Received: 6 March 2020 | Accepted: 28 March 2021

DOI: 10.1111/rssa.12699

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

ORIGINAL ARTICLE

Estimation of causal effects with small data in the presence of trapdoor variables

Jouni Helske | **Santtu Tikka** | **Juha Karvanen**

Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

Correspondence
Jouni Helske, Department of Mathematics and Statistics, University of Jyväskylä, Finland.
Email: jouni.helske@jyu.fi

Funding information
Academy of Finland, Grant/Award Number: 311877

Abstract

We consider the problem of estimating causal effects of interventions from observational data when well-known back-door and front-door adjustments are not applicable. We show that when an identifiable causal effect is subject to an implicit functional constraint that is not deducible from conditional independence relations, the estimator of the causal effect can exhibit bias in small samples. This bias is related to variables that we call *trapdoor variables*. We use simulated data to study different strategies to account for trapdoor variables and suggest how the related trapdoor bias might be minimized. The importance of trapdoor variables in causal effect estimation is illustrated with real data from the Life Course 1971–2002 study. Using this data set, we estimate the causal effect of education on income in the Finnish context. Bayesian modelling allows us to take the parameter uncertainty into account and to present the estimated causal effects as posterior distributions.

KEYWORDS

Bayesian estimation, bias, causality, functional constraint, identifiability

1 | INTRODUCTION

Understanding causal relations forms the basis of decision-making in society. The role of statistics is to provide tools that allow us to estimate the causal effects of planned interventions. Instead of

modelling just associations, a causal model describes the functional relationships present in the system of interest. A core feature of causal models (Pearl, 2009) is the capability to represent the effects of actions on the model via symbolic interventions. These interventions are assumed modular in the sense that they only modify the target of the intervention leaving other causal mechanisms in the model intact. The resulting probability distribution of this post-interventional causal model is defined as the causal effect of the intervention.

Various methods have been suggested for estimating causal effects in different settings. Propensity score matching (Imbens, 2000; Rosenbaum & Rubin, 1983), inverse probability weighting (Rosenbaum, 1987; Rosenbaum & Rubin, 1983) and *g*-methods (Robins et al., 1992) are some of the most well-known methods. These methods try to mimic a randomized experiment by creating pseudo-samples or by weighting the original sample in various ways, and it is typically assumed that there are no unobserved confounders present. In the classical structural equation modelling (SEM) approach (see, e.g. Kline, 2011), observed variables are assumed to be Gaussian, and unobserved confounders are treated as correlations between observed variables (there are also some extensions, such as LiNGAM for linear non-Gaussian cases (Shimizu, 2014)).

Another approach to causal inference is based on causal graphs, where the notion of identifiability plays a central role (see Section 2). For certain graphs, the well-known back-door and front-door adjustment criteria (Pearl, 1995) allow for identification of interventional distributions from observational data in the presence of unobserved confounders. The back-door criterion tells us whether a set of variables forms an admissible set, so that we only need to condition on these variables when estimating the causal effect, whereas the similar front-door criterion can be applied in the presence of a mediator between the interventional variable and the response variable. More generally, do-calculus (Pearl, 1995) can be used to assess whether the interventional distribution of interest is identifiable given only the known causal graph, without any parametric assumptions about the distributions of the variables or the form of the effects they have on each other. Do-calculus provides an identifying functional that is a nonparametric formula for the interventional distribution consisting of terms that represent observational distributions. Although identifiability does not in general guarantee estimability (Maclaren & Nicholson, 2019), it is usually possible to obtain an estimator for the causal effect by replacing the terms present in the identifying functional by suitable parametric or nonparametric estimators and then combining the results accordingly.

We study the estimation of causal effects with small data in scenarios where standard adjustment criteria are not applicable. By small data, we refer to a case where parameter estimation exhibits non-negligible uncertainty due to the sample size. In addition, we consider the presence of functional equality constraints known as *Verma-constraints* (Robins, 1986; Tian & Pearl, 2002; Verma & Pearl, 1990). Under certain conditions (see Section 2), these constraints are related to special variables that we call *trapdoor variables*. These variables can bias the causal effect estimator for finite samples and we refer to this form of bias as *trapdoor bias*. We demonstrate the practical ramifications of trapdoor variables for the estimation of causal effects via simulations in a number of synthetic scenarios with small sample sizes and compare a variety of estimation strategies in both nonparametric and parametric settings.

As a motivating example, we consider Bayesian estimation of the causal effect of education on yearly income using real data from the Finnish Life Course 1971–2002 study (Kuusinen, 2018). We construct a causal model for this study where we take into account the grade point average (GPA) from primary school, language skills, gender and the socioeconomic status (SES) of the parents. In the causal model, we find that GPA is in fact a trapdoor variable due to a functional equality constraint on the causal effect of interest. Bayesian modelling allows us to estimate the full post-interventional distribution of the income on different levels of education, which indicate a clear positive causal effect of education on income. We combine Bayesian estimation with a specialized Monte Carlo approach in order to take the effect of the

trapdoor variable into account in a number of scenarios including the life course model. All analysis was done in the R environment (R Core Team, 2020), and the codes for the simulation experiments, the life course example, and the figures for the simulation results (created with the `ggplot2` package (Wickham, 2016)) of this paper are available at https://github.com/helske/trapdoor.

The paper is structured as follows. Section 2 introduces the notation, gives a definition for the trapdoor variables and present examples on causal models where such variables are present. Section 3 focuses on various aspects related to the estimation of causal effects in the presence of trapdoor variables including a Bayesian approach and demonstrates how a trapdoor variables manifests under a linear-Gaussian model. Section 4 considers the effect of trapdoor variables and the trapdoor bias of causal effect estimators via simulation in a model with binary variables, a linear-Gaussian model and a synthetic scenario based on the life course model. The analysis using the real life course data is presented in Section 5. Section 6 provides some concluding remarks.

2 | THEORY

2.1 | Notation and basic definitions

Our analysis is based on the framework of *structural causal models* (SCM) and directed graphs, and we assume the reader to be familiar with these concepts and their core probabilistic and graphical properties. For a more detailed discussion on SCMs and graph theoretic concepts, we refer the reader to works such as (Pearl, 2009) and (Koller & Friedman, 2009).

We use capital letters to denote variables (V) and small letters to denote their values (v). Bold letters are used to denote sets of variables ($\mathbf{V}$) and value assignments ($\mathbf{v}$). The set of all possible value assignments to $\mathbf{V}$ is denoted by $val(\mathbf{V})$. Set difference of sets $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\mathbf{A} \setminus \mathbf{B}$. We use shorthand notation $P(Y|x)$ to denote the probability distribution $P_\theta(Y \mid X = x)$ where we typically omit the dependence of (unknown) model parameters θ.

Each SCM M over a set of variables $\mathbf{V}$ is associated a joint probability distribution $P(\mathbf{V})$ in a population of interest and a *causal graph* G over $\mathbf{V}$ where directed edges between two observed variables in $\mathbf{V}$ correspond to direct causal relationships which are assumed to not form any cycles. Bidirected edges between two observed variables in $\mathbf{V}$ are used to denote confounding by an unobserved common cause. In this framework, interventions are represented using the do$(\cdot)$-operator. An intervention do$(\mathbf{X} = \mathbf{x})$ forces the variables in $\mathbf{X}$ to take the values specified by $\mathbf{x}$ while leaving other mechanisms of the model intact. This intervention induces a submodel $M_{\mathbf{x}}$ with the interventional distribution $P(\mathbf{V}|\text{ do}(\mathbf{X} = \mathbf{x}))$. A causal effect $P(\mathbf{Y}|\text{do}(\mathbf{X} = \mathbf{x}))$ is said to be *identifiable* in G if it is uniquely computable from $P(\mathbf{V})$ in any SCM that induces G. A variable Y in the post-intervention model $M_{\mathbf{x}}$ is denoted as $Y(\mathbf{x})$. For an identifiable causal effect, an *identifying functional* is a function f such that $f(P(\mathbf{V})) = P(\mathbf{Y}|\text{ do}(\mathbf{X} = \mathbf{x}))$. We assume that $P(\mathbf{V} = \mathbf{v}) > 0$ for values $\mathbf{v} \in val(\mathbf{V})$ making all conditional distributions and interventions well-defined.

Note that identifiability only indicates the existence of an estimator and does not take into account the potential problems stemming from finite data. Therefore, even though determining the identifiability of $P(\mathbf{Y}|\text{ do}(\mathbf{X} = \mathbf{x}))$ is an important first step, it does not guarantee that we can estimate the causal effect in practice without additional stability assumptions (Maclaren & Nicholson, 2019). Note also that through this paper, we assume that our causal graph is correct, while in practice, we are rarely certain of this.

Interventional distributions exhibit conditional independence constraints, which can be characterized by d-separation (Pearl, 1988) in the associated causal graph of the model. However, identifiable causal effects can be subject to Verma-constraints. As an example of such a constraint, we show that

in the causal graph of Figure 1, a causal effect does not depend on the value of a variable W despite it appearing in the identifying functional of the interventional distribution. The causal effect of X on Y is identifiable in this graph which can be verified using do-calculus or by applying an identifiability algorithm such as the one by Huang and Valtorta (2006) or the ID algorithm by Shpitser and Pearl (2006). Application of the ID algorithm implemented in the `R` package `causaleffect` (Tikka & Karvanen, 2017) provides the formula

$$P(Y|\ \mathrm{do}(X = x)) = \sum_z P(Y|x,\ z,\ w)P(z|w),$$

where the left-hand side depends on the value of X and Y, but on the right-hand the variable W is also present and it is not subject to summation, unlike the variable Z. However, the right-hand side cannot depend on the value of W, as there is an admissible set Z which gives us an alternative formula

$$P(Y|\ \mathrm{do}(X = x)) = \sum_z P(Y|x,\ z)P(z),$$

that is a simple back-door formula where variable W is not present.

Perhaps surprisingly, Verma-constraints can be expressed as conditional independences in the interventional distribution (Shpitser & Pearl, 2008) and can be used to give an alternative definition for nested Markov models in acyclic directed mixed graphs (Richardson et al., 2017). Verma-constraints have been used for testing edges (Shpitser et al., 2009) and and for marginalization via variable elimination (Shpitser et al., 2011).

2.2 | Trapdoor variables

Here we give a broad definition that captures the notion that a set of variables $\mathbf{Z}$ may appear in an identifying functional of a causal effect, but the value of the causal effect is not dependent on the value of $\mathbf{Z}$.

Definition 1. (Functional independence) Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subset \mathbf{V}$ be disjoint sets and let $P(\mathbf{Y}|\ \mathrm{do}(\mathbf{X} = \mathbf{x}))$ be an identifiable causal effect with an identifying functional $g(\mathbf{v}) = f(P(\mathbf{v}))$. If the domain of g is $val(\mathbf{X}) \times val(\mathbf{Y}) \times val(\mathbf{Z})$, and $g(\mathbf{x}, \mathbf{y}, \mathbf{z}_1) = g(\mathbf{x}, \mathbf{y}, \mathbf{z}_2)$ for all $\mathbf{x} \in val(\mathbf{X})$, $\mathbf{y} \in val(\mathbf{Y})$ and $\mathbf{z}_1, \mathbf{z}_2 \in val(\mathbf{Z})$, then g is *functionally independent* from $\mathbf{Z}$.

The constraint defined above is specific to the given identifying functional. In some instances, we may be able to find identifying functionals that do not exhibit functional equality constraints for any subset $\mathbf{Z}$ of $\mathbf{V}$. Our interest lies in the opposite direction, where every identifying functional exhibits a

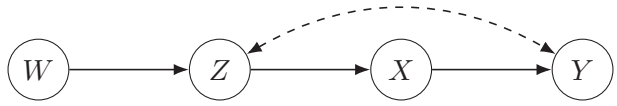

FIGURE 1 A causal graph where the identifying functional of $P(Y|\ \mathrm{do}(X = x))$ obtained by an application of the ID algorithm does not depend on the value of of W and there is an admissible set Z enabling back-door adjustment

specific type of functional independence. Before characterizing this property of interest, we must first define an operation known as the latent projection of a causal graph (Pearl & Verma, 1991).

Definition 2. (Latent projection) Let G be a causal graph over a set of vertices $\mathbf{V} \cup \mathbf{L}$. The *latent projection* $L(G, \mathbf{V})$ is a causal graph over $\mathbf{V}$ where for every pair of distinct vertices $Z, W \in \mathbf{V}$ it holds that

1. $L(G, \mathbf{V})$ contains an edge $Z \longrightarrow W$ if there exists a directed path $Z \longrightarrow \cdots \longrightarrow W$ in G on which every vertex except Z and W is in $\mathbf{L}$.
2. $L(G, \mathbf{V})$ contains an edge $Z \longleftrightarrow W$ if there exists a path from Z to W in G that does not contain the pattern $Z \longrightarrow M \& \longleftarrow W$ (a collider), on which every vertex except Z and W is in $\mathbf{L}$, the first edge of the path has an arrowhead into W and the last edge has an arrowhead into Z.

Latent projections can be used to derive identifying functionals for causal effects such that they do not contain a specific variable that is the variable is considered latent, and the causal effect of interest is identified in the corresponding latent projection (Tikka & Karvanen, 2018). For this reason, the presence of a functional constraint in some identifying functional does not rule out the possibility of obtaining another identifying functional that is not subject to the same constraint (Recall the example on the identifying functional of $P(Y|\, \mathrm{do}(X = x))$ in the graph of Figure 1).

We rule out this possibility of finding alternative identifying functionals by restricting our attention to settings where a *trapdoor variable* is present.

Definition 3. (Trapdoor variables) If $P(\mathbf{Y}|\, \mathrm{do}(\mathbf{X}{=}\mathbf{x}))$ is identifiable in G from $P(\mathbf{V})$, its identifying functional $f(P(\mathbf{V}))$ is functionally independent of $\mathbf{Z}$ where $\mathbf{Z} \subset \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$, and $P(\mathbf{Y}|\, \mathrm{do}(\mathbf{X}{=}\mathbf{x}))$ is not identifiable in $L(G, \mathbf{V} \setminus \mathbf{Z})$ from $P(\mathbf{V} \setminus \mathbf{Z})$, then $\mathbf{Z}$ is a set of *trapdoor variables* with respect to $f(P(\mathbf{V}))$ in G.

Finding trapdoor variables is straightforward as outlined by their definition. Given a causal effect of interest, we first determine its identifiability and whether the identifying functional is subject to functional equality constraints. If this is the case, we proceed to verify whether the causal effect might be identifiable when the set $\mathbf{Z}$ is considered latent using a suitable latent projection. The algorithm by Tian and Pearl (2002) can be used to systematically enumerate functional equality constraints implied by a causal model. Evans (2018) showed that this constraint-finding algorithm is complete for categorical variables, but it is not known whether the algorithm is complete in general, that is, whether all such constraints can be found by the algorithm.

If there exists a trapdoor variable for an identifying functional of a causal effect of interest, then the estimate of the causal effect may depend on the value of the trapdoor variable. This dependency can introduce bias in the estimate.

Definition 4. (Trapdoor bias) Let $\widehat{g}(\mathbf{v})$ be an estimator of an identifying functional $g(\mathbf{v})$ of a causal effect $P(\mathbf{Y}|\, do(\boldsymbol{X} = \boldsymbol{x}))$ and let $\mathbf{Z}$ be a set of trapdoor variables with respect to $g(\mathbf{v})$. Let $B(\widehat{g}(\mathbf{v}))$ denote the bias of this estimator. If there exists $\mathbf{z}_1, \mathbf{z}_2 \in val(\mathbf{Z})$ so that $B(\widehat{g}(\mathbf{x}, \mathbf{y}, \mathbf{z}_1)) \neq B(\widehat{g}(\mathbf{x}, \mathbf{y}, \mathbf{z}_2))$, then $\widehat{g}(\mathbf{v})$ exhibits *trapdoor bias* with respect to $\mathbf{Z}$.

For a consistent estimator, the effect of the trapdoor bias becomes negligible as the sample size grows, but for small samples the choice of how to take the trapdoor variables into account may be significant.

2.3 | Example on trapdoor variables

Consider the three causal graphs in Figure 2. We are interested in estimating the causal effect of X on Y. For example, in our application in Section 5, X will be the education level and Y is the yearly income. Furthermore, Z corresponds to the GPA from primary school, and W to the SES of the parents. In all three graphs, we have an arrow from X to Y, meaning that we assume that there is a direct causal effect of education on income. In addition, we assume that there may be some unobserved confounders between SES of the parents and income. In the graphs of Figure 2b and c, we also assume that there is confounding between SES of the parents and education level of the participant. In Figure 2c, we assume that the effect of SES on the education level is mediated by the GPA. We will further extended the third graph with additional variables, such as gender, in Section 5.

Now consider the estimation of the interventional distribution $P(Y|\,\mathrm{do}(X = x))$, that is the distribution of Y when we intervene on X by setting it to x, which differs from a simple conditional distribution of $P(Y|X = x)$ in these graphs. In order to estimate $P(Y|\,\mathrm{do}(X = x))$, we need to find a formula for it in terms of the observed variables only, that is Y, X, W and Z in our example. In the graph of Figure 2a, we obtain the so called back-door adjustment formula

$$P(Y|\ \mathrm{do}(X = x)) = \sum_w P(Y|\,x,\ w)P(w).$$

By conditioning on W, we block all back-door paths from X to Y that is those paths between X and Y which have arrows into X. Thus by estimating the parameters of the terms $P(Y|w,x)$ and $P(W)$, we can estimate the interventional distribution of interest, $P(Y|\,\mathrm{do}(X = x))$. For example, assuming that all variables are Gaussian and their relationships are linear, we have $Y(x) \sim N(a_y + b_{yw}a_w + b_{yx}x,\ s_y^2 + b_{yw}^2 s_w^2)$, with b_{ij} denoting the estimated regression coefficient of variable j on variable i, a_i denoting the intercept term and s_i^2 corresponding to the estimated residual variance.

Now consider the graph of Figure 2b. In this case, conditioning on W blocks the path $Y \longleftarrow W \longrightarrow X$ but opens the path $Y \longleftarrow W \longleftarrow X$ meaning that we cannot apply the back-door adjustment here. In fact, adding the unobserved confounder between W and X renders the causal effect nonidentifiable.

In the graph of Figure 2c, we assume that we have obtained data on variable Z which lies on the directed path from W to X. Given this additional information, the causal effect $P(Y|\,\mathrm{do}(X = x))$ is again identifiable, and we have

$$P(Y|\ \mathrm{do}(X = x)) = \frac{\sum_w P(Y|\,x,\ z,\ w)P(x|\,z,\ w)P(w)}{\sum_w P(x|\,z,\ w)P(w)}. \tag{1}$$

However, there is no term for the distribution of Z in Equation (1). Thus after estimating the distributions $P(Y|x, z, w)$, $P(X|z, w)$ and $P(W)$, Z is essentially reduced to a fixed but unknown parameter in the context

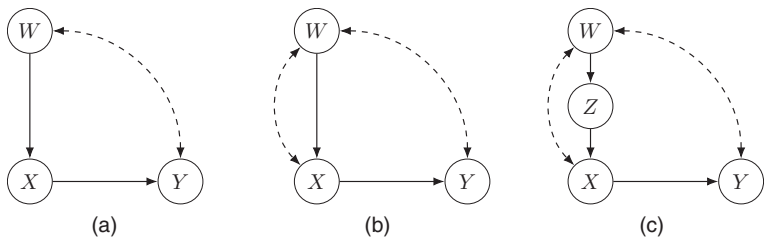

FIGURE 2 Three causal graphs of increasing complexity, where the interest is in the causal effect of X on Y

of $P(Y|\,\mathrm{do}(X = x))$. This graph is also considered by Tian and Pearl (2002), Pearl and Mackenzie (2018), and Jung et al. (2020). Tian and Pearl (2002) show that this graph contains a functional equality constraint of an interventional distribution which cannot be expressed as a conditional independence constraint using the observed variables. The constraint states that the formula for $P(Y|\mathrm{do}(X = x))$ given in Equation (1) is functionally independent of Z, meaning that its value is independent on the choice of z, just as we would intuitively expect. However, when estimating equation (1) from the data, we clearly must choose some value for Z, even though the constraint states that the actual value should not matter. In fact, Z is a trapdoor variable with respect to the identifying functional of Equation (1) in this graph. This follows from the functional equality constraint and the fact that the causal effect is not identifiable in the causal graph of Figure 2b which is the latent projection of Figure 2c when Z is considered latent.

Trapdoor variables are in no way limited to the scenario of Figure 2c. Figure 3a presents a generalization of Figure 2c with an additional set of variables $\mathbf{B} = \{B_1, \ldots, B_k\}$, that are possibly connected to each other with directed arrows or through unobserved confounders. In addition, any member of $\mathbf{B}$ is possibly connected to W via a bidirected edge or to Z, X or Y via a directed edge. Figure 4 depicts three additional examples of graphs where trapdoor variables are present. In the following sections we investigate the challenges that trapdoor variable s impose on the estimation of causal effects.

3 | ESTIMATION

3.1 | Plug-in estimation

A straightforward strategy for causal effect estimation given the assumed causal graph is to construct a parametric model for the conditional distributions that appear in the formula of an identifying functional of a causal effect. As an example, consider the graph of Figure 2c, and assume that we wish to estimate $E(Y|\,\mathrm{do}(X = x))$. We need to estimate the unknown model parameters θ corresponding to the causal effect of Equation (1) and we let $P_{\hat{\theta}}(\cdot)$ denote the density where the unknown parameters θ are assigned some fixed values such as their maximum likelihood (ML) estimates. The plug-in estimator for the expected value of the interventional distribution takes the following form

$$\widehat{E}_{\hat{\theta},z}(Y|\,\mathrm{do}(X = x)) = \sum_y y \frac{\sum_w P_{\hat{\theta}}(y|\,w,\,z,\,x) P_{\hat{\theta}}(x|\,w,\,z) P_{\hat{\theta}}(w)}{\sum_w P_{\hat{\theta}}(x|\,w, z) P_{\hat{\theta}}(w)}. \tag{2}$$

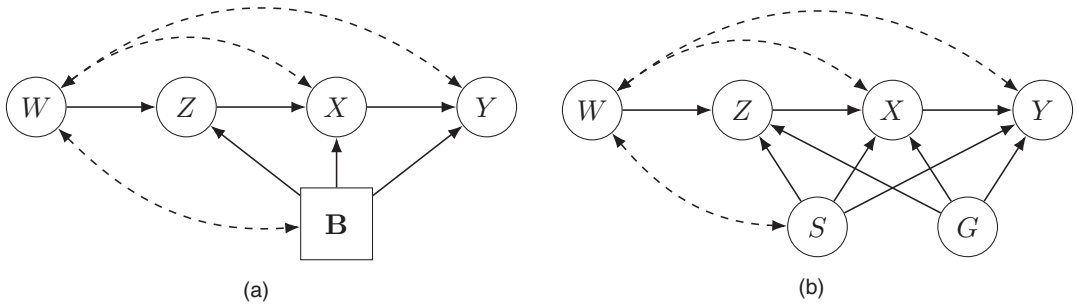

FIGURE 3 A generalization of the setting presented in Fig. 2c with an additional set of variables $\mathbf{B}$ is shown in (a). The square node for $\mathbf{B}$ denotes an arbitrary causal graph over $\mathbf{B}$. Edges to and from $\mathbf{B}$ mean that such an edge may exist between any member of $\mathbf{B}$ and the other endpoint of the edge. An instance of (a) is shown in (b) with $\mathbf{B} = \{S, G\}$

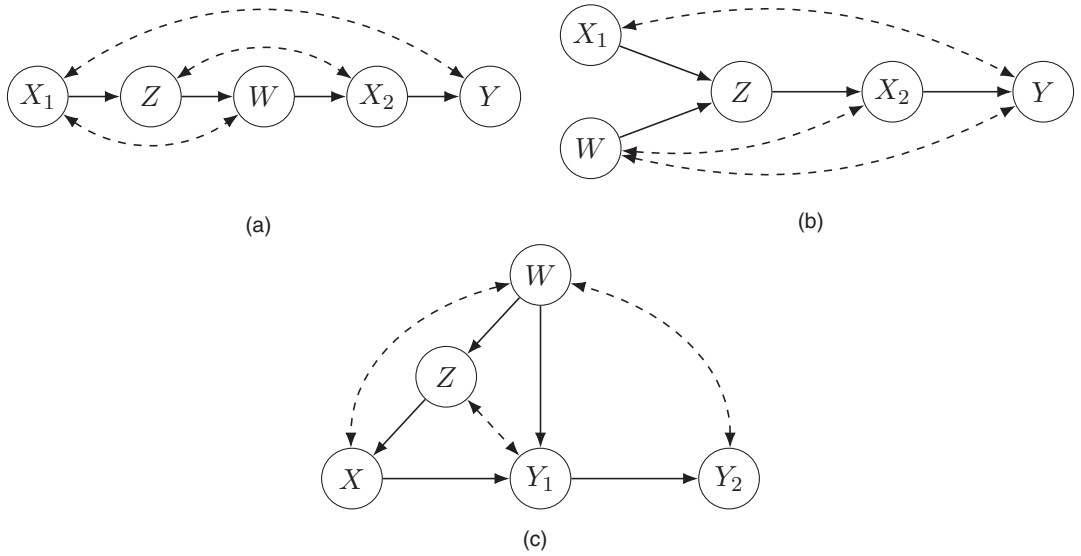

FIGURE 4 Examples on graphs where trapdoor variables are present. In (a) and (b), Z is a trapdoor variable for $P(Y|\ \mathrm{do}(X_1 = x_1, X_2 = x_2))$. In (c), Z is a trapdoor variable for $P(Y_1, Y_2 \,|\, \mathrm{do}(X = x))$

Note that this estimate may depend on the value of z when using a small sample estimate for θ. This dependency and the corresponding trapdoor bias vanish when the models are correctly specified and the true parameter values are used (by the definition of functional independence). However, even if we assume an unbiased estimator $\widehat{\theta}$, the estimator (2) can still exhibit bias since it is a nonlinear function of $\widehat{\theta}$ that depends on the choice of z. For large enough data sets, the choice of z in Equation (2) should not have a large impact, but we will show that for small samples, the value of the trapdoor variable z can have a crucial role in estimating the correct causal effect of X on Y.

An alternative strategy for defining the plug-in estimator is to eliminate the functional constraint by averaging the identifying functional over a (conditional) distribution of the trapdoor variable. If Z is a trapdoor variable and $\mathbf{T} \subseteq (\mathbf{V} \setminus Z)$, then

$$P(Y|\ \mathrm{do}(X = x)) = f(P(\mathbf{v})) \cdot 1 = f(P(\mathbf{v})) \sum_z P(z \,|\, \mathbf{t}) = \sum_z P(z \,|\, \mathbf{t}) f(P(\mathbf{v})),$$

since $f(P(\mathbf{v}))$ is functionally independent of Z. This leads to the following alternative estimator for the expected value of the interventional distribution in the causal graph of Figure 2c

$$\widehat{E}_{\widehat{\theta}}(Y|\ \mathrm{do}(X = x)) = \sum_y y \sum_z P_{\widehat{\theta}}(z \,|\, \mathbf{t}) \frac{\sum_w P_{\widehat{\theta}}(y \,|\, w,\, z,\, x) P_{\widehat{\theta}}(x \,|\, w,\, z) P_{\widehat{\theta}}(w)}{\sum_w P_{\widehat{\theta}}(x \,|\, w, z) P_{\widehat{\theta}}(w)}. \tag{3}$$

Note that while this strategy eliminates the problem of having to choose a specific value for z, it also introduces a new problem of selecting the set $\mathbf{T}$ and estimating the model parameters for the distribution $P_\theta(Z|\mathbf{t})$. As before, this estimator is a nonlinear function of the parameters that can result in bias. Note that in a frequentist framework, there is no way to divide the total bias of the final estimate into trapdoor bias, plug-in bias, or other possible sources of bias.

3.2 | Bayesian estimation of causal effects

We advocate the use of Bayesian methods for jointly estimating the parameters of the distributions of identifying functionals of interest. By drawing samples from the joint posterior of the model parameters (using any generic Bayesian inference engine, typically some type of Markov chain Monte Carlo (MCMC) algorithm), we propagate the parameter estimation uncertainty to the final causal effect estimates and avoid the plug-in bias due to the nonlinear formula of the identifying functional with respect to model parameters. This allows us to focus on the effects of the trapdoor variable and the trapdoor bias of the estimators. We can also obtain samples from the full posterior of $P(Y|\ \mathrm{do}(X = x))$ which can be used for straightforward evaluation of any properties of interest (such as mean and variance) of this posterior.

In parametric estimation of causal effects, we typically do not have an analytical formula for the types of plug-in estimators that are shown in Equations (2) and (3). In these cases, we can use a Monte Carlo approach to draw samples from $P_{\hat{\theta}}(Y|\ \mathrm{do}(X = x))$, and estimate the desired quantity using these samples. The specific Monte Carlo algorithm depends on the causal graph, the identifying functional and the corresponding conditional distributions. In simple cases, we simulate variables from their (conditional) distributions with fixed x, whereas, for example in the case of Equation (1) where $P(X|\cdot)$ is present, we need additional weighting of the samples. Consider our extended example graph in Figure 3a, which leads to the following formula for $P(Y|\ \mathrm{do}(X = x))$:

$$\sum_{\mathbf{b}} P(\mathbf{b}) \frac{\sum_w P(Y \mid x, z, w, \mathbf{b}) P(x \mid z, w, \mathbf{b}) P(\mathbf{b} \mid w) P(w)}{\sum_w P(x \mid z, w, \mathbf{b}) P(\mathbf{b} \mid w) P(w)}. \tag{4}$$

Algorithm 1 describes the Monte Carlo algorithm for Equation (4) based on the second approach discussed in Section 3.1, given the parameter vector $\widehat{\theta}$ (the same algorithm is also suitable for (1) by omission of the references to the variables in $\mathbf{B}$). Algorithm 1 draws samples from the marginal and conditional distributions defined in (4) and gives us $N \times M$ weighted replications from $P_{\hat{\theta}}(Y|\ \mathrm{do}(X = x))$ which allows us to compute, for example

$$\widehat{E}_{\hat{\theta}}(Y|\,\mathrm{do}(X = x)) = \sum_{i=1}^{N} \sum_{j=1}^{M} \overline{\gamma}^{ij} y^{ij}, \tag{5}$$

where the weights $\overline{\gamma}^{ij}$ are defined in the last step of Algorithm 1.

Algorithm 1 Monte Carlo algorithm for sampling from $P_{\hat{\theta}}(Y \mid \mathrm{do}(X = x))$ defined by equation (4) with $N \times M$ Monte Carlo samples.

1: For $i = 1, \ldots, N$:
2: Sample $\mathbf{b}^i \sim P_{\hat{\theta}}(\mathbf{B})$
3: Set or sample the value of the trapdoor variable, for example as $z^i = E_{\hat{\theta}}(Z)$ or $z^i \sim P_{\hat{\theta}}(Z \mid x, \mathbf{b}^i)$
4: For $j = 1, \ldots, M$:
5: Sample $w^{ij} \sim P_{\hat{\theta}}(W)$
6: Sample $y^{ij} \sim P_{\hat{\theta}}(Y \mid x, z^i, w^{ij}, \mathbf{b}^i)$
7: Compute $\gamma^{ij} = P_{\hat{\theta}}(x \mid z^i, w^{ij}, \mathbf{b}^i) P_{\hat{\theta}}(\mathbf{b}^i \mid w^{ij})$
8: Compute the normalized weights:

$$\bar{\gamma}^{ij} = \frac{\gamma^{ij}}{\frac{1}{M} \sum_{i=1}^{N} \gamma^{ij}}, \quad i = 1, \ldots, N,\ j = 1, \ldots, M$$

Here the trapdoor variable has a more crucial role than in the case of analytical formulas. With poor choices of z, our weights $\overline{\gamma}^{ij}$ can become degenerate, that is, most of the weights are near zero. In order to avoid this, it is natural to condition the trapdoor variable on x and perhaps other variables such as members of $\mathbf{b}^i$ in Algorithm 1 which should make the weights well behaved.

The suitable number of Monte Carlo samples $N \times M$ can determined by computing the Monte Carlo standard error (MCSE) of our causal effect estimate. The MCSE measures the additional uncertainty in our estimate due to the finite Monte Carlo sample size. For example, the MCSE for estimator (4) can be computed as

$$MCSE\left(\widehat{E}_{\widehat{\theta}}(Y|\ \mathrm{do}(X=x))\right) = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M}\left[\overline{\gamma}^{ij}\left\{y^{ij} - \widehat{E}_{\widehat{\theta}}(Y|\ \mathrm{do}(X=x))\right\}\right]^2}.$$

When combining Algorithm 1 with Bayesian parameter estimation, the algorithm is used at each MCMC iteration given the current values of the model parameters θ.

With the MCMC approach, we can compute functions of $Y(x)$ at each iteration, giving us samples from its posterior distribution, or we can store all $N \times M$ weighted Monte Carlo samples leading to posterior distribution of $Y(x)$, that is, the variable Y in the post-interventional distribution, which can be further used to evaluate functions of interest. In the latter case, it can be practical to resample (using the corresponding weights) and store only one replication y^{ij} at each iteration if memory constraints limit the storing of all samples.

3.3 | Analytical causal effect for a linear-Gaussian model

We will now consider the estimation of causal effects in a common linear-Gaussian case for the causal graph of Figure 2c. We assume that the underlying model is defined as

$$\begin{aligned}
U &\sim N(\mu_u, \sigma_u^2),\\
V &\sim N(\mu_v, \sigma_v^2),\\
(W|U=u, V=v) &\sim N(\alpha_w + \beta_{wu}u + \beta_{wv}v, \sigma_w^2),\\
(Z|W=w) &\sim N(\alpha_z + \beta_{zw}w, \sigma_z^2),\\
(X|Z=z, V=v) &\sim N(\alpha_x + \beta_{xz}z + \beta_{xv}v, \sigma_x^2),\\
(Y|X=x, U=u) &\sim N(\alpha_y + \beta_{yx}x + \beta_{yu}u, \sigma_y^2).
\end{aligned} \tag{6}$$

Here all the parameters $\mu_\cdot$, $\alpha_\cdot$, $\beta_{\cdot\cdot}$ and $\sigma_\cdot$ are unknown, and U and V are unobserved confounders. Our observational model needed for estimating $P(Y|\ \mathrm{do}(X=x))$ is

$$\begin{aligned}
W &\sim N(a_w, s_w^2),\\
(X|Z=z, W=w) &\sim N(a_x + b_{xz}z + b_{xw}w, s_x^2),\\
(Y|X=x, Z=z, W=w) &\sim N(a_y + b_{yx}x + b_{yz}z + b_{yw}w, s_y^2),
\end{aligned} \tag{7}$$

where parameters $a_\cdot$, $b_{\cdot\cdot}$ and $s_\cdot$ are unknown and have to be estimated from the data.

Now using equations of model (7) to Equation (1) yields

$$E(Y|\ \mathrm{do}(X=x)) = a_y + \frac{b_{yw}s_x^2}{b_{xw}^2 s_w^2 + s_x^2}a_w - \frac{b_{yw}b_{xw}s_w^2}{b_{xw}^2 s_w^2 + s_x^2}a_x + \left(b_{yx} + \frac{b_{yw}b_{xw}s_w^2}{b_{xw}^2 s_w^2 + s_x^2}\right)x + \left(b_{yz} - \frac{b_{yw}b_{xw}s_w^2}{b_{xw}^2 s_w^2 + s_x^2}b_{xz}\right)z \quad (8)$$

and

$$Var(Y|\ \mathrm{do}(X=x)) = \frac{b_{xw}^2 s_y^2 s_w^2 + s_x^2\left(b_{yw}^2 s_w^2 + s_y^2\right)}{b_{xw}^2 s_w^2 + s_x^2}.$$

While the variance $Var(Y|\ \mathrm{do}(X=x))$ does not contain the trapdoor variable Z, the expected value $E(Y|\ \mathrm{do}(X=x))$ does, which is not surprising since $P(Y|\ \mathrm{do}(X=x)) = P(Y|\ \mathrm{do}(X=x,Z=z))$ in the graph of Figure 2c. Also, if the interest is only in the difference $E(Y|\ \mathrm{do}(X=x+1)) - E(Y|\ \mathrm{do}(X=x))$ then the effect of trapdoor variable cancels out in this linear case.

As our model is linear-Gaussian, we can apply path analysis (Wright, 1934) to our causal graph (see, e.g. Pearl (2013)) to find the marginal covariance matrix Σ for (Y, X, Z, W) (shown in Appendix). From Σ, using the properties of multivariate normal distribution, we can obtain the theoretical formulas for the parameters $(a_\cdot, b_{\cdot\cdot}, s_\cdot)$ of model (7) in terms of the true parameters $(\mu_\cdot, \alpha_\cdot, \beta_{\cdot\cdot}, \sigma_\cdot)$ of the causal graph (see Appendix for details). Plugging these into Equation (8), we obtain

$$E(Y|\ \mathrm{do}(X=x)) = \alpha_y + \beta_{yu}\mu_u + \beta_{yx}x,$$

and

$$Var(Y|\ \mathrm{do}(X=x)) = \beta_{yu}^2\sigma_u^2 - 2\beta_{wu}\beta_{zw}\beta_{xz}\beta_{yx}\beta_{yu}\sigma_u^2 + \sigma_y^2.$$

As expected, given the true causal model and its known parameters, the effect of the trapdoor variable cancels out. Nevertheless, with a finite data set, the estimate of expected value (8) depends on z, unless $\left(b_{yz} - \frac{b_{yw}b_{xw}s_w^2}{b_{xw}^2 s_w^2 + s_x^2}b_{xz}\right)$ happens to estimate to zero.

4 | SIMULATION EXPERIMENTS

4.1 | Binary model

We will now illustrate different choices for the trapdoor variable with nonparametric estimation of $P(Y|\ \mathrm{do}(X=x))$ in a case where all variables are binary. Consider Bernoulli variables defined in accordance with Figure 2c as

$$\begin{aligned} V &\sim \mathrm{Bernoulli}(0.5),\\ U &\sim \mathrm{Bernoulli}(0.5),\\ (W|V=v, U=u) &\sim \mathrm{Bernoulli}(0.4u+0.4v),\\ (Z|W=w) &\sim \mathrm{Bernoulli}(0.4+0.4w),\\ (X|Z=z, V=v) &\sim \mathrm{Bernoulli}(0.4z+0.4v),\\ (Y|X=x, U=u) &\sim \mathrm{Bernoulli}(0.4x+0.4u), \end{aligned} \quad (9)$$

where V and U correspond to the bidirected edges between X and W and between W and Y, respectively. By solving Equation (1) analytically, we obtain

$$P(Y=y|\,\mathrm{do}(X=x)) = (0.2+0.4x)^y(0.8-0.4x)^{1-y}, \quad y\in\{0,1\},$$

which does not depend on z. However, in practice, when U and V are unobserved and we need to estimate distributions $P(W)$, $P(X|z, w)$, and $P(Y|x, z, w)$ from finite data, we will show how the estimate of $P(Y|\text{do}(X = x))$ can depend on the chosen value z.

As an example, we simulated 100,000 data sets according to model (9) with sample sizes 100, 300 and 500, and estimated $P(Y|\text{do}(X = x))$ nonparametrically with various methods for dealing with the trapdoor variable. Besides fixing z to zero or one and using the estimator (2), we also computed a weighted average of these estimates, where the weights were based on the estimated marginal distribution $P(Z)$, or the conditional distribution $P(Z|x)$, corresponding to the estimator (3). With sample sizes 100 and 300, there were some cases where fixing z to zero or one lead to undefined probabilities (for example, there were no observations for which $x = 0$, $z = 0$ and $w = 0$). In these cases (about 10% of cases with sample size 100, and less than 0.1% with sample size 300), the entire replication was discarded. Figure 5 shows the results of the simulation. We see that using the fixed values with $z = 0$ or $z = 1$ leads to some bias. The weighted average estimators perform better, and the one based on $P(Z|x)$ outperforms the estimator based on $P(Z)$. Overall, we are not far off from the ground truth in this simple setting.

4.2 | Linear-Gaussian model

Now consider a model based on Equation (6) with

$$\begin{aligned} U &\sim N(1,\ 1),\\ V &\sim N(1,\ 1),\\ (W|U=u,\ V=v) &\sim N(1+u+v,\ 1),\\ (Z|W=w) &\sim N(1+w,\ 1),\\ (X|Z=z,\ V=v) &\sim N(1+z+v,\ 1),\\ (Y|X=x,\ U=u) &\sim N(1+x+u,\ 0.01). \end{aligned} \tag{10}$$

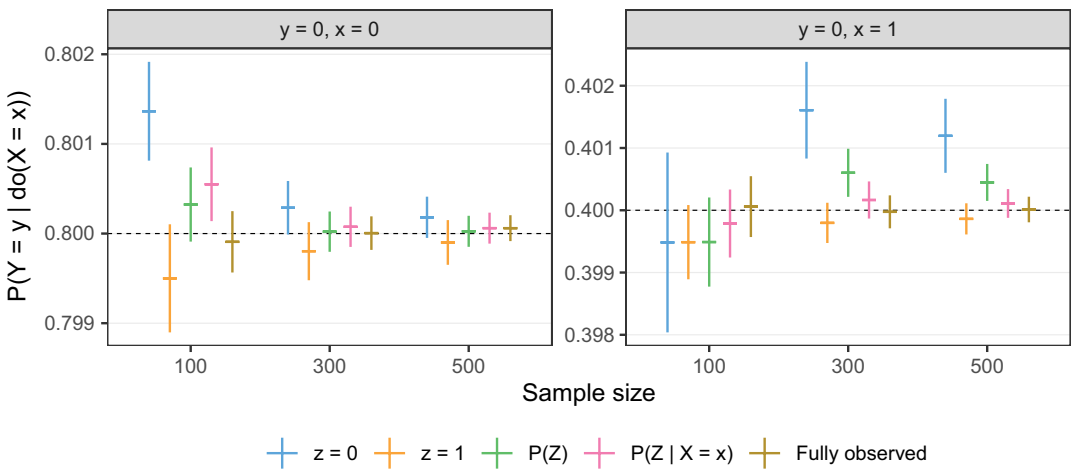

FIGURE 5 Average estimates of $P(Y = 0|\text{do}(X = x))$ and $\pm 2SE$ over 100,000 replications for the Binary model with varying sample sizes and different strategies to account for the trapdoor variable Z. The dashed lines show the true causal effects

We will compare various approaches to account for the trapdoor variable Z. Instead of the nonparametric approach used in Section 4.1, we now switch to parametric Bayesian modelling. For comparative purposes, we use uniform priors for all unknown model parameters ($a_\cdot, b_{\cdot,\cdot}$ and $s_\cdot$ in Equation (7)). As stated in Section 3, the Bayesian approach takes into account the uncertainty in $P(Y|\,\mathrm{do}(X = x))$ due to parameter estimation by integrating over the posterior distribution of the parameters and avoids the plug-in bias due to the nonlinearity of Equation (8).

For model estimation, we wrote a Stan model (Carpenter et al., 2017; Stan Development Team, 2019) which simultaneously estimates all unknown model parameters and $E(Y|\mathrm{do}(X = x))$ using MCMC. We estimate the true expected causal effect $E(Y|\,\mathrm{do}(X = x))$ using the causal graph with observed confounders, with $x \in \{0, 3, 6, 9\}$, and compare it to the estimates obtained by six different methods:

1. Fix the trapdoor variable Z to the constant $z = 0$ in Equation (8).
2. Marginalize over $P(Z)$ that is in addition of estimating model (7) also estimate $P(Z)$, so that the estimator (8) uses $z = E(Z)$ at each MCMC iteration.
3. As above, but estimate $P(Z|x)$ and use $z = E(Z|x)$.
4. Use a constraint $b_{yz} = (b_{xz}b_{yw}b_{xw}s_w^2)/(b_{xw}^2 s_w^2 + s_x^2)$ so that the contribution of z is fixed to zero.
5. Fit a structural equation model (SEM), a common approach for linear-Gaussian causal modelling (Kline, 2011). We used the `R` package `lavaan` (Rosseel, 2012) for this purpose.
6. Use the composition of weighting operators (CWO) (Jung et al., 2020), which applies weighted regression to estimate functions of interventional distributions for arbitrary identifying functionals.

Based on model 1000 we have $E(X) = 6$, $E(Z) = 4$, and $E(Z|X = 0) = 0.25$, $E(Z|X = 3) = 2.1255$, $E(Z|X = 6) = 4$ and $E(Z|X = 9) = 5.875$.

We sampled 1000 replications of varying sample size from model (10) and for each replication estimated $E(Y|\,\mathrm{do}(X = x))$ using a single MCMC chain with 10,000 post-warmup iterations and averaging over the iterations (thus taking account the uncertainty from parameter estimation), except for the SEM and CWO approaches which are based on the ML estimates. Figure 6 shows how, despite theoretical equivalence, results depend heavily on the chosen strategy. Perhaps the most natural choices of $z = 0$ and $z = E(Z)$ are prone to bias which depends on the value x, but the case where z is adapted based on x performs well. The SEM approach, which explicitly models the covariance structures between variables, also performs well due to the structure of the graph (see Shpitser et al. (2018) for more details), but is not applicable in more general graphs with non-Gaussian or nonlinear equations. The CWO method shows considerable bias when $x \neq E(X)$.

4.3 | Analytical versus Monte Carlo approach for the linear-Gaussian model

Consider again model (10) but now with

$$(X \,|\, Z = z, V = v) \sim N(1 + z + v,\, 0.01),$$

that is smaller measurement error in X, leading also to a narrower density of $P(X|z, w)$. Because of this, a poor choice for the value of the trapdoor variable can cause most of the normalized weights $\overline{\gamma}$ in Algorithm 1 to be close to zero, leading to inefficient Monte Carlo sampling. As an example, we simulated one replication of size $n = 100$ from this model, and estimated the posterior distribution of $E(Y|\,\mathrm{do}(X = 0))$, both using the analytical formula and Monte Carlo with $N = 500$. For MCMC, we

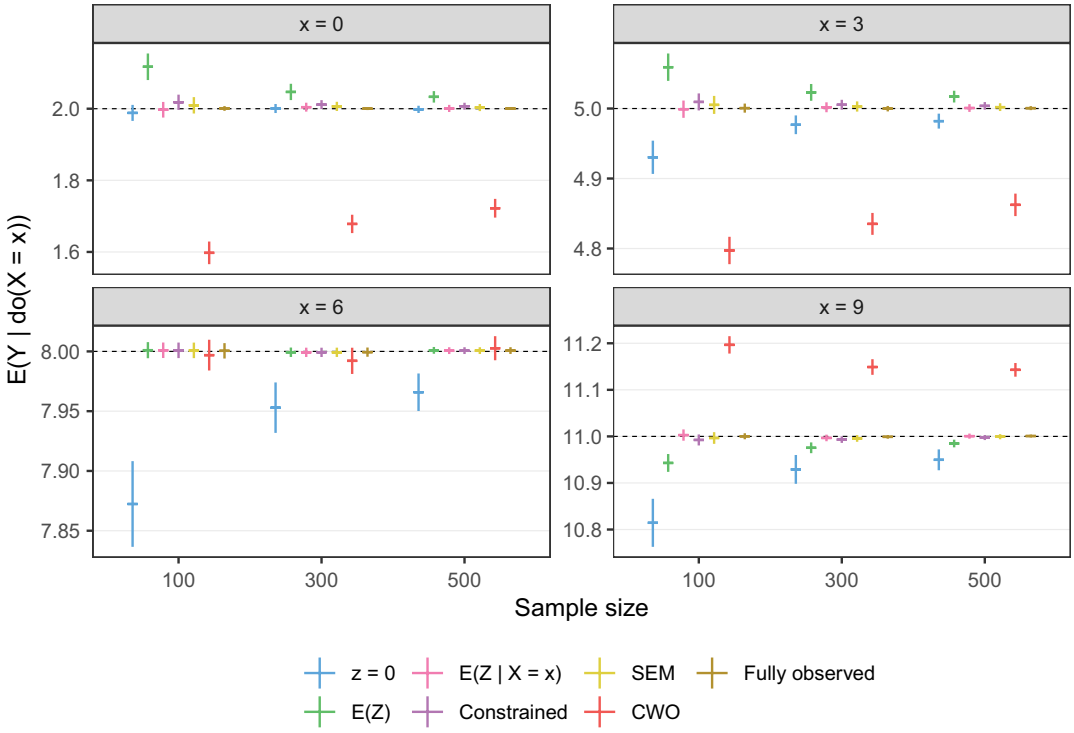

FIGURE 6 Average estimates of $E(Y|\,\mathrm{do}(X = x))$ and ± 2 SE over 1000 replications with sample sizes 100, 300 and 500 and different choices for taking the trapdoor variable into account in the simulation study of Section 4.2. The dashed lines show the true expected values

ran 4 chains with a total of 100,000 post-warmup iterations. In addition, at each MCMC iteration, we sampled one replication from the posterior predictive density $P(Y|\,\mathrm{do}(X = 0))$ from the weighted Monte Carlo sample. Figure 7 shows the posterior distributions. We can see that with suitable $z = E(Z|X = 0)$ both analytical and Monte Carlo methods give approximately equal results. A poor choice of $z = E(Z)$ biases results compared to the analytical solution (which is also biased) as can be seen from the discrepancy between the posterior distributions of $E(Y|\,\mathrm{do}(X = 0))$ based on Monte Carlo and analytical approaches. With $N = 500$, the average MCSE (over posterior samples) was 0.08 for $z = E(Z|X = 0)$ and 0.22 for $z = E(Z)$. The additional variation due to the Monte Carlo sampling can inflate the posterior distribution of $E(Z|X = 0)$, but here with $N = 500$ the proportion of MCSE of the total posterior uncertainty is negligible: when using $z = E(Z|X = 0)$, the posterior standard deviations of analytical and Monte Carlo methods were 0.40 and 0.41, respectively.

4.4 | Non-Gaussian model

In Section 5, we study the causal effect of education level on income using real data. Here we perform a simulation experiment where we emulate the real-data case using the assumed graph of the real-data case and define the data generating process so that it reflects the nature of the variables in Section 5. We use graph of Figure 3b where we obtain the following formula for $P(Y|\,\mathrm{do}(X = x))$:

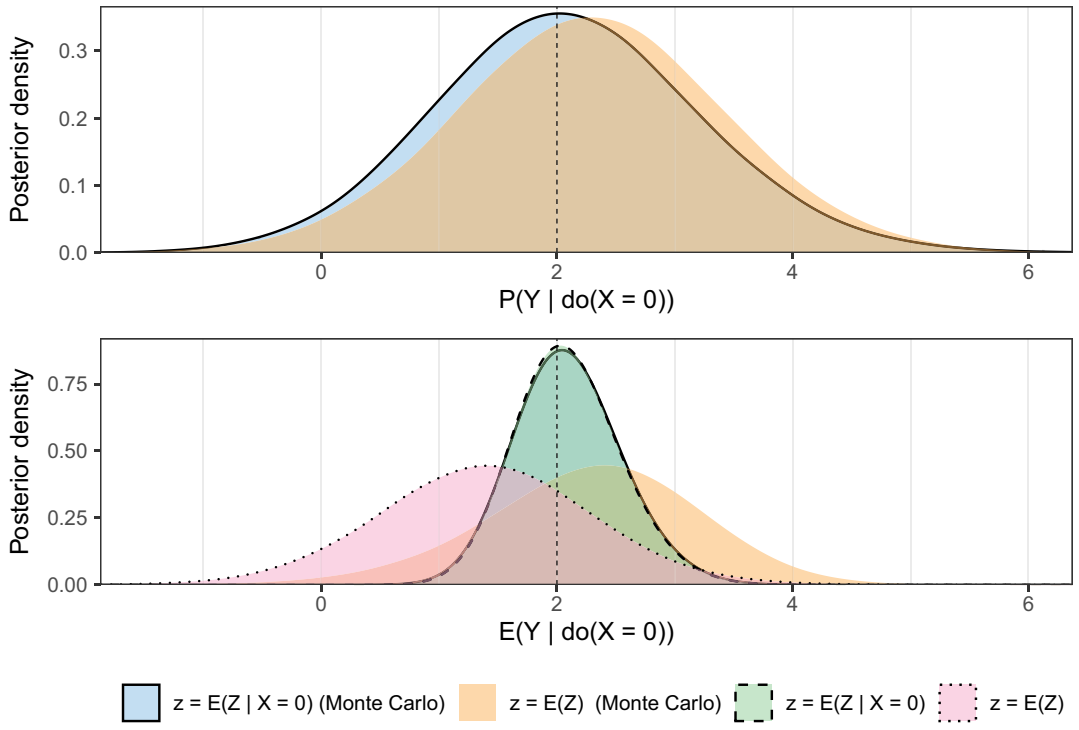

FIGURE 7 Posterior predictive distribution $P(Y|\,\mathrm{do}(X=0))$ and posterior distribution of $E(Y|\,\mathrm{do}(X=0))$ of the linear-Gaussian model for different estimation methods in the Monte Carlo simulation of Section 4.3

$$\sum_{s,g} P(g)P(s)\frac{\sum_w P(Y\,|\,x,z,w,s,g)P(x\,|\,z,w,s,g)P(s\,|\,w)P(w)}{\sum_w P(x\,|\,z,\ w,\ s,\ g)P(s\,|\,w)P(w)}. \tag{11}$$

Let variables U_1, U_2 and U_3 correspond to the confounders between W and Y, W and X, and W and S, respectively. For our simulation purposes, we define a following data generating process:

$$\begin{aligned}
U_1 &\sim N(0,1),\\
U_2 &\sim N(0,1),\\
U_3 &\sim N(0,1),\\
G &\sim \text{Bernoulli}(0.5)\\
(S|U_3=u_3) &\sim N(36+3u_3\ ,\ 25)\\
(W|U_1=u_1,\ U_2=u_2, U_3=u_3) &= \begin{cases} 1 & \text{if } u_1+u_2+u_3 \le -1.1\\ 2 & \text{if } -1.1 < u_1+u_2+u_3 \le 1.9\\ 3 & \text{otherwise}\end{cases}\\
(Z|W=w,\ S=s,\ G=g) &\sim \text{Beta}(\mu_z\phi_z,\ (1-\mu_z)\phi_z),\\
(X|Z=z,\ U_2=u_2,\ S=s,\ G=g) &\sim \text{SM}(-0.5g+0.04s+13.5z+2u_2, (12.5,\ 14)),\\
(Y|X=x, U_1=u_1,\ S=s,\ G=g) &\sim \text{Gamma}(10000,\ 10000/\mu_y),
\end{aligned}$$

where

$$\mu_z = \exp(-1.2 + 0.4g + 0.05s + 0.1\mathrm{I}(w=2) + 0.3\mathrm{I}(w=3)),$$
$$\phi_z = \exp(2.2 + 0.2g),$$
$$\mu_y = \exp(9.3 + 0.02s - 0.5g + 0.2\mathrm{I}(x=1) + 0.5\mathrm{I}(x=2) + 0.4u_1),$$

and SM(η,τ) is a sequential model (Bürkner & Vuorre, 2019; Tutz, 1990) with a linear predictor η and a threshold vector τ.

Compared to the linear-Gaussian experiment of Section 4.2, here additional difficulties arise due to the fact that in addition to the unknown parameters θ, our distributional assumptions for the terms in Equation (11) are not correct. For example, while $(Y|X = x,\ U_1 = u_1,\ S = s,\ G = g)$ is Gamma distributed by definition, $(Y|X = x,\ S = s,\ G = g,\ W = w,\ Z = z)$ might not be. This can naturally bias our causal estimates, but the question remains whether different choices for the trapdoor variable affect the bias or precision of the causal estimates.

For the terms in (11), we assume a Gamma distribution for Y, and model its expected value given other variables via log-link, assuming a monotonic effect (Bürkner & Charpentier, 2020) of X and W. We treat X and W as ordinal variables and model them with ordered logistic regression and a sequential model with a logit-link (Bürkner & Vuorre, 2019; Tutz, 1990), respectively. We use a normal distribution for S, and a Bernoulli distribution for G.

We simulated 1000 replications of sample size 500 from this model, and estimated the posterior mean of $P(Y|\ \mathrm{do}(X = x))$ using the four different strategies for the trapdoor variable: $Z \sim P(Z)$, $Z \sim P(Z|\,S = s^i,\ G = g^i)$, $Z \sim P(Z|\,S = s^i,\ G = g^i,\ X = x)$, and $Z \sim P(Z|X = x)$. We use a beta distribution for Z in all cases by modelling the expected value via a logit-link and precision via a log-link. Based on the data generating process, the true causal effects are 19,690; 24,049 and 32,463 for $x = 0$, 1, 2, respectively. The root mean square error (RMSE) and bias of our causal estimates compared to ground truth are shown in Table 1. We see that RMSE increases with respect to the intervention variable X and the differences between trapdoor strategies are relatively small except when $x = 2$, where conditioning on the intervention variable results in approximately 20% smaller RMSE than when using the marginal distribution of the trapdoor variable or when conditioning only with covariates S and G. The differences in bias estimates are within the Monte Carlo error in the case of $X = 0$. With $X = 1$, the trapdoor strategies without conditioning on X results in positive bias whereas conditioning on X leads to a bias of similar magnitude but in a different direction. For $X = 2$, the bias

TABLE 1 RMSEs and biases of the estimates for the causal effect of X on Y in the non-Gaussian simulation experiment using four different strategies for sampling the trapdoor variable Z, and the true data generating process with observed confounders. MCSEs (computed with `simhelpers` R package (Joshi & Pustejovsky, 2020)) were between 0.4 and 2 for the RMSE estimates, and between 14 and 62 for the bias estimates, increasing with respect to X in both cases

	RMSE			Bias		
Trapdoor strategy	**X = 0**	**X = 1**	**X = 2**	**X = 0**	**X = 1**	**X = 2**
$P(Z\|x, s, g)$	528	875	1770	121	−149	209
$P(Z\|x)$	528	890	1729	123	−233	4
$P(Z\|s,g)$	540	914	2238	106	110	1028
$P(Z)$	544	912	2173	114	101	942
Data generating process	443	541	734	−3	−3	−3

is substantially larger with trapdoor strategies not involving X. In this simulation experiment, conditioning on the covariates $\mathbf{B}$ does not improve the estimates.

5 | CAUSAL EFFECT OF EDUCATION ON INCOME

As an example with real data, we analyse Life Course 1971–2002 data set from Finnish Social Science Data Archive (Kuusinen, 2018). This longitudinal study consists of life courses of 634 Finnish children born in 1964–1968 in Jyväskylä, Finland. In the early 1970s, when the children were aged between 3 and 9, the Illinois Test of Psycholinguistic Abilities (ITPA) was used to test their verbal intelligence. Participants were then followed up, with further information on their life events gathered in 1984, 1991 and 2002. While the data set has been used in various studies regarding the intelligence and school achievement (see, e.g. Kuusinen & Leskinen, 1988), we use this data to study the causal effect of education level (secondary or less, lowest/lower tertiary, or higher tertiary) on yearly income (euros, in 2000). Note that due to the limited geographical scope of the data, the participants do not necessarily form a fully representative sample of the corresponding birth cohorts in Finland.

In addition to earned income Y and education level X, we have information on the GPA from primary school Z, SES of the parents W (low, middle and high), gender G and the ITPA score S (the General Language composite variable, a combination of the subtests). Out of the 634 participants in the full data set, we use 509 participants with fully observed aforementioned variables and avoid considerations related to missing data. The assumed causal graph is shown in Figure 3b, and as before, all analyses are based on the assumption that this graph is correct. We use same distributional assumptions as in our simulation experiment in Section 4.4, after transforming the original scale of GPA from 4–10 to 0–1.

It could be argued that there should be direct arrows from parental SES to child's education level and income. However, in Finland, it is likely that the effect of SES to education is strongly mediated by the child's school achievement (variable Z in our graph) (Acacio-Claro et al., 2018). Also, the SES in our data was coded based on the occupations of the parents, which in turn depends on their education. We thus assume that the unobserved confounder between W and X contains, among other things, the education levels of the parents. Similarly, there are studies suggesting that in Finland, after accounting for a child's education, the effect of family income on children's income is low (Österbacka, 2001). Finally, the intergenerational mobility of education and income are among the highest in Finland (and Nordic countries in general) compared to other countries (Björklund et al., 2002; Pfeffer, 2008). This suggests that even if there is a direct arrow from W to X or Y, these direct effects are likely negligible. Nevertheless, it is, of course, possible (and perhaps likely) that our causal graph is an oversimplification of the complex causal mechanisms related to education and income.

We used four chains with total of 100,000 post-warmup iterations and $N = M = 250$, leading to MCSE around 50–80, depending on x with both trapdoor strategies.

Table 2 shows the posterior summary statistics for $median(Y|\,\mathrm{do}(X = x))$ and $E(Y|\,\mathrm{do}(X = x))$ for all education groups. We see a clear discrepancy between the estimates based on the conditional trapdoor variable $Z \sim P(Z|x, s, g)$ and the marginal trapdoor variable $Z \sim P(Z)$. Finally, Figure 8 shows the full posterior distributions $P(Y|\,\mathrm{do}(X = x))$. The two strategies for the trapdoor variable give somewhat different results: When conditioning the trapdoor variable grade on the education level, the interventional distributions between educational levels differ more than with sampling the trapdoor variable

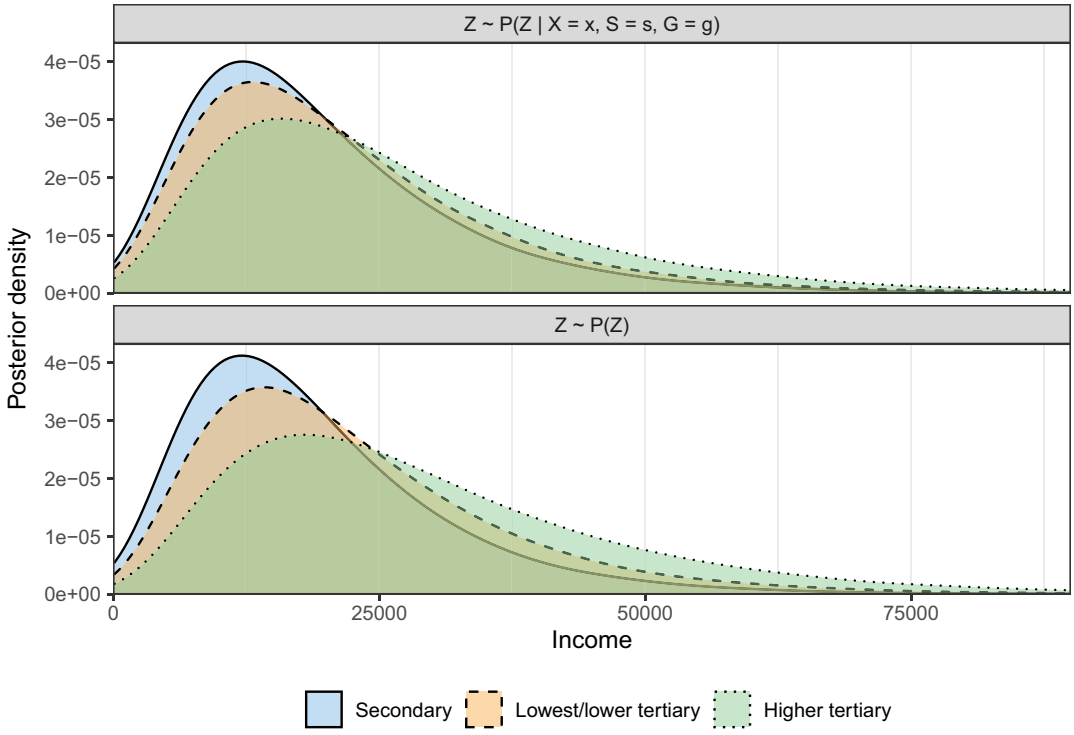

FIGURE 8 Posterior predictive distribution $P(Y|\ \mathrm{do}(X = x))$ for the income model using the estimators with the conditional trapdoor variable (upper panel) and the marginal trapdoor variable (lower panel)

TABLE 2 Posterior mean, median, and their posterior standard deviations for the causal effect of education on income, in hundred euros

Education level	Trapdoor strategy	Mean	Median	SD (Mean)	SD (Median)
Secondary or less	$P(Z)$	195	166	8	6
Secondary or less	$P(Z\|x,s,g)$	187	161	7	6
Lowest/lower tertiary	$P(Z)$	216	183	12	11
Lowest/lower tertiary	$P(Z\|x,s,g)$	223	192	13	11
Higher tertiary	$P(Z)$	266	226	21	18
Higher tertiary	$P(Z\|x, s, g)$	295	253	22	19

from $P(Z)$. For example, we estimate the mean annual income for the highest education level as 29,500 euros using the conditional trapdoor variable compared to 26,600 euros when using the marginal trapdoor variable. On the basis of the simulation results of Section 4.4, we prefer the estimates with $Z \sim P(Z|x, s, g)$. Unsurprisingly, obtaining a higher education level has a positive causal effect on income. Our causal estimates also differ somewhat from the observed incomes in different education

groups, with the observed median incomes being (from lowest to highest) 17,700, 19,300 and 26,800 euros, and the mean incomes 18,700, 21,800 and 28,700 euros.

6 | CONCLUSION

We have shown how it is possible to estimate causal effects when the back-door and front-door adjustments are not applicable and highlighted the potential issues related to the application of theoretical identifying formulas to finite data sets. Our real-data example on the effect of education on income illustrates how Bayesian causal inference can be applied to complex causal graphs with multiple types of variables, and how trapdoor variables can have a substantial effect on the estimated causal effects.

The basic structure that leads to implicit functional constraints and trapdoor variables was presented in the graph of Figure 2c. This graph has been considered in the literature earlier but according to our knowledge, it has not been fully analysed from the viewpoint of estimation. This basic structure can be extended in several directions so that the causal effect is still identifiable while retaining the implicit functional constraint. The graph for the Life Course 1971–2002 study (Figure 3b) is an example of such an extension. There are other interesting graphs such as those shown in Figure 4 with similar or even more complex identifying functionals which exhibit the problems related to trapdoor variables.

Determining the optimal method to account for the presence of a trapdoor variable remains a challenging problem. The trapdoor bias exhibited for small sample sizes depends on the (possibly parametric) assumptions about the causal model as well as the properties of the estimators used for the conditional distributions appearing in the identifying functional. Even in the simple case of the linear model with a single trapdoor variable, the optimal method is not apparent. It may also be the case that minimizing the trapdoor bias might result in a large variance of the causal effect estimator necessitating further considerations about the estimation problem at hand.

Our simulation experiments in Bernoulli and linear-Gaussian cases illustrated how choosing the value of the trapdoor variable can have substantial effect in estimating the interventional distribution $P(Y|\,\mathrm{do}(X = x))$. Our results suggest that a good default for the trapdoor variable should be based on its conditional distribution given the interventional variable X. On the other hand, for nonlinear and non-Gaussian models the effect of the trapdoor variable can have nonlinear effects on $P(Y|\,\mathrm{do}(X = x))$ which can further bias the results. Therefore as a general sensitivity check, we recommend computing the causal effect using various strategies to account for the trapdoor variables and reporting how sensitive the results are with respect to these strategies.

ACKNOWLEDGEMENTS

This work belongs to the thematic research area 'Decision analytics utilizing causal models and multiobjective optimization' (DEMO) supported by Academy of Finland (grant number 311877). We thank Yonghan Jung for providing example codes for the CWO method, and Satu Helske for providing insights on the intergenerational mobility.

REFERENCES

Acacio-Claro, P.J., Doku, D.T., Koivusilta, L.K. & Rimpelä, A.H. (2018) How socioeconomic circumstances, school achievement and reserve capacity in adolescence predict adult education level: A three-generation study in Finland. *International Journal of Adolescence and Youth*, 23, 382–397. https://doi.org/10.1080/02673843.2017.1389759.

Björklund, A., Eriksson, T., Jäntti, M., Raaum, O. & Österbacka, E. (2002) Brother correlations in earnings in Denmark, Finland, Norway and Sweden compared to the United States. *Journal of Population Economics*, 15, 757–772.

Bürkner, P.-C. & Charpentier, E. (2020) Modelling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology*, 73, 420–451. https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/bmsp.12195.

Bürkner, P.-C. & Vuorre, M. (2019) Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2, 77–101.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M. et al. (2017) Stan: A probabilistic programming language. *Journal of Statistical Software*, 76, 1–32.

Evans, R.J. (2018) Margins of discrete Bayesian networks. *Annals of Statistics*, 46, 2623–2656.

Huang, Y. & Valtorta, M. (2006) Pearl's calculus of intervention is complete. In: *Proceedings of the 22nd conference on uncertainty in artificial intelligence*. 217–224. AUAI Press.

Imbens, G.W. (2000) The role of the propensity score in estimating dose-response functions. *Biometrika*, 87, 706–710.

Joshi, M. & Pustejovsky, J. (2020) Simhelpers: Helper functions for simulation studies.Available at https://CRAN.R-project.org/package=simhelpers. R package version 0.1.0.

Jung, Y., Tian, J. & Bareinboim, E. (2020) Estimating causal effects using weighting-based estimators. In: *Proceedings of the 34th AAAI conference on artificial intelligence*. New York, NY: AAAI Press.

Kline, R.B. (2011) *Principles and practice of structural equation modeling*. New York: Guilford Press.

Koller, D. & Friedman, N. (2009) *Probabilistic graphical models: Principles and techniques*. Cambridge: MIT Press.

Kuusinen, J. (2018) Life course 1971–2002 dataset. Available at https://services.fsd.uta.fi/catalogue/FSD2076?study_language=en. Finnish Social Science Data Archive.

Kuusinen, J. & Leskinen, E. (1988) Latent structure analysis of longitudinal data on relations between intellectual abilities and school achievement. *Multivariate Behavioral Research*, 23, 103–118.

Maclaren, O.J. & Nicholson, R. (2019) What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems. ArXiv preprint arxiv:1904.02826.

Österbacka, E. (2001) Family background and economic status in Finland. *The Scandinavian Journal of Economics*, 103, 467–484. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9442.00255.

Pearl, J. (1988) *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

Pearl, J. (1995) Causal diagrams for empirical research. *Biometrika*, 82, 669–688.

Pearl, J. (2009) *Causality: Models, reasoning, and inference*, 2nd edn. Cambridge: Cambridge University Press.

Pearl, J. (2013) Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference*, 1, 155–170.

Pearl, J. & Mackenzie, D. (2018) The book of why: The new science of cause and effect, 1st edn. New York, NY: Basic Books, Inc.

Pearl, J. & Verma, T.S. (1991) A theory of inferred causation. In: *Principles of knowledge representation and reasoning: Proceedings of the second international conference*, pp. 441–452.

Pfeffer, F.T. (2008) Persistent inequality in educational attainment and its institutional context. *European Sociological Review*, 24, 543–565. https://doi.org/10.1093/esr/jcn026.

R Core Team (2020) *R: A language and environment for statistical computing*. Austria, Vienna: R Foundation for Statistical ComputingAvailable at: https://www.R-project.org/.

Richardson, T.S., Evans, R.J., Robins, J.M. & Shpitser, I. (2017) Nested Markov properties for acyclic directed mixed graphs. ArXiv preprint arxiv:1701.06686.

Robins, J.M. (1986) A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7, 1393–1512.

Robins, J.M., Mark, S.D. & Newey, W.K. (1992) Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48, 479–495.

Rosenbaum, P.R. (1987) Model-based direct adjustment. *Journal of the American Statistical Association*, 82, 387–394.

Rosenbaum, P.R. & Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

Rosseel, Y. (2012) lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, 1–36.

Shimizu, S. (2014) Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41, 65–98.

Shpitser, I. & Pearl, J. (2006) Identification of joint interventional distributions in recursive semi-Markovian causal models. In: *Proceedings of the 21st national conference on artificial intelligence*, vol. 2, AAAI Press, pp. 1219–1226.

Shpitser, I. & Pearl, J. (2008) Dormant independence. In: *Proceedings of the 23rd national conference on artificial intelligence*, vol. 2, pp. 1081–1087.

Shpitser, I., Richardson, T.S. & Robins, J.M. (2009) Testing edges by truncations. In: *International joint conference on artificial intelligence*, vol. 21, pp. 1957–1963.

Shpitser, I., Richardson, T.S. & Robins, J.M. (2011) An efficient algorithm for computing interventional distributions in latent variable causal models. In: *Proceedings of the 27th conference on uncertainty in artificial intelligence*.

Shpitser, I., Evans, R.J., & Richardson, T.S. (2018) Acyclic linear SEMs obey the nested Markov property. In *Proceedings of the 34th conference on uncertainty in artificial intelligence*. AUAI Press.

Stan Development Team. (2019) RStan: the R interface to Stan. Available at: http://mc-stan.org/. R package version 2.19.2.

Tian, J. & Pearl, J. (2002) On the testable implications of causal models with hidden variables. In: *Proceedings of the 18th conference on uncertainty in artificial intelligence*, 519–527. Morgan Kaufmann.

Tikka, S. & Karvanen, J. (2017) Identifying causal effects with the R package causaleffect. *Journal of Statistical Software*, 76, 1–30.

Tikka, S. & Karvanen, J. (2018) Enhancing identification of causal effects by pruning. *Journal of Machine Learning Research*, 18, 1–23.

Tutz, G. (1990) Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43, 39–55.

Verma, T.S. & Pearl, J. (1990) Equivalence and synthesis of causal models. In: *Proceedings of the 6th conference on uncertainty in artificial intelligence*, pp. 255–270.

Wickham, H. (2016) *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. Available at: http://ggplot2.org.

Wright, S. (1934) The method of path coefficients. *Annals of Mathematical Statistics*, 5, 161–215.

How to cite this article: Helske J, Tikka S, Karvanen J. Estimation of causal effects with small data in the presence of trapdoor variables. *J R Stat Soc Series A*. 2021;00:1–22. https://doi.org/10.1111/rssa.12699

APPENDIX A

THEORETICAL PARAMETER ESTIMATES FOR THE LINEAR-GAUSSIAN MODEL OF SECTION 3.3

Marginal covariance matrix Σ of Y, X, Z and W can be found by applying path analysis in the graph of Figure 2c, which leads to

$$\Sigma = \begin{pmatrix} \sigma_{yy} & \cdot & \cdot & \cdot \\ \sigma_{xx}\beta_{yx} + \sigma_{uu}\beta_{xz}\beta_{zw}\beta_{wu}\beta_{yu} & \sigma_{xx} & \cdot & \cdot \\ \sigma_{zz}\beta_{xz}\beta_{yx} + \sigma_{uu}\beta_{zw}\beta_{wu}\beta_{yu} + \sigma_{vv}\beta_{zw}\beta_{wv}\beta_{xv}\beta_{yx} & \sigma_{zz}\beta_{xz} + \sigma_{vv}\beta_{zw}\beta_{wv}\beta_{xv} & \sigma_{zz} & \cdot \\ \sigma_{ww}\beta_{zw}\beta_{xz}\beta_{yx} + \sigma_{uu}\beta_{wu}\beta_{yu} + \sigma_{vv}\beta_{wv}\beta_{xv}\beta_{yx} & \sigma_{ww}\beta_{zw}\beta_{xz} + \sigma_{vv}\beta_{wv}\beta_{xv} & \sigma_{ww}\beta_{zw} & \sigma_{ww}. \end{pmatrix}$$

From Σ, we can obtain the following

$$b_y = \begin{pmatrix} b_{yx} \\ b_{yz} \\ b_{yw} \end{pmatrix} = \Sigma_{1,2:4}\Sigma_{2:4,2:4}^{-1}$$

$$= \begin{pmatrix} \beta_{yx} - \dfrac{\beta_{wu}\beta_{wv}\beta_{xv}\beta_{yu}\sigma_u^2\sigma_v^2}{\beta_{wu}^2\beta_{xv}^2\sigma_u^2\sigma_v^2 + \sigma_w^2(\beta_{xv}^2\sigma_v^2 + \sigma_x^2) + \beta_{wu}^2\sigma_u^2\sigma_x^2 + \beta_{wv}^2\sigma_v^2\sigma_x^2} \\ \beta_{xz}\dfrac{\beta_{wu}\beta_{wv}\beta_{xv}\beta_{yu}\sigma_u^2\sigma_v^2}{\beta_{wu}^2\beta_{xv}^2\sigma_u^2\sigma_v^2 + \sigma_w^2(\beta_{xv}^2\sigma_v^2 + \sigma_x^2) + \beta_{wu}^2\sigma_u^2\sigma_x^2 + \beta_{wv}^2\sigma_v^2\sigma_x^2} \\ \dfrac{\beta_{wu}\beta_{yu}\sigma_u^2(\beta_{xv}^2\sigma_v^2 + \sigma_x^2)}{\beta_{wu}^2\beta_{xv}^2\sigma_u^2\sigma_v^2 + \sigma_w^2(\beta_{xv}^2\sigma_v^2 + \sigma_x^2) + \beta_{wu}^2\sigma_u^2\sigma_x^2 + \beta_{wv}^2\sigma_v^2\sigma_x^2}, \end{pmatrix}$$

$$b_x = \begin{pmatrix} b_{xz} \\ b_{xw} \end{pmatrix} = \Sigma_{2,3:4}\Sigma_{3:4,3:4}^{-1} = \begin{pmatrix} \beta_{xz} \\ \dfrac{\beta_{wv}\beta_{xv}\sigma_v^2}{\beta_{wu}^2\sigma_u^2 + \beta_{wv}^2\sigma_v^2 + \sigma_w^2}, \end{pmatrix}$$

$$a_w = \alpha_w + \beta_{wu}\mu_u + \beta_{wv}\mu_v,$$

$$a_x = \alpha_x + \frac{\beta_{xv}\mu_v(\beta_{wu}^2\sigma_u^2 + \sigma_w^2) - \beta_{wv}\beta_{xv}\sigma_v^2(\alpha_w + \beta_{wu}\mu_u)}{\beta_{wu}^2\sigma_u^2 + \beta_{wv}^2\sigma_v^2 + \sigma_w^2},$$

$$\begin{aligned} a_y = \alpha_y - \{&\beta_{wu}\beta_{yu}(\alpha_w\sigma_u^2\sigma_x^2 + \alpha_w\beta_x v^2\sigma_u^2\sigma_v^2 + \beta_{wv}\mu_v\sigma_u^2\sigma_x^2 - \alpha_x\beta_{wv}\beta_x v\sigma_u^2\sigma_v^2) \\ &-\beta_{yu}\mu_u(\beta_{wv}^2\sigma_v^2\sigma_x^2 + \beta_x v^2\sigma_v^2\sigma_w^2 + \sigma_w^2\sigma_x^2)\}/ \\ &(\beta_{wu}^2\beta_x v^2\sigma_u^2\sigma_v^2 + \beta_{wu}^2\sigma_u^2\sigma_x^2 + \beta_{wv}^2\sigma_v^2\sigma_x^2 + \beta_x v^2\sigma_v^2\sigma_w^2 + \sigma_w^2\sigma_x^2), \end{aligned}$$

$$s_x^2 = \Sigma_{2,2} - \Sigma_{2,3:4}\Sigma_{3:4,3:4}^{-1}\Sigma_{2,3:4}' = \sigma_x^2 + \beta_{xv}^2\sigma_v^2\frac{\beta_{wu}^2\sigma_u^2 + \sigma_w^2}{\beta_{wu}^2\sigma_u^2 + \beta_{wv}^2\sigma_v^2 + \sigma_w^2},$$

$$s_w^2 = \Sigma_{4,4} = \beta_{wu}^2\sigma_u^2 + \beta_{wv}^2\sigma_v^2 + \sigma_w^2.$$