Y(a')]$, and assume there are no baseline factors W, the above reduces to

$$\sum_m \big\{\mathbb{E}[Y|m,a] - \mathbb{E}\big[Y|m,a'\big]\big\}p\big(m|a'\big)$$

which recovers the well-known *mediation formula* [12].

The independence assumptions which were necessary to derive this functional, namely $(Y(m,a) \perp\!\!\!\perp M(a') \perp\!\!\!\perp A)$, are implied by the MWM for the graph in Figure 2(a). It is possible to consider such assumptions independently of a graph. However, the advantage of graphs is their ability to encode assumptions of this type *systematically*, which allowed us to derive such functionals for a wide variety of problems, and moreover, to give simple visual characterizations of when such derivations are possible.

5.3. *Identification of path interventions.* As we saw in the previous section, identification of responses to edge interventions under the SWM requires node consistency, while any joint response to any edge intervention is identified under the MWM. In this section, we show that path interventions are identified under the MWM as long as *edge consistency* holds, that is, as long as a path intervention can be expressed as an edge intervention. Lack of edge consistency will result

in nonidentification under the MWM. The presence of a "recanting witness" in a path-specific effect [1] can be viewed as a special case of the lack of edge consistency.

A set of paths $\boldsymbol{\alpha}$ live for $\mathbf{Y}$ is called *consistent* for $\mathbf{Y}$ if for every edge $(AB)_{\rightarrow}$ that is an edge prefix of $\alpha \in \boldsymbol{\alpha}$, if $(AB)_{\rightarrow}$ is in $\beta \in \text{rel}_{\mathcal{G}}(\mathbf{Y}|\boldsymbol{\alpha})$, then $(AB)_{\rightarrow}$ is an edge prefix of a prefix subpath of β in $\boldsymbol{\alpha}$.

For a set of paths $\boldsymbol{\alpha}$ live and consistent for $\mathbf{Y}$, we call a path intervention $\pi_{\mathfrak{a}_{\boldsymbol{\alpha}}}$ *edge consistent* for $\mathbf{Y}$ if for every edge $(AB)_{\rightarrow}$, all paths in $\boldsymbol{\alpha}$ with $(AB)_{\rightarrow}$ as a prefix are assigned the same value (say a). Any path intervention that is not edge consistent we call edge inconsistent, including any path intervention on a set of paths not consistent for an outcome set of interest.

The path set $\{(WAMY)_{\rightarrow}\}$ in Figure 2(a) is live but not consistent for $\{Y\}$, thus any path intervention on this set is inconsistent for $\{Y\}$. A path intervention corresponding to $Y((wAMY)_{\rightarrow}, (wAY)_{\rightarrow})$ is edge consistent for Y, while a path intervention corresponding to $Y((wAMY)_{\rightarrow}, (w'AY)_{\rightarrow})$ is consistent for Y, but not edge consistent for Y.

For a path intervention $\pi_{\mathfrak{a}_{\boldsymbol{\alpha}}}$ edge consistent for $\mathbf{Y}$, define the set of edges $\boldsymbol{\alpha}_1 \equiv \{(AB)_{\rightarrow} | (AB)_{\rightarrow}$ is a prefix for $\alpha \in \boldsymbol{\alpha}\}$. Let $\eta_{\mathfrak{a}_{\boldsymbol{\alpha}_1}}$ be the induced edge intervention for $\pi_{\mathfrak{a}_{\boldsymbol{\alpha}}}$, where η assigns $(AB)_{\rightarrow} \in \boldsymbol{\alpha}_1$ to the value assigned by π to all $\alpha \in \boldsymbol{\alpha}$ which have $(AB)_{\rightarrow}$ as an edge prefix.

LEMMA 5.7. *Given a DAG $\mathcal{G}$ with vertices $\mathbf{V}$, and a path intervention $\pi_{\mathfrak{a}_{\boldsymbol{\alpha}}}$ edge consistent for $\mathbf{Y} \subseteq \mathbf{V}$, $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}})) = p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}_1}))$.*

COROLLARY 5.2. *If $\pi_{\mathfrak{a}_{\boldsymbol{\alpha}}}$ is edge consistent for $\mathbf{Y}$, then the distribution $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}}))$ is identified as a functional of $p(\mathbf{V})$ under the MWM model via the appropriate marginal of the edge g-formula for the response to the corresponding induced edge intervention.*

We will show that responses to edge inconsistent path interventions are not identifiable under the MWM using the same strategy as we used for node inconsistent edge interventions. First, we reproduce a result stating that a joint response to a conflicting exposure is not identifiable. Then we extend this result to the general case we need.

LEMMA 5.8. *The distributions $p(B(a), B(a'))$ and $p(B(a), B)$ are not identifiable from $p(A, B)$ under the MWM for the DAG in Figure 3(a).*

THEOREM 5.2. *Consider a DAG $\mathcal{G}$ with vertices $\mathbf{V}$, and a proper set of paths $\boldsymbol{\alpha}$ live for $\mathbf{Y}$. Then $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}}))$ is identifiable from $p(\mathbf{V})$ under the MWM for $\mathcal{G}$ if and only if $\pi_{\mathfrak{a}_{\boldsymbol{\alpha}}}$ is edge consistent. Moreover, if $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}}))$ is identifiable, it is equal to the appropriate marginal of the edge g-formula for $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}_1}))$, the response to the induced edge intervention.*

5.4. *A model where responses to path interventions are identified.* Though we have shown that responses to path interventions that cannot be expressed as responses to edge interventions are not in general identified under the MWM, there exist submodels of the MWM where all responses to path interventions are identified. In particular, consider the *linear structural equation model* (*SEM*), which is an MWM where the mapping from $\mathbf{v}_{\mathrm{pa}_{\mathcal{G}}(V)} \in \mathfrak{X}_{\mathrm{pa}_{\mathcal{G}}(V)}$ to $V(\mathbf{v}_{\mathrm{pa}_{\mathcal{G}}(V)})$ is a linear function of $\mathbf{v}_{\mathrm{pa}_{\mathcal{G}}(V)}$ and an error term ε_V, where such error terms are normally distributed and mutually independent.

THEOREM 5.3. *Let $\pi_{\mathfrak{a}_\alpha}$ be a path intervention. Then $p(\mathbf{Y}(\mathfrak{a}_\alpha))$ is identified under the linear SEM.*

This follows as a corollary of results in [2]. The reason even edge-inconsistent path interventions are identified is that linearity, normality and independence are such strong assumptions that we can directly evaluate even counterfactuals of the form $p(W(a), W(a'))$ using the algorithm in [2]. A fruitful open question if whether there are other interesting (for instance maximal) submodels of the MWM where all responses to path interventions are identified.

5.5. *Targets not representable as path interventions.* We have shown that a wide class of targets of interest in causal inference can be expressed as responses to path interventions. Nevertheless, there exist targets of interest which are known not to be representable in this way, such as principal stratification effects. For instance, the principal stratum direct effect (PSDE) [22, 23] is defined to be a treatment contrast only among those individuals for whom the mediator assumes a particular value for both active and baseline treatment levels. In Figure 2(a), the PSDE is a contrast of the form

$$\mathbb{E}\big[Y(a,m)|M(a)=M\big(a'\big)=m\big]-\mathbb{E}\big[Y\big(a',m\big)|M(a)=M\big(a'\big)=m\big].$$

Under the MWM, we obtain independences $Y(a,m) \perp\!\!\!\perp \{M(a), M(a')\}$, and $Y(a',m) \perp\!\!\!\perp \{M(a), M(a')\}$, which implies the PSDE is equal to the controlled direct effect contrast under the MWM: $\mathbb{E}[Y(a,m)] - \mathbb{E}[Y(a',m)]$. Under the SWM, the PSDE contrast is not identified without more assumptions. In either case, it is not possible to express the condition defining the principal strata, namely $M(a) = M(a') = m$ as a response to a path intervention, since this will entail assigning conflicting values to a directed edge from A to M. This is perhaps not surprising, since responses to path interventions are meant to encode effects *along particular causal pathways* which is not something principal strata effects encode. Note that despite this, the MWM allows us to rephrase the PSDE as a node intervention.

6. The edge g-formula and single world intervention graphs. A connection between the SWM, node interventions, the extended g-formula and a type of graph with split nodes called the Single World Intervention Graph (SWIG) was given in [13].

If a set of responses $\mathbf{V}$ to a node intervention $\nu_{\mathbf{a}}$ includes all variables (including $\mathbf{A}$), then, under the SWM, the response is linked to the observed distribution via (2), and can be viewed as a kind of Markov factorization [9] of the joint response $\mathbf{V}(\mathbf{a})$, where terms $p(V \mid \mathrm{pa}_{\mathcal{G}}(V))$ with $\mathrm{pa}_{\mathcal{G}}(V) \cap \mathbf{A} \neq \varnothing$ are replaced with $p(V \mid \mathrm{pa}_{\mathcal{G}}(V) \setminus \mathbf{A}, \mathbf{a}_{\mathrm{pa}_{\mathcal{G}}(V) \cap \mathbf{A}})$. SWIGs are a graphical representation of this factorization, in the sense that independences in $p(\mathbf{V}(\mathbf{a}))$ can be read off from the corresponding SWIG. Since A occurs both as a treatment and a response, SWIGs split the vertex A into a random and fixed versions (we draw fixed vertices as squares).

For example, the SWIG in Figure 4(a) represents $p(\{Y, M, W, A\}(a))$ in the SWM corresponding to Figure 2(a). We can check independences of counterfactuals in the joint $p(\{Y, M, W, A\}(a))$, via a simple modification of the d-separation criterion [9]. For instance, $Y(a) \perp\!\!\!\perp A \mid W$, since all d-connected paths from Y to A are blocked by W.

Similarly, if a set of responses $\mathbf{V}$ to an edge intervention $\eta_{\mathfrak{a}}$ includes all variables (including $\mathbf{A}$), then, under the MWM, the response is linked to the observed distribution via (2), and can be viewed as a kind of Markov factorization [9] of the joint response $\mathbf{V}(\mathfrak{a})$, where terms $p(V \mid \mathrm{pa}_{\mathcal{G}}(V))$ with $\mathrm{pa}_{\mathcal{G}}(V) \cap \mathrm{so}_{\mathcal{G}}(\boldsymbol{\alpha}) \neq \varnothing$ are replaced with $p(V \mid \mathrm{pa}_{\mathcal{G}}^{\overline{\boldsymbol{\alpha}}}(V), \mathfrak{a}_{(WV)_{\rightarrow} \in \boldsymbol{\alpha}})$. It is possible to generalize SWIGs to give a graphical representation of this factorization. Instead of splitting the vertices into the fixed and random versions, we instead shatter every intervened-on vertex into a set corresponding to distinct values (including the natural value) that vertex assumes when defining the response. For example, the graph in Figure 4(b) represents $p(\{Y, M, W, A\}((wM)_{\rightarrow}, (w'A)_{\rightarrow}, (aY)_{\rightarrow}))$ in the MWM corresponding to Figure 2(a). We can check independences of counterfactuals in this joint via a simple modification of d-separation: $Y((aY)_{\rightarrow}, (wM)_{\rightarrow}, (w'A)_{\rightarrow}) \perp\!\!\!\perp A((w'A))|M((w'A)_{\rightarrow}, (wM)_{\rightarrow})$ since all d-connected paths from Y to A are blocked by M. Note that we shatter W in Figure 2(a) into three vertices, and A into two, where the random vertex has an outgoing arrow to M. This is because there are two treatment values for W, and W is also a response, while A is a response for

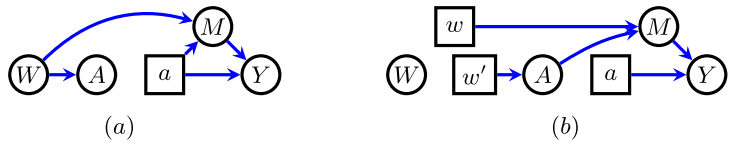

FIG. 4. (a) *A SWIG for* $\{Y, M, W, A\}(a)$. (b) *An edge intervention version for* $\{Y, M, W, A\}((wM)_{\rightarrow}, (w'A)_{\rightarrow}, (aY)_{\rightarrow})$.

the purposes of the $(AM)_{\rightarrow}$ edge and a treatment for the purposes of the $(AY)_{\rightarrow}$ edge. That responses to edge interventions factorize according to these kinds of "shattered graphs" under the MWM (but not SWM) follows as a straightforward generalization of the proof of proposition 11 in [13]. In fact, these shattered graphs can be viewed as SWIGs defined on an augmented graph where a treatment vertex is split into copies, corresponding to (individually intervenable) components of the treatment associated with direct and indirect effects. For examples of such graphs, and associated discussion; see [19], Section 6 and Figure 6.

Thus, the edge g-formula can be viewed as the MWM analogue of the extended g-formula, and it is possible to construct graphs that stand in the same relation to edge interventions, the edge g-formula, and the MWM as SWIGs do to node interventions, the extended g-formula and the SWM. In the interest of space, we do not derive this formally, nor pursue this connection further here.

7. The edge g-formula and causal effects in hidden variable DAGs. If some variables in a causal model of a DAG are unobserved, not every response to a node intervention is identified, since (2) cannot always be directly applied. A complete theory for identifying $\mathbf{Y}(\mathbf{a})$ from $p(\mathbf{V})$, where $\mathbf{A}$ and $\mathbf{Y}$ are disjoint, was given in [25, 33]. In this section, we show that certain identifying functionals for $p(\mathbf{Y}(\mathbf{a}))$ correspond to marginals of the edge g-formula (5).

For example, it can be shown that $p(Y(a))$ is identified via the *front-door functional* $\sum_{m,a'} p(Y|a',m)p(m|a)p(a')$ under the SWM shown in Figure 5(a), where H is not observable. If we replace H and its outgoing arrows by an arrow from A to Y, we obtain the DAG in Figure 5(c). A straightforward consequence of (5) is that $p(Y((aM)_{\rightarrow}))$ is identified via the same functional under the MWM for Figure 5(c). In this section, we give a general condition for case when this correspondence of functionals occurs.

We introduce additional terminology to help us formulate our results. Rather than defining the identification problem on hidden variable DAGs directly, we will define it on acyclic directed mixed graphs (ADMGs) which represent a class of hidden variable DAGs that all share identifying functionals. An ADMG is a mixed

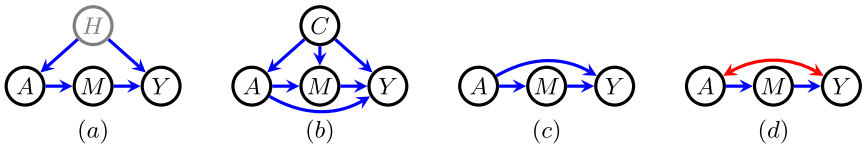

FIG. 5. (a) *A hidden variable DAG where the causal effect* $p(Y(a) = y)$ *is identified via the front–door formula* $\sum_{m,a'} p(y|a',m)p(m \mid a)p(a')$. (b) *A DAG for a simple setting in mediation analysis where multiply robust estimators for functionals derived from* (5) *for* $Y((aY)_{\rightarrow}, (a'M)_{\rightarrow})$ *are known.* (c) *A DAG where* $Y((aM)_{\rightarrow})$ *is identified via the front-door formula in* (a). (d) *A latent projection ADMG of the DAG in* (a) *onto* $\{A, M, Y\}$.

graph with directed ($\rightarrow$) and bidirected edges ($\leftrightarrow$), with no directed cycles. ADMGs represent classes of hidden variable DAGs via a latent projection operation onto a graph defined only on observed variables [34]. For example, this operation applied to Figure 5(a) results in an ADMG shown in Figure 5(d). Connected components in a graph obtained from an ADMG $\mathcal{G}$ by dropping all directed edges are called *districts* of $\mathcal{G}$. For example, the sets $\{A, Y\}$ and $\{M\}$ are districts of the graph in Figure 5(d). The set of districts of $\mathcal{G}$ is denoted by $\mathcal{D}(\mathcal{G})$. If a set $\mathbf{S}$ is in a district of $\mathcal{G}$, we denote that district by $\text{dis}_{\mathcal{G}}(\mathbf{S})$.

For an ADMG $\mathcal{G}$ with vertices $\mathbf{V}$, and $\mathbf{A} \subseteq \mathbf{V}$, let $\mathcal{G}_{\mathbf{A}}$ be a subgraph consisting only of vertices in $\mathbf{A}$ and edges between them. Let $\text{an}_{\mathcal{G}}(V) = \{A | A \rightarrow \cdots \rightarrow V \text{ is in } \mathcal{G}\}$. A total order $\prec_{\mathcal{G}}$ on $\mathbf{V}$ in $\mathcal{G}$ is *topological* if whenever $V_1 \prec_{\mathcal{G}} V_2$, $V_2 \notin \text{an}_{\mathcal{G}}(V_1)$. For a total order $\prec$ on $\mathbf{V}$, for any $V \in \mathbf{V}$, let $\text{pre}_{\prec}(V) \equiv \{W \in \mathbf{V} \setminus \{V\} | W \prec V\}$.

In the remainder of this section, we will, without loss of generality, restrict attention to identification problems where $\mathbf{V} \subseteq \text{an}_{\mathcal{G}}(\mathbf{Y})$, and $\mathbf{V} \setminus \text{an}_{\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}}}(\mathbf{Y}) \subseteq \mathbf{A}$, and consider responses to node interventions that yield an identifying functional in a particular convenient form that only involves conditional distributions derived from $p(\mathbf{V})$.

DEFINITION 2. Given $p(\mathbf{V})$, for any total order $\prec$ on $\mathbf{V}$, and $\mathbf{v} \in \mathfrak{X}_{\mathbf{V}}$, a functional of $p(\mathbf{V})$ of the form $\sum_{\mathbf{V}' \setminus \mathbf{Y}} \prod_{V \in \mathbf{V}'} p(V | \mathbf{S}_V, \mathbf{v}_{\text{pre}_{\prec}(V) \setminus \mathbf{S}_V})$, where $\mathbf{Y} \subseteq \mathbf{V}' \subseteq \mathbf{V}$, and $\mathbf{S}_V \subseteq \text{pre}_{\prec}(V) \cap \mathbf{V}'$ is called a *g-functional*.

The output of the g-computation algorithm [14], mentioned in Section 2.4, is always a g-functional, but some identifying functionals for responses to node interventions are g-functionals that cannot arise from g-computation. For instance, consider the front-door functional $\sum_{m,a'} p(Y | a', m) p(m | a) p(a')$, which identifies $p(Y(a))$ in the graph in Figure 5(a). If we let $\mathbf{V} = \mathbf{V}' = \{A, M, Y\}$, $\mathbf{Y} = \{Y\}$, take $\prec$ to be the topological ordering for the graph, and let $\mathbf{S}_Y = \{M, A\}$ and $\mathbf{S}_M = \mathbf{S}_A = \{\}$, we see that this satisfies Definition 2 and so is a functional. However, g-computation cannot be used to identify responses to node interventions in cases where intervened on variables have unobserved causes in common with responses, as is the case in Figure 5(a).

We give a sufficient condition on ADMGs $\mathcal{G}$, and on response sets $\mathbf{Y}$ to node interventions on $\mathbf{A}$, such that the output is a g-functional, and then show that it is possible to construct a DAG $\mathcal{G}^{\dagger}$ from $\mathcal{G}$ where a certain response to an edge intervention is identified via the same g-functional via (5).

For a particular treatment set $\mathbf{A}$ in $\mathcal{G}$, let $\mathbf{D}_{\mathbf{S},\mathbf{A},\mathcal{G}} = \text{dis}_{\mathcal{G}_{\text{an}_{\mathcal{G}}(\mathbf{S})}}(\mathbf{S})$ for each $\mathbf{S} \in \mathcal{D}(\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}})$. We will omit $\mathbf{A}$ and $\mathcal{G}$ from the subscript if they are obvious, to yield $\mathbf{D}_{\mathbf{S}}$, and let $\mathbf{A}_f = \mathbf{A} \setminus \bigcup_{\mathbf{S} \in \mathcal{D}(\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}})} \mathbf{D}_{\mathbf{S}}$. In words, $\mathcal{D}(\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}})$ is the districts in a graph where treatments $\mathbf{A}$ are removed. For instance, in Figure 5(d), with treatment A, these districts are $\{M\}$ and $\{Y\}$. For each such district $\mathbf{S}$, $\mathbf{D}_{\mathbf{S}}$ is a (possibly larger)

district containing all of $\mathbf{S}$ in a graph containing ancestors of $\mathbf{S}$. For instance, $\mathbf{D}_{\{Y\}}$ in Figure 5(d) is $\{A, Y\}$. $\mathbf{A}_f$ is all treatments not in any such $\mathbf{D}_\mathbf{S}$. Since A is the only treatment in Figure 5(d) and is in $\mathbf{D}_{\{Y\}}$, $\mathbf{A}_f = \{\}$ in this case.

LEMMA 7.1. *If* $(\forall \mathbf{S}_1, \mathbf{S}_2 \in \mathcal{D}(\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}}))(\mathbf{D}_{\mathbf{S}_1} \cap \mathbf{D}_{\mathbf{S}_2} \neq \varnothing) \Rightarrow (\mathbf{S}_1 = \mathbf{S}_2)$, *then the sets* $\{\mathbf{D}_\mathbf{S} | \mathbf{S} \in \mathcal{D}(\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}})\}$ *partition* $\mathbf{V} \setminus \mathbf{A}_f$.

Given Lemma 7.1, for every $V \in \mathbf{V} \setminus \mathbf{A}_f$, let $\mathbf{D}_V = \mathbf{D}_\mathbf{S}$ for the unique $\mathbf{D}_\mathbf{S}$ such that $V \in \mathbf{D}_\mathbf{S}$.

The following lemma gives two conditions sufficient to yield an identifying g-functional. First, any district $\mathbf{S} \in \mathcal{G}_{\mathbf{V}\setminus\mathbf{A}}$ must not have parents not in $\mathbf{S}$ as elements of $\mathbf{D}_\mathbf{S}$, and second the sets $\mathbf{D}_\mathbf{S}$ must partition $\mathbf{V} \setminus \mathbf{A}_f$ as in Lemma 7.1. This is satisfied by Figure 5(d), since $\mathbf{D}_{\{Y\}} = \{Y, A\}$, $\mathbf{D}_{\{M\}} = \{M\}$, and $\text{pa}_\mathcal{G}(\{M\}) = \{A\}$, $\text{pa}_\mathcal{G}(\{Y\}) = \{M\}$.

LEMMA 7.2. *Fix* $\mathbf{A}, \mathbf{Y}, \mathcal{G}$ *such that*:

1. $(\forall \mathbf{S} \in \mathcal{D}(\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}}))$, $(\text{pa}_\mathcal{G}(\mathbf{S}) \setminus \mathbf{S}) \cap \mathbf{D}_\mathbf{S} = \varnothing$, *and*
2. $(\forall \mathbf{S}_1, \mathbf{S}_2 \in \mathcal{D}(\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}}))$, $(\mathbf{D}_{\mathbf{S}_1} \cap \mathbf{D}_{\mathbf{S}_2} \neq \varnothing) \Rightarrow (\mathbf{S}_1 = \mathbf{S}_2)$.

Then $p(\mathbf{Y}(\mathbf{a}) = \mathbf{y})$ *is identified by a g-functional*

$$(6) \qquad \sum_{\mathbf{v}\mathbf{V}\setminus(\mathbf{Y}\cup\mathbf{A}_f)} \prod_{V\in\mathbf{V}\setminus\mathbf{A}_f} p\big((\mathbf{y}\cup\mathbf{v})_V | \mathbf{a}_{\text{pre}_{\prec_\mathcal{G}}(V)\cap(\mathbf{A}\setminus\mathbf{D}_V)}, (\mathbf{y}\cup\mathbf{v})_{\text{pre}_{\prec_\mathcal{G}}(V)\setminus(\mathbf{A}\setminus\mathbf{D}_V)}\big).$$

Finally, given that preconditions given by Lemma 7.2 are satisfied by an ADMG $\mathcal{G}$, the following result claims we can modify $\mathcal{G}$ into a DAG $\mathcal{G}^\dagger$, where there is some edge intervention with a response identified by the same g-functional as given by lemma 7.2. This DAG for Figure 5(d) is Figure 5(c).

LEMMA 7.3. *For an ADMG* $\mathcal{G}$ *with vertex set* $\mathbf{V}$, *fix disjoint* $\mathbf{Y}, \mathbf{A} \subseteq \mathbf{V}$ *that satisfy the preconditions of Lemma 7.2. Then there exists a DAG* $\mathcal{G}^\dagger$ *with vertex set* $\mathbf{V}$, *and an edge intervention* $\eta_{\mathfrak{a}_{\boldsymbol{\alpha}}}$ *on a set of edges* $\boldsymbol{\alpha}$ *in* $\mathcal{G}^\dagger$ *such that* $p(\mathbf{Y}(\mathfrak{a}_{\boldsymbol{\alpha}}))$ *is identified under the MWM for* $\mathcal{G}^\dagger$ *via a margin of the functional in* (5) *that is equal to the identifying g-functional for* $p(\mathbf{Y}(\mathbf{a}))$ *in terms of* $p(\mathbf{V})$ *in* $\mathcal{G}$.

A natural question raised by Lemma 7.3 is the converse—is it the case that every identifying functional for an edge intervention corresponds to an identifying functional of a response to a node intervention. We leave this question for future work.

The fact that a class of causal effects identified via a g-functional, even those effects with unobserved causes of treatments, corresponds to responses to edge interventions in a DAG gives an additional reason to study estimation theory of the edge g-formula (5). Furthermore, this connection gives another setting in which front-door type functionals may arise—the context of mediation analysis where the baseline treatment is not a constant value, but a naturally occurring value in the population.

8. A multiply robust estimator for a special case of the edge g-formula. We have shown that the edge g-formula (5) encodes a wide class of identified targets in causal inference. Here, we give an example of how a response to an edge-consistent path intervention is represented as an edge intervention, identified via a marginal of (5), and reexpressed as a contrast parameter for which an estimator exists which is robust to misspecification of parts of the likelihood function. We consider discrete state spaces, but extensions to continuous state spaces are straightforward in this case.

Consider the graph in Figure 5(b), which represents a simple mediation setting, with A an exposure, Y an outcome, M a mediator and C a set of baseline covariates. We might be interested in a direct or indirect effect of A on Y. As discussed in Section 4, we may represent such effects as contrasts obtained from a response to a path intervention $p(Y((aMY)_{\to}, (a'Y)_{\to}))$. This path intervention is natural, and edge consistent, and the response of Y to it is equal to the response to an edge intervention $p(Y((aM)_{\to}, (a'Y)_{\to}))$, which is identified as a marginal of (5), namely $\sum_c p(Y|a', m, c)p(m|a, c)p(c)$. Let $\Upsilon(a, a', c) = \sum_m \mathbb{E}(Y|a', m, c) \cdot p(m|a, c)$. Then the mean response is $\Phi(a, a') = \sum_c \Upsilon(a, a', c) \cdot p(c)$, and the *efficient influence function* of $\Phi(a, a')$ under the saturated model $\mathcal{P}_s$, that is, the set of all densities $p(Y, A, M, C)$, is

$$\begin{aligned} U^{\text{eff}}_{\mathcal{P}_s}\big(\Phi(a, a')\big) = {} & \frac{\mathbb{I}(A = a)p(M|a', C)}{p(a|C)p(M|a, C)}\big\{Y - \mathbb{E}(Y|C, M, a)\big\} \\ & + \frac{\mathbb{I}(A = a')}{p(a'|C)}\big\{\mathbb{E}(Y|C, M, a) - \Upsilon\big(a, a', C\big)\big\} \\ & + \Upsilon\big(a, a', C\big) - \Phi\big(a, a'\big), \end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function for an event [31].

To represent direct and indirect effects as contrasts, we also need to consider the response of Y to A being set to a for the purposes of all pathways from A to Y, which simply corresponds to $p(Y(a))$, which is identified via a marginal of (2), namely $\sum_c p(Y|a, c)p(c)$. The mean response is then $\Phi(a, a) = \sum_c \mathbb{E}(Y|a, c)p(c)$. The efficient influence function of $\Phi(a, a)$ under the saturated model $\mathcal{P}_s$ is simply $U^{\text{eff}}_{\mathcal{P}_s}(\Phi(a, a))$, which simplifies to

$$\frac{\mathbb{I}(a)}{p(a|C)}\big\{Y - \Upsilon(a, a, C)\big\} + \Upsilon(a, a, C) - \Phi(a, a),$$

the efficient influence function derived in the context of total effects in [20].

Natural direct and indirect effects may be defined on the difference scale as $\Phi(a, a) - \Phi(a, a')$, and $\Phi(a, a') - \Phi(a', a')$. Alternatively, for binary outcomes we may also define such effects in a natural way on the risk ratio or odds ratio scale.

Estimating these parameters using an unrestricted likelihood is not a feasible strategy in settings with a high dimensional vector of baseline covariates, which

means we must resort to modeling. An approach in [31] is to assume models $\{\mathbb{E}^{\text{par}}(Y|a,m,c;\hat{\alpha}),\ f^{\text{par}}(m|a,c;\hat{\beta}),\ f^{\text{par}}(a|c;\hat{\gamma})\}$, and use a substitution estimator which solves the estimating equations

$$\mathbb{P}_n\big(\hat{U}^{\text{eff}}_{\mathcal{P}_s}(\Phi(a,a'))\big)=0,$$

where $\mathbb{P}_n(\cdot)$ is the empirical average (for sample size n), and $\hat{U}^{\text{eff}}_{\mathcal{P}_s}$ is equal to $U^{\text{eff}}_{\mathcal{P}_s}$ evaluated at $\{\mathbb{E}^{\text{par}}(Y|a,m,c;\hat{\alpha}),\ f^{\text{par}}(m|a,c;\hat{\beta}),\ f^{\text{par}}(a|c;\hat{\gamma})\}$.

The resulting estimator exhibits the property of *triple robustness*, that is, it remains consistent in the union model where any two of the above three parametric models is correct. This estimator is combined with a similarly defined doubly robust estimator for $\Phi(a,a)$ derived in [20] to yield a triply robust estimator for the direct and indirect parameters on the difference scale. This was extended to the semi-parametric models for direct effects on the additive and multiplicative scales [32].

Since our results show that the edge g-formula encompasses a wide range of causal inference targets, including effects of treatments on the multiply treated, path-specific effects and causal effects with unobserved causes of treatments, an interesting avenue of future work is to generalize estimation theory for simple instances of the edge g-formula, like above, to more general cases, for instance longitudinal cases like that shown in Figure 2(b).

9. Discussion. We have defined an inclusion hierarchy of interventions associated with graphical features: node interventions corresponding to standard treatment interventions, edge interventions corresponding to intervening on a portion of the treatment mechanism associated with a particular outgoing edge and path interventions corresponding to intervening on a portion of the treatment mechanism associated with a particular outgoing causal pathway. We have shown that a variety of causal inference targets of interest, including effects of treatment on the multiply treated, and path-specific effects can be viewed as special cases of responses to path interventions. In addition, we have shown that edge interventions are in some sense naturally associated with the MWM of Pearl as the responses to such interventions are naturally identified under the assumptions of this model, just as node interventions are naturally associated with the SWM of Robins. The question of whether a particular causal inference target is identified, and under what model thus reduces to expressing the target as a path intervention, and then considering whether the path intervention is natural, and whether it can be reexpressed as an edge intervention or a node intervention. This process is summarized in a flowchart shown in Figure 6.

An obvious extension of our work is to consider identification of responses in our hierarchy in hidden variable DAG models in terms of observed marginal distributions. Existing results on mediation analysis [24] and ETT identification [28] would be subsumed as special cases under this framework, but it would entail

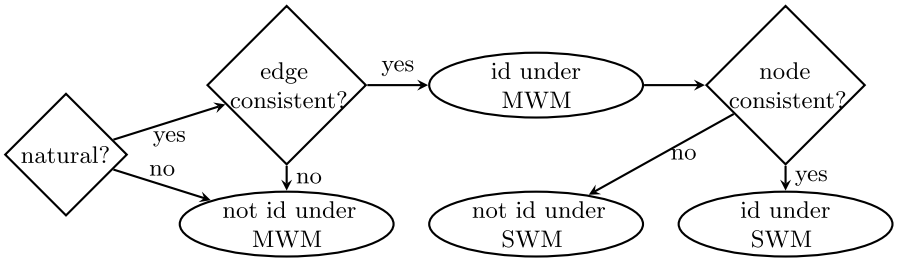

FIG. 6. *A flowchart for identification results for path interventions under the MWM and the SWM.*

novel identification results for any new target expressible as a path intervention response. In addition, an interesting question is whether *all* identifying functionals for responses to node interventions in a hidden variable DAG model correspond to some sort of identified response to an edge intervention, although possibly not in a DAG but an ADMG. If true, this would recast *any* identified causal effect as a certain type of identified mediated effect.

While estimation theory of functionals derived from the extended g-formula (2) has received attention in the literature [18], multiply robust estimators for functionals obtained from the edge g-formula (5) are known only in very special cases such as the point treatment setting we discussed in Section 8 [31]. As we have shown in this paper, developing estimators for general functionals obtained from the edge g-formula (5) results in estimators for a wide class of targets of interest in causal inference, including path-specific effects, effects of treatment on the multiply treated, effects of treatments on the indirectly treated and causal effects in the presence of unobserved causes of treatments.

Our results thus not only provide a unifying view of identification, under various models, of a large class of targets of interest in causal inference, but also motivate the development of estimation theory for a more general functional than the g-formula.

SUPPLEMENTARY MATERIAL

Supplement to "Causal inference with a graphical hierarchy of interventions" (DOI: 10.1214/15-AOS1411SUPP; .pdf). Our supplementary materials contain detailed arguments for most of our claims, and some auxiliary definitions. In addition, we provide a detailed rationale for the use of path interventions.

REFERENCES

[1] AVIN, C., SHPITSER, I. and PEARL, J. (2005). Identifiability of path-specific effects. In *Proceedings of the International Joint Conference on Artificial Intelligence* **19** 357–363. Morgan Kaufmann, San Francisco, CA.

[2] BALKE, A. and PEARL, J. (1994). Probabilistic evaluation of counterfactual queries. In *Proceedings of the Conference on Artificial Intelligence* **12** 230–237. Morgan Kaufmann, San Francisco, CA.

[3] BARON, R. M. and KENNY, D. A. (1986). The moderator-mediator variable distinction in social psychology research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51** 1173–1182.
[4] HUBBARD, A. E. and VAN DER LAAN, M. J. (2008). Population intervention models in causal inference. *Biometrika* **95** 35–47. MR2409713
[5] IMAI, K., TINGLEY, D. and YAMAMOTO, T. (2013). Experimental designs for identifying causal mechanisms. *J. Roy. Statist. Soc. Ser. A* **176** 5–51. MR3042176
[6] MOODIE, E. E. M., RICHARDSON, T. S. and STEPHENS, D. A. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics* **63** 447–455. MR2370803
[7] MURPHY, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 331–366. MR1983752
[8] NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 463–472.
[9] PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*: *Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA. MR0965765
[10] PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence* **17** 411–420. Morgan Kaufmann, San Francisco, CA.
[11] PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166
[12] PEARL, J. (2011). The causal mediation formula—A guide to the assessment of pathways and mechanisms. Technical Report R-379, Cognitive Systems Laboratory, Univ. California, Los Angeles.
[13] RICHARDSON, T. S. and ROBINS, J. M. (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Preprint. Available at http://www.csss.washington.edu/Papers/wp128.pdf.
[14] ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Modelling* **7** 1393–1512. MR0877758
[15] ROBINS, J. M. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J. Chronic. Dis.* **40** 139–161.
[16] ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability of direct and indirect effects. *Epidemiology* **3** 143–155.
[17] ROBINS, J. M., HERNÁN, M. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.
[18] ROBINS, J. M., HERNÁN, M. A. and SIEBERT, U. (2004). Effects of multiple interventions. In *Comparative Quantification of Health Risks*: *Global and Regional Burden of Disease Attributable to Selected Major Risk Factors* **2** 2191–2230. World Health Organization, Geneva.
[19] ROBINS, J. M. and RICHARDSON, T. S. (2010). Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology*: *Finding the Determinants of Disorders and Their Cures* (P. Shrout, ed.). Oxford University Press, Oxford.
[20] ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730
[21] RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.
[22] RUBIN, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scand. J. Stat.* **31** 161–170. MR2066246
[23] RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. MR2166071

[24] SHPITSER, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science* (*Rumelhart Special Issue*) **37** 1011–1035.
[25] SHPITSER, I. and PEARL, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *National Conference on Artificial Intelligence* **21**. AAAI Press, Palo Alto, CA.
[26] SHPITSER, I. and PEARL, J. (2006). Identification of conditional interventional distributions. In *Uncertainty in Artificial Intelligence* **22**. AUAI Press, Corvallis, OR.
[27] SHPITSER, I. and PEARL, J. (2008). Complete identification methods for the causal hierarchy. *J. Mach. Learn. Res.* **9** 1941–1979. MR2447308
[28] SHPITSER, I. and PEARL, J. (2009). Effects of treatment on the treated: Identification and generalization. In *Uncertainty in Artificial Intelligence* **25**. AUAI Press, Corvallis, OR.
[29] SHPITSER, I. and TCHETGEN TCHETGEN, E. (2015). Supplement to "Causal inference with a graphical hierarchy of interventions." DOI:10.1214/15-AOS1411SUPP.
[30] SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (1993). *Causation, Prediction, and Search*. *Lecture Notes in Statistics* **81**. Springer, New York. MR1227558
[31] TCHETGEN TCHETGEN, E. J. and SHPITSER, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *Ann. Statist.* **40** 1816–1845. MR3015045
[32] TCHETGEN TCHETGEN, E. J. and SHPITSER, I. (2014). Estimation of a semiparametric natural direct effect model incorporating baseline covariates. *Biometrika* **101** 849–864. MR3286921
[33] TIAN, J. and PEARL, J. (2002). On the testable implications of causal models with hidden variables. In *Uncertainty in Artificial Intelligence* **18** 519–527. AUAI Press, Corvallis, OR.
[34] VERMA, T. S. and PEARL, J. (1990). Equivalence and synthesis of causal models. Technical Report R-150, Dept. Computer Science, Univ. California, Los Angeles.
[35] WRIGHT, S. (1921). Correlation and causation. *J. Agric. Res.* **20** 557–585.
[36] YOUNG, J. G., HERŃAN, M. A. and ROBINS, J. M. (2014). Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. *Epidemiol Method* **3** 1–19.

DEPARTMENT OF COMPUTER SCIENCE
JOHNS HOPKINS UNIVERSITY
3400 N. CHARLES STREET
BALTIMORE, MARYLAND 21218
USA
E-MAIL: ilyas@cs.jhu.edu

SCHOOL OF PUBLIC HEALTH
HARVARD UNIVERSITY
677 HUNTINGTON AVENUE
KRESGE BUILDING
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: etchetge@hsph.harvard.edu