European Journal of Epidemiology
https://doi.org/10.1007/s10654-020-00687-4

TECHNICAL REPORT
R-492
October 2020

ESSAY

Generalizing experimental results by leveraging knowledge of mechanisms

Carlos Cinelli[1] · **Judea Pearl**[1]

Received: 30 December 2019 / Accepted: 22 September 2020

Abstract
We show how experimental results can be generalized across diverse populations by leveraging knowledge of local mechanisms that produce the outcome of interest, only some of which may differ in the target domain. We use structural causal models and a refined version of selection diagrams to represent such knowledge, and to decide whether it entails the invariance of *probabilities of causation* across populations, which then enables generalization. We further provide: (i) bounds for the target effect when some of these conditions are violated; (ii) new identification results for probabilities of causation and the transported causal effect when trials from multiple source domains are available; as well as (iii) a Bayesian approach for estimating the transported causal effect from finite samples. We illustrate these methods both with simulated data and with a real example that transports the effects of Vitamin A supplementation on childhood mortality across different regions.

Keywords Generalizability · Probability of causation · Transportability · Causal inference · Mechanisms

Introduction

Generalizing results of randomized control trials (RCT) is critical in many empirical sciences and demands an understanding of the conditions under which such generalizations are feasible. When the mechanisms that determine the outcome differ between the study population and the target population, generalization requires measuring the variables responsible for such differences or, if this is not possible, isolating them away by measuring other variables [20]. Recent work [9–11] describes an interesting situation under which transportability across populations is feasible without such measurements. This feasibility, however, is not immediately inferable using a standard (non-parametric) selection diagram [1, 20], because it relies on the invariance of only some components of the outcome mechanism, but not all.

In this paper, we use the theory of Structural Causal Models (SCM) [17] to show how generalization in these settings can be modeled using ordinary structural equations, counterfactual logic and selection diagrams. We demonstrate that it requires two key assumptions: (i) the independence of causal factors that affect the outcome; and, (ii) *functional constraints* on how these factors interact to produce the outcome. The combination of these assumptions may entail the invariance of certain *probabilities of causation* [16, 26] across domains, thus allowing the transport of causal effects in settings where non-parametric generalization is otherwise impossible.

We further extend the results of existing literature by: (i) relaxing the monotonicity assumption and providing bounds for the causal effect in the target domain; (ii) deriving novel identification and over-identification results for probabilities of causation, as well as the transported causal effect, when trials from multiple source domains are available; and, (iii) providing a Bayesian framework for estimating the transported causal effect from finite samples. We illustrate these methods both in simulated data and in a real example that

We thank Anders Huitfeldt, Ricardo Silva, and anonymous reviewers for valuable comments and feedback. This research was supported in parts by grants from Defense Advanced Research Projects Agency [#W911NF-16-057], National Science Foundation [#IIS-1302448, #IIS-1527490, and #IIS-1704932], and Office of Naval Research [#N00014-17-S-B001].

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10654-020-00687-4) contains supplementary material, which is available to authorized users.

✉ Carlos Cinelli
carloscinelli@ucla.edu

Judea Pearl
judea@cs.ucla.edu

[1] Departments of Statistics and Computer Science, University of California, Los Angeles, Los Angeles, USA

Published online: 18 October 2020

(a) Coarse causal diagram **(b)** Coarse selection diagram

Fig. 1 Coarse causal (**a**) and selection (**b**) diagrams of the Russian Roulette trial. The presence of $S \rightarrow Y$ in **b** correctly prohibits the naive transportation of the interventional distribution $P(Y_x)$ from the source Π (Los Angeles) to the target environment Π^* (New York)

generalizes the effects of Vitamin A supplementation on childhood mortality across different regions [14, 25, 28]. Open source software for `R` implements the methods discussed in this paper.[1]

Motivating example

To fix ideas, we borrow the "Russian Roulette" example from Huitfeldt [9]. Although stylized, this intuitive example illustrates the key features of the problem.

A Russian Roulette trial

Suppose the city of Los Angeles decides to run a randomized control trial (RCT) to assess the effect of playing "Russian Roulette" on mortality.[2] After running the experiment, the mayor of Los Angeles discovers that "Russian Roulette" is harmful: among those assigned to play Russian Roulette, 17.5% of the people died, as compared to only 1% among those who were not assigned to play the game (people can die due to other causes during the trial, for example, prior poor health conditions).

After hearing the news about the Los Angeles experiment, the mayor of New York City (a dictator) wonders what the overall mortality rate would be if the city forced everyone to play Russian Roulette. Currently, the practice of Russian Roulette is forbidden in New York, and its mortality rate is at 5% (4% higher than LA). The mayor thus asks the city's statistician to decide *whether* and *how* one could use the data from from Los Angeles to predict the mortality rate in New York, once the new policy is implemented.

Intuitively, our causal knowledge of the domain permits us to answer the question posed by the NYC mayor. Mortality is a consequence of two "independent" processes (the game of Russian Roulette and prior health conditions of the individual), and while the first factor remains unaltered across cities, the second intensifies by a known amount (5% vs 1%). Moreover, we can safely assume that the two processes interact disjunctively, namely, that death occurs if and only if at least one of the two processes takes effect. From these two assumptions and elementary probability theory, we can conclude that mortality in NYC would be 20.8%. In the section "Building the structural model" we will cast this intuition into a formal setting, define this notion of "independence," and show how the data from NYC and LA should be combined to match our expectation. But before that, let us examine how this intuition clashes with the conclusion of a coarse analysis using selection diagrams.

An "impossibility" result

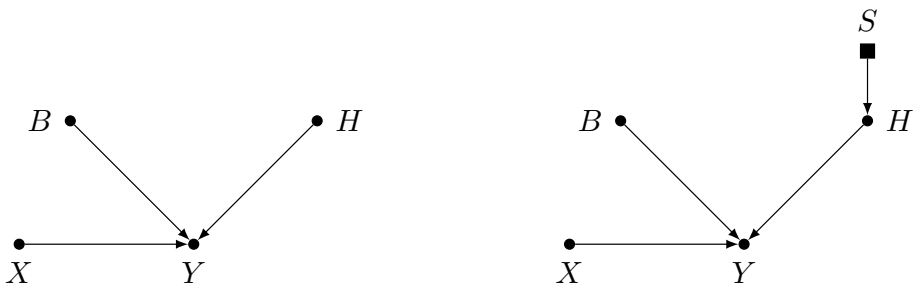
Selection diagrams are causal diagrams enriched with "selection nodes" S, usually represented by square nodes (■). These new nodes are used by the analyst to indicate which *local mechanisms* are suspected to differ between two environments (in our example, the mortality mechanism is suspected to differ between Los Angeles and New York). More importantly, the absence of a selection node pointing to a variable represents the *assumption* that the local mechanism responsible for assigning the value to that variable is the same in the two populations [1, 15, 17, 20].

To build our selection diagram, we need to introduce some notation. The population of Los Angeles will be denoted by Π (the "source population") and that of New York by Π^* (the "target population"). The random variable Y stands for mortality, with events $Y = 1$ denoting "death" and $Y = 0$ denoting "survival;" the random variable X stands for the "treatment" assignment, with events $X = 1$ denoting "play Russian Roulette" and $X = 0$ denoting "not play Russian Roulette." The random variable Y_x denotes the potential response of Y when the treatment X is experimentally set to x. Thus, mathematically, the findings of the RCT can be translated to $P(Y_1 = 1) = 17.5\%$ and $P(Y_0 = 1) = 1\%$, and the available data from New York is $P^*(Y_0 = 1) = 5\%$. Our task is to estimate $P^*(Y_1 = 1)$.

The coarsest causal diagram of the Russian Roulette trial comprises only the treatment X and the outcome Y, as shown in Fig. 1a. To move from the causal diagram to the selection diagram, we need to think of what may differ between LA and NYC. Since we already know from the data that $P(Y_0 = 1) \neq P^*(Y_0 = 1)$, we suspect there are differences in the way mortality is determined in the two cities (for example, people in New York may be in poorer health conditions,

[1] Available in https://github.com/carloscinelli/generalizing.

[2] Russian Roulette consists of loading a bullet into a revolver, spinning the cylinder, pointing the gun at one's own head and then pulling the trigger. We do not recommend attempting this.



(a) Causal diagram **(b)** Selection diagram

Fig. 2 New causal (**a**) and selection (**b**) diagrams explicitly including the variables "health conditions" (H) and "bad luck" (B) when playing Russian Roulette. Here the analyst asserts (using the selection node S) that H may differ between LA and NYC, but assumes that the mechanism triggering B is the same between the two cities. Also important is the absence of a directed edge or a bidirected edge between H and B

or the air quality may be worse). Thus, the selection diagram must contain a selection node S pointing to the mortality variable Y to indicate this disparity, as shown in Fig. 1b.

Graphically, checking whether a causal relationship is transportable from one environment to another involves checking whether there exists a set of measurements that d-separates [17] the source of disparity (the selection node S) from our target quantity. The presence of the selection node pointing directly into Y prevents the separation of S from Y, and leads us to conclude that transportability is impossible without further assumptions. On the other hand, the intuition that led us to predict the new mortality rate in NYC tells us that such assumptions, once formalized, could license transportability. This intuition, as we discussed, was based on two assumptions that are not shown in the coarse selection diagram of Fig. 1. The diagram represents only the existence of a disparity between LA and NYC, not the fact that it is localized to one cause of death (prior health factors), and that it does not extend to the other cause (the game of Russian Roulette). As a result, the diagram correctly warns us that, absent further assumptions, we are not authorized to make any generalization between the two cities.

Building the structural model

We now explicate formally what we know about the game of "Russian Roulette" and health factors, and show how this knowledge renders transportability possible.

Prior health conditions *versus* physical mechanism

To represent the two causes of death, we refine our model by defining two extra random variables, B and H: (i) B denotes "bad luck" when playing Russian Roulette, and its values represent a match ($B = 1$) or mismatch ($B = 0$) between the trigger and the location of the bullet in the cylinder; (ii) and H denotes *all* other health factors producing death ($H = 1$) or survival ($H = 0$). Accordingly, our causal diagram will contain two new edges, $H \to Y$ and $B \to Y$, since both "health conditions" and "bad luck" are key determinants of mortality Y. The updated causal diagram is shown Fig. 2a. Note the absence of a directed or bidirected edge between H and B, which encodes our assumption that these two mechanisms are activated independently of each other.[3]

The new model helps us see more clearly the commonalities and disparities between LA and NYC. First, since there is a multitude of factors that can affect prior health conditions, and those are likely to differ between the two cities (as suggested by the observed difference $P(Y_0 = 1) \neq P^*(Y_0 = 1)$), we again introduce a selection node pointing to H. Moreover, to encode the assumption that the probability of "bad luck" occurring is the same in both cities, we do not connect B to a selection node.[4] The new selection diagram is shown in Fig. 2b.

The diagram of Fig. 2b now guides us toward leveraging the data obtained in LA to make predictions in NYC. If we can find a way to *block the source of disparity originating from* H, we would be left with the invariant physical mechanism shared by both cities. However, since H is unobserved, blockage is impossible without further assumptions. We now ask whether our understanding of how the

[3] The arrow $X \to Y$ comprises, of course, many intermediate mechanisms (such as loading the gun, spinning the cylinder, pulling the trigger) that are not modeled explicitly.

[4] Note that, although reasonable, one cannot take this assumption for granted—it could be the case that revolvers used for Russian Roulette in New York have a different number of chambers than those used in Los Angeles. *The absence of a selection node* pointing to B encodes the *assumption* that this is not the case.

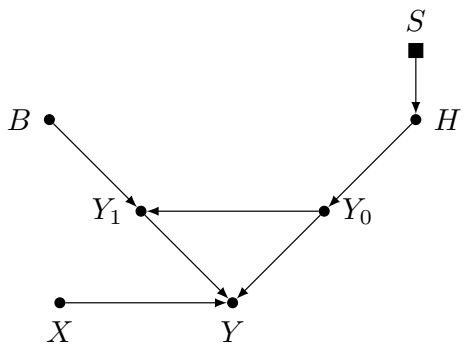


Fig. 3 Selection diagram explicitly showing the potential outcomes Y_0 and Y_1 as implied by the functional constraints. Note that $Y_1 \perp S \mid Y_0$

two mechanisms interact in producing Y would permit us to estimate $P^*(Y_1 = 1)$.

Leveraging functional constraints

Our understanding that mortality is caused by *either one* of the two processes (prior health conditions or bad luck in the game), dictates the following *functional specification* for the *structural equation* of Y,

$$Y = H \vee (X \wedge B) \tag{1}$$

Where $\vee$ denotes the logical "or" operator, and $\wedge$ denotes the logical "and" operator. Like any structural equation, Eq. 1 defines the potential outcomes Y_0 and Y_1 [17, Ch.7] which we may now find useful to encode explicitly. Its first implication is that $Y_0 = H$ and $Y_1 = H \vee B = Y_0 \vee B$. This tells us that, once we know the potential response of units under no treatment (Y_0) we do not need to know anything else about their previous health condition (H) to determine the value of Y_1—B would suffice.[5] We can represent this fact in a modified selection diagram, in which the potential outcomes are now also shown explicitly (Fig. 3). The diagram reveals that Y_0 blocks the source of health disparities between the two populations, and we conclude that $Y_1 \perp S \mid Y_0$.[6]

More concretely, consider the counterfactual quantity

$$\mathrm{PS}_{01} := P(Y_1 = 1 \mid Y_0 = 0)$$

which stands for the share of people who would die if forced to play Russian Roulette, among those who would not have died if not forced to do so. In other words, PS_{01} represents the probability that the game of Russian Roulette is *sufficient* to *kill* a person *during the trial.* The acronym PS_{01} was chosen to emphasize its relation to the "probability of sufficiency" (PS), $\mathrm{PS} = P(Y_1 = 1 | Y = 0, X = 0)$, as defined and analyzed in [16] and [26]. In our context, since the treatment is randomized, the two quantities coincide,

$$\begin{aligned} P(Y_1 = 1 | Y_0 = 0) &= P(Y_1 = 1 | Y_0 = 0, X = 0) \\ &= P(Y_1 = 1 | Y = 0, X = 0) \end{aligned}$$

where the first equality is licensed by the randomization of X and the second equality is due to consistency. In general, however, PS_{01} need not be the same as PS—the later measures the probability of fatal treatment among those who, given the choice, would *choose* not to be treated and survive; the former measures the probability of fatal treatment among those who would survive had they not been *assigned* for treatment.[7] Similar reasoning holds for $\mathrm{PS}_{10} := P(Y_1 = 0 \mid Y_0 = 1)$, which stands for the probability that playing Russian Roulette is *sufficient* to *save* a person who would die if denied treatment. In our example, this probability is obviously zero as we shall formally show below. The condition $Y_1 \perp S \mid Y_0$, implied by the diagram, states that these *probabilities of causation* are invariant across cities.[8] This feature of invariance, which is important in its own right, follows solely from our structural assumption about the mechanisms involved.

A second implication of Eq. 1 is that the treatment effect is *monotonic*, that is $Y_1 \geq Y_0$ for all individuals. This, in turn, implies $\mathrm{PS}_{10} = 0$; in other words, an individual that would have died of other causes during the trial, would still die if forced to play Russian Roulette. It has been shown that monotonicity is sufficient for identifying PS_{01} in this setting [10, 16, 26]. Indeed, by the law of total probability,

$$P(Y_1 = 1) = (1 - \mathrm{PS}_{10})P(Y_0 = 1) + \mathrm{PS}_{01}(1 - P(Y_0 = 1))$$

[5] Although here we have $Y_0 = H$ for simplicity, this need not be the case. The same argument would hold, for instance, if we define H to be a random variable with arbitrary cardinality and $Y = g(H) \vee (X \wedge B)$, where $g(H) \in \{0, 1\}$. Likewise, "see the appendix" for an example where the treatment variable X is continuous and the same strategy adopted here can be employed.

[6] Since some relationships in the graph may be deterministic, conditional independencies other than those revealed by *d*-separation (with lower-case d) may be present. A complete criterion for DAGs with deterministic nodes is given by the *D*-separation criterion (with capital D) of [5]. Moreover, note arrows between potential outcomes need not convey causal influence; their purpose is merely to ensure that the correct conditional independencies among variables are encoded in the graph, *as derived from the structural equations*. Finally, here we are not treating the question of how scientists acquire scientific knowledge in the form of a functional specification such as Eq. 1. Rather, our task is more modest: given that scientists sometimes have knowledge of mechanisms, how can we leverage some of that knowledge for identification.

[7] For example, in legal settings, where acts are executed by *choice*, conditioning on the *observed* X gives a more appropriate measure of an agent's responsibility, as argued in Pearl [17, Ch. 9] and Pearl [18].

[8] Probabilities of causation have been extensively studied elsewhere under a different context. See [16, 17, 26].

The quantity $P(Y_0 = 1)$ is given from the RCT (1%) and, due to monotonicity, $\text{PS}_{10} = 0$. Thus, we have:

$$PS_{01} = \frac{P(Y_1 = 1) - P(Y_0 = 1)}{1 - P(Y_0 = 1)} = \frac{17.5\% - 1\%}{99\%} = 1/6$$

This is not surprising; the probability that the "treatment" is *sufficient* to kill an individual who would have otherwise survived indeed equals 1/6—the probability of having "bad luck" in the game of Russian Roulette, using a revolver with six chambers.[9]

Thus far we have established that $\text{PS}_{10} = \text{PS}^*_{10}$, $\text{PS}_{01} = \text{PS}^*_{01}$, and that $\text{PS}_{10} = 0$, $\text{PS}_{01} = 1/6$. Combining these results with the current baseline mortality from NYC, that is, $P^*(Y_0 = 1) = 5\%$, we can finally evaluate our target quantity $P^*(Y_1 = 1)$,

$$\begin{aligned} P^*(Y_1 = 1) &= (1 - \text{PS}^*_{10})P^*(Y_0 = 1) + \text{PS}^*_{01}(1 - P^*(Y_0 = 1)) \\ &= (1 - \text{PS}_{10})(5\%) + \text{PS}_{01}(95\%) \\ &= (1)(5\%) + (1/6)(95\%) = 20.8\% \end{aligned}$$

Which matches the intuitive answer obtained before.

As a brief remark, note that, if instead of $Y_1 \perp\!\!\!\perp S \mid Y_0$ we had obtained the condition $Y_0 \perp\!\!\!\perp S \mid Y_1$, we would conclude that the probabilities $\text{PN}_{01} := P(Y_0 = 0 \mid Y_1 = 1)$ and $\text{PN}_{10} := P(Y_0 = 0 \mid Y_1 = 1)$ are the same across trials. These quantities represent the probability that the treatment is *necessary* for causing (PN_{01}) or preventing (PN_{10}) the outcome during the experiment. All results of this paper hold in this setting, with minor modifications. Therefore, for simplicity of exposition, in the remainder of the text we discuss the case of $Y_1 \perp\!\!\!\perp S \mid Y_0$ only.[10]

[9] The right-hand side of this expression is known as the "relative difference," or "susceptibility." Simple algebra shows that $\frac{P(Y_1=1)-P(Y_0=1)}{1-P(Y_0=1)} = 1 - \frac{1-P(Y_1=1)}{1-P(Y_0=1)}$, where the quantity $\frac{1-P(Y_1=1)}{1-P(Y_0=1)}$ is known as the "survival ratio." Since under the assumption of monotonicity these estimands identify PS_{01}, and PS_{01} is invariant across domains, it thus follows that the "relative difference" and the "survival ratio" will also be equal between populations. Huitfeldt et al. [10] suggested using this fact as a rationale for assuming homogeneity of effect measures across domains, a common heuristic among epidemiologists for approaching generalizability problems. These equivalences, however, break down without monotonicity; in that case, the "relative difference" is a lower bound for the probability of sufficiency [26], as we discuss next.

[10] For example, under the assumption of monotonicity, we have that $\text{PN}_{01} = \frac{P(Y_1=1)-P(Y_0=1)}{P(Y_1=1)}$ [16]. This last estimand is known as the "excess-risk-ratio," and algebra also shows that $\frac{P(Y_1=1)-P(Y_0=1)}{P(Y_1=1)} = 1 - \frac{1}{P(Y_1=1)/P(Y_0=1)}$, where $\frac{P(Y_1=1)}{P(Y_0=1)}$ is the "risk ratio." Thus in this setting, both the "excess-risk-ratio" and the "risk ratio" would be equal across domains. Without monotonicity, the "excess-risk-ratio" is a lower bound on the probability of necessity [26].

Bounds without monotonicity

A key step in obtaining a point estimate for $P^*(Y_1 = 1)$ was the monotonicity property, which emanates from the functional form of Eq. 1. Monotonicity allowed us to identify the probabilities of sufficiency PS_{01} and PS_{10}, which, as advertised by the assumptions in the selection diagram of Fig. 3, are invariant across domains. The monotonicity property holds trivially in our example of the Russian Roulette, when Y represents death, but it may not hold for other outcomes or, more generally, it may not hold in contexts beyond our stylized example.

Remarkably, however, even in the absence of monotonicity, one can still assess the transported causal effect, albeit in the form of a *bound*. The next theorem shows that the counterfactual independence $Y_1 \perp\!\!\!\perp S \mid Y_0$ by itself is strong enough for bounding the causal effect in the target domain. These results improve the bias analysis performed by Huitfeldt et al. [10], and provide an exact characterization of the inferences compatible with the assumption of $Y_1 \perp\!\!\!\perp S \mid Y_0$.

Theorem 1 *Consider a source domain* Π *and a target domain* Π^*. *Let* $P_{ij} := P(Y_i = j)$, $P^*_{ij} := P^*(Y_i = j)$, *and let* $RR = \frac{P_{11}}{P_{01}}$ *denote the risk-ratio in the trial of the source domain* Π. *If* $Y_1 \perp\!\!\!\perp S \mid Y_0$, *then* P^*_{11} *of* Π^* *is bounded by* $P^{*L}_{11} \le P^*_{11} \le P^{*U}_{11}$, *with,*

$$P^{*L}_{11} = RR \times P^*_{01} + \min\left\{\left(\frac{P_{01} - P^*_{01}}{P_{01}}\right)\text{PS}^L_{01}, \left(\frac{P_{01} - P^*_{01}}{P_{01}}\right)\text{PS}^U_{01}\right\},$$

$$P^{*U}_{11} = RR \times P^*_{01} + \max\left\{\left(\frac{P_{01} - P^*_{01}}{P_{01}}\right)\text{PS}^L_{01}, \left(\frac{P_{01} - P^*_{01}}{P_{01}}\right)\text{PS}^U_{01}\right\}$$

where $\text{PS}^L_{01} = \max\left\{0, \frac{P_{11}-P_{01}}{1-P_{01}}\right\}$ *and* $\text{PS}^U_{01} = \min\left\{\frac{P_{11}}{1-P_{01}}, 1\right\}$ *are the lower and upper bounds on* PS_{01}, *respectively.*

Proof The bounds are obtained by solving a linear optimization problem, as detailed in the appendix. □

Theorem 1 can be better understood as a two-stage process. First, with a little algebra, it is possible to re-express $P^*(Y_1 = 1)$ as a function of PS_{01} alone, resulting in,

$$P^*(Y_1 = 1) = RR \times P^*(Y_0 = 1) + \left(\frac{P(Y_0 = 1) - P^*(Y_0 = 1)}{P(Y_0 = 1)}\right)\text{PS}_{01} \tag{2}$$

Where $RR = P(Y_1 = 1)/P(Y_0 = 1)$ denotes the *risk-ratio* obtained in the trial of the source domain Π. The first term of this expression, $RR \times P^*(Y_0 = 1)$, consists of the "naive" prediction for $P^*(Y_1 = 1)$ that one would have obtained by assuming a constant risk ratio across populations. The second term adjusts this naive prediction, by taking into account both the excess risk-ratio of contrasting the baseline

mortality between Π and Π^*, as well as the probability of sufficiency shared across environments, PS_{01}.

After this, note that, although the probability of sufficiency PS_{01} in Eq. 2 cannot be point identified, it can be bounded by (see the appendix as well as [26])

$$\max\left\{0, \frac{P(Y_1=1)-P(Y_0=1)}{1-P(Y_0=1)}\right\} \leq \text{PS}_{01} \leq \min\left\{\frac{P(Y_1=1)}{1-P(Y_0=1)}, 1\right\} \tag{3}$$

Thus, by substituting PS_{01} with its bounds, we obtain the desired bounds for the target quantity $P^*(Y_1=1)$.

For instance, in our Russian Roulette example, regardless of whether monotonicity holds, PS_{01} can be bounded by

$$16.7\% \leq \text{PS}_{01} \leq 17.7\%$$

And this assures us that $P^*(Y_1=1)$ must lie between,

$$16.8\% \leq P^*(Y_1=1) \leq 20.8\%$$

To put it another way, the results of the trial in LA tells us that implementing the policy in NYC would cause *at least* an increase of $16.8\% - 5\% = 11.8\%$ and *at most* an increase of $20.8\% - 5\% = 15.8\%$ in mortality. Note that, here, substituting the lower bound for PS_{01} (16.7%) actually translates to the *upper bound* for $P^*(Y_1=1)$ (20.8%). This happens because the baseline risk in the target population Π^* is *higher* than that of the source population Π, and thus the adjustment due to PS_{01}, in Eq. 2, is negative.

These considerations naturally lead to the question: in general, how informative are the bounds on $P^*(Y_1=1)$? It turns out that the width of the bounds have a simple characterization. Consider the case in which the bounds for PS_{01} are not zero nor one. Now let $P^{*U}(Y_1=1)$ and $P^{*L}(Y_1=1)$ denote the upper and lower bound on $P^*(Y_1=1)$, respectively. After some algebra, it is possible to show that (see the appendix),

$$P^{*U}(Y_1=1) - P^{*L}(Y_1=1) = \frac{|P(Y_0=1) - P^*(Y_0=1)|}{1-P(Y_0=1)}$$

That is, in this setting, the width of the bounds depends on the baseline risks $P(Y_0=1)$ and $P^*(Y_0=1)$ alone. Moreover, even if the bounds for PS_{01} happen to be "wide," if the baseline risks are close enough across populations, the bounds for $P^*(Y_1=1)$ can still be "narrow." In the section "A Bayesian approach to estimation" we illustrate this fact with a real data example in which the bounds are narrow enough to imply a positive effect of the treatment.

Identification with trials from multiple source domains

In Theorem 1 we learned that the existence of experimental data from *one* source population leads to bounds on the transported causal effect of the target population, although it is not enough for its point identification. Surprisingly, however, if we can obtain experimental data from an additional source population, this suffices to change the picture. With *two* source trials, it is possible to obtain a point estimate for the probabilities of sufficiency, and, consequently, for $P^*(Y_1=1)$ without invoking monotonicity, nor any further assumptions beyond $Y_1 \perp S \mid Y_0$. Moreover, multiple source trials entail strong testable implications that can be used to *falsify* this "cross-world" assumption.[11]

To illustrate, consider our Russian Roulette example, and suppose we learn that the city of Chicago has also performed an RCT. In that trial, 25% of those assigned to play the game died, in contrast to 10% of those not assigned to play. If the selection diagram contrasting NYC with Chicago is the same as that of Fig. 3, we can combine the results from LA and Chicago to estimate the probabilities of sufficiency shared across cities. By the law of total probability, expand the expression for $P(Y_1=1)$, both for LA and Chicago, to obtain a system of two equations and two unknowns:

$$\text{(LA):} \quad 0.175 = (1-\text{PS}_{10}) \times 0.01 + \text{PS}_{01} \times 0.99$$

$$\text{(Chicago):} \quad 0.250 = (1-\text{PS}_{10}) \times 0.10 + \text{PS}_{01} \times 0.90$$

This system can then be solved for PS_{10} and PS_{01}

$$\text{PS}_{10} = 0, \qquad \text{PS}_{01} = 1/6$$

Put differently, the *only* values for PS_{10} and PS_{01} that are compatible with the observed data from *both* trials (LA and Chicago) are that: (i) the "treatment" cannot save anyone from dying; and, that (ii) the treatment kills 1/6 of those who would not have died otherwise. These are the same numeric values as before, but with an important difference—we did not assume monotonicity to obtain point identification; instead, we learned *from the data* that the treatment effect must be monotonic. Once we have these numbers, we can use the same strategy as before to predict the causal effect in NYC, which amounts to, again, 20.8%.

[11] Similar observations regarding testable implications when combining information from multiple studies have also been made in Hartman et al. [8], Lu et al. [12] and Dahabreh et al. [4].

Furthermore, since PS_{10} and PS_{01} must be valid probabilities, not all observed values are compatible with the assumption that $Y_1 \perp S \mid Y_0$. For instance, suppose that instead of 10%, the observed baseline mortality rate in Chicago were 5%. This would imply the impossible value $\text{PS}_{10} = -1.03$, thus *falsifying* the assumption of invariance across domains. It is also easy to see that with three or more source domains we obtain over-identification, since each population pair implies different estimates for PS_{10} and PS_{01}. If those estimates are discordant, this calls into question the assumption of $Y_1 \perp S \mid Y_0$. These results are somewhat reassuring. They tell us that, despite its "cross-world" nature, the assumption of invariance of probabilities of causation across domains may have strong testable implications, and can thus be subjected to empirical scrutiny.

We formalize the previous considerations with the next two theorems.

Theorem 2 *Consider two source domains* Π^a *and* Π^b. *Let the probabilities of sufficiency be the same across the two populations, that is,* $\text{PS}^a_{01} = \text{PS}^b_{01} = \text{PS}_{01}$ *and* $\text{PS}^a_{10} = \text{PS}^b_{10} = \text{PS}_{10}$. *Then,*

$$\text{PS}_{10} = 1 - \frac{P^a_{11}P^b_{00} - P^b_{11}P^a_{00}}{P^a_{01}P^b_{00} - P^b_{01}P^a_{00}}$$

$$\text{PS}_{01} = \frac{P^b_{11}P^a_{01} - P^a_{11}P^b_{01}}{P^a_{01}P^b_{00} - P^b_{01}P^a_{00}}$$

Where $P^a_{ij} := P^a(Y_i = j)$ *and* $P^b_{ij} := P^b(Y_i = j)$. *Moreover, the experimental probabilities of necessity, and probability of necessity and sufficiency* [26] *of both populations are also identifiable from experimental data of* Π^a *and* Π^b.

Proof As explained in the text, we can use the law of total probability for each domain to obtain two linear equations with two unknowns, PS_{01} and PS_{10}. We can thus (generically) solve the system of equations for those quantities. Interestingly, in this setting, not only the probabilities of sufficiency, but *all* remaining probabilities of causation (as discussed in [26]), are also identifiable. See details in the appendix. □

Next, the causal effect for a target population Π^* can be transported by appealing again to the law of total probability.

Theorem 3 *Consider two source domains* Π^a, Π^b, *and a target domain* Π^*. *Let the probabilities of sufficiency be the same across populations, that is,* $\text{PS}^a_{01} = \text{PS}^b_{01} = \text{PS}^*_{01}$ *and* $\text{PS}^a_{10} = \text{PS}^b_{10} = \text{PS}^*_{10}$. *Then, the causal effect* P^*_{11} *in* Π^* *is given by,*

$$P^*_{11} = \frac{P^a_{11}P^b_{00} - P^b_{11}P^a_{00}}{P^a_{01}P^b_{00} - P^b_{01}P^a_{00}} \times P^*_{01} + \frac{P^b_{11}P^a_{01} - P^a_{11}P^b_{01}}{P^a_{01}P^b_{00} - P^b_{01}P^a_{00}} \times P^*_{00}$$

A Bayesian approach to estimation

The previous results focused on *identification*, that is, they are "asymptotic," and assume that the measured quantities are representative of their corresponding quantities in the population. In practice, however, researchers need to take sampling uncertainty into account. In this section, we describe a Bayesian framework that practitioners can easily put to use for finite sample inference. A Bayesian approach is especially suited for this setting—when the target quantity $P^*(Y_1 = 1)$ is not identifiable from the data alone, preference for any value of the parameter within the identified bounds must rely on prior knowledge.

Model specification

The Bayesian specification of our model can be simplified if we use *counts*. For the source population Π, let n_0 denote the *sum* of individuals with $Y = 1$ in the control group, and let n_1 denote the *sum* of individuals with $Y = 1$ in the treatment group. Likewise, let n^*_0 and n^*_1 denote those quantities for the target population Π^*. Note that n^*_1 is not observed, since the target population is under the "no-treatment" regime.

Now let us use the same notation of Theorem 1 to denote population parameters, that is: $P_{11} := P(Y_1 = 1)$, $P_{01} := P(Y_0 = 1)$, $P^*_{01} := P^*(Y_0 = 1)$, $P^*_{11} := P^*(Y_1 = 1)$. Given that the outcome variable Y is binary, the sum of individuals with $Y = 1$ follows a binomial distribution, and we can write the model for the observed data $\mathcal{D} = \{n_0, n_1, n^*_0\}$ as,

$$n_0 \sim \text{Binomial}(N_0, P_{01})$$

$$n_1 \sim \text{Binomial}(N_1, P_{11})$$

$$n^*_0 \sim \text{Binomial}(N^*_0, P^*_{01})$$

where N_0 denotes the total number of individuals in the control arm, and N_1 the total number of individuals in the treatment arm of the trial in the source population; N^*_0 denotes the total sample size of the target population (which is under the no-treatment regime). We treat N_0, N_1 and N^*_0 as *known* fixed quantities. Note the observed data depends *only* on the parameters P_{01}, P_{11} and P^*_{01}.

We now need to specify the prior distribution of the parameters and the target quantities of interest. Here we describe two general alternatives, depending on whether

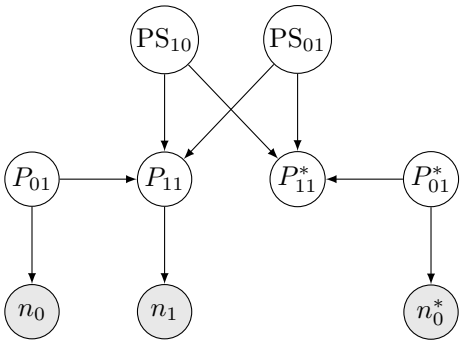


Fig. 4 Probabilistic graphical model for Bayesian inference when the quantity of interest is P^*_{11}. Gray nodes (n_0, n_1, n^*_0) denote observed variables. White notes denote latent parameters (P_{01}, P_{11}, PS_{10}, PS_{01}, P^*_{11}, P^*_{01}). Note that P_{11} and P^*_{11} share the parameters PS_{10} and PS_{01}, which are invariant across populations

the researcher is interested in making inferences directly on P^*_{11} (which in general will not be identified from the data), or on its bounds (which are identified)—we believe these two approaches are complementary, and we encourage investigators to explore both options (see also [7, 23, 24]).

Inference on P^*_{11}. As discussed in the previous section, we have that P_{11} is a deterministic function of PS_{10}, PS_{01} and P_{01}, that is, $P_{11} = (1 - \text{PS}_{10})P_{01} + \text{PS}_{01}(1 - P_{01})$. Therefore, we need only to specify priors for the parameters P_{01}, P^*_{01}, PS_{10} and PS_{01}. For example, an "uninformative" (or "flat") prior consists of a uniform distribution over 0 and 1 for all parameters. Another option is to choose a prior that incorporates the assumption of monotonicity, by setting a point mass on $\text{PS}_{10} = 0$. Users have the flexibility of picking anything in between, such as setting a prior that puts most, but not all, of the mass on $\text{PS}_{10} = 0$, for instance. The target of inference is the *posterior distribution* of P^*_{11}, which is, again, a transformation of the parameters P^*_{01}, PS_{10} and PS_{01},

$$P^*_{11} = (1 - \text{PS}_{10})P^*_{01} + \text{PS}_{01}(1 - P^*_{01})$$

As we shall see, with a "flat" prior, as the sample size increases the posterior distribution converges to the identified bounds; whereas with a prior that assumes monotonicity the posterior converges to the identified point estimate. Other quantities of interest may be the posterior distribution of certain *effect measures*, such as the risk difference $RD^* = P^*_{11} - P^*_{01}$ or the risk ratio $RR^* = P^*_{11}/P^*_{01}$. Figure 4 shows the probabilistic graphical model of this setup, with observed variables in gray, and latent parameters in white. The known fixed parameters N_0, N_1 and N^*_0 are omitted for clarity.

Inference on bounds. When making inferences on P^*_{11} (which is not identified), the shape of its posterior will be dependent on (but not completely determined by) the shape of the prior of the unidentified quantities PS_{01} and PS_{10}, regardless of sample size. For this reason, users may also find useful to perform inference directly on the bounds P^{*L}_{11} and P^{*U}_{11} (which are identified). While the previous framework can still be used for such inferences, we note that, if interest lies on the bounds alone, there is a simpler alternative—as the bounds are functionals of the observed data, inference about P^{*L}_{11} and P^{*U}_{11} only requires priors on the identified parameters P_{01}, P_{11} and P^*_{01} [23, 24].

Sampling. Given the observed data $\mathcal{D}$ and a prior distribution on the parameters, one can obtain the posterior distribution of the target quantities using Gibbs sampling. Here we use the Gibbs sampler JAGS [22]. Extending the model to two (or more) source populations follows the same logic, thus we defer its discussion to the appendix. Next, we demonstrate the method using: (i) simulated data from the Russian Roulette example; and, (ii) real data from trials that investigate the effects of vitamin A supplementation on childhood mortality. Code for replicating all results is also provided in the online supplemental material.

Simulated data example

To illustrate the method, we start by applying our tools to simulated data drawn from a process with the same proportions as the Russian Roulette example, with various sample sizes. We show the posterior distribution of $P^*(Y_1 = 1)$ using both a "flat" prior for all parameters, and a prior assuming monotonicity. The results are shown in Figs. 5 and 6.

Let us start by examining Fig. 5. Here we set "flat" priors for *all* parameters. Note that, as per Theorem 1, the posterior distribution remains spread in the asymptotic bounds of 16.8% and 20.8% regardless of sample size. Moving to Fig. 6, we now set a point mass prior on $\text{PS}_{10} = 0$, representing the assumption of monotonicity. The remaining parameters continue to have a "flat" prior. As expected, the posterior distribution now concentrates around 20.8% as the number of cases increases.

Real data example

We now illustrate our method with a real data example. We investigate three experiments designed to determine the effects of vitamin A supplementation on childhood mortality. The first trial was carried out in the Aceh province at the northern tip of Sumatra, Indonesia [25]; the second trial was conducted in the West Java province, in Java, also in Indonesia [14]. Finally, the third trial took place in the district of Sarlahi, Nepal [28]. The results from the studies are shown in Table 1. Our exercise in this section consists of using the results of earlier trials, along with the baseline risk of the target population, to predict mortality under treatment in the target population.

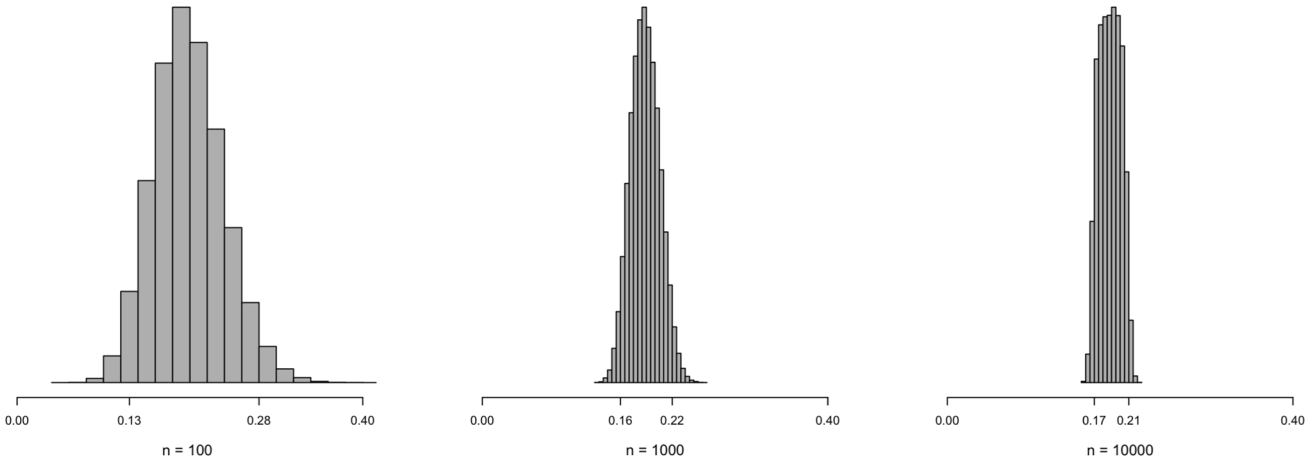


Fig. 5 Histograms of the posterior samples of $P^*(Y_1 = 1)$ for a simulation of the Russian Roulette data, considering different sample sizes 100, 1000 and 10,000. Here all parameters have a "flat" prior. Note that, as the sample size increases, the posterior distribution does not concentrate on a point; rather, the posterior remains spread on the identified bound of 16.8% to 20.8%, as per Theorem 1

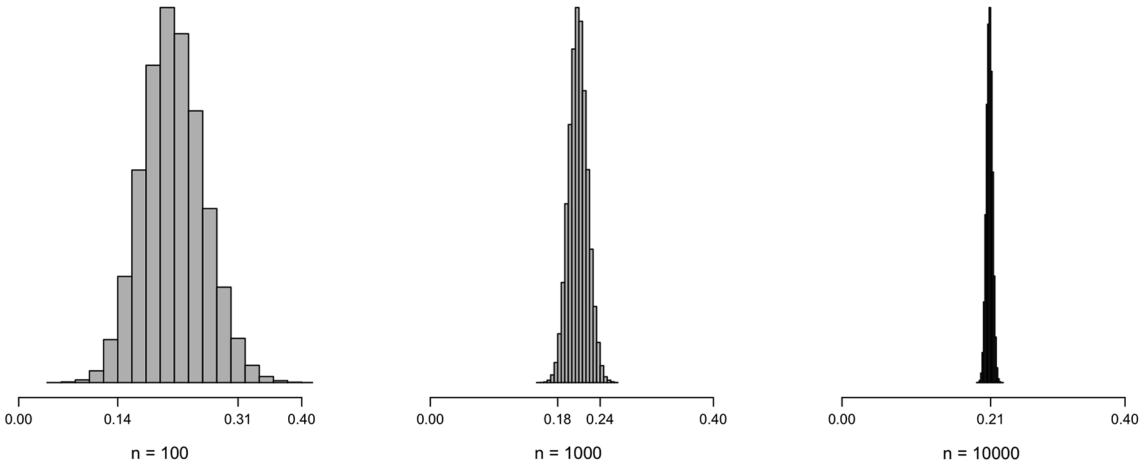


Fig. 6 Histograms of the posterior samples of $P^*(Y_1 = 1)$ for a simulation of the Russian Roulette data, considering different sample sizes 100, 1000 and 10,000. Here we put a point mass prior on PS_{10}, corresponding to the assumption of monotonicity. The remaining parameters have a "flat" prior. Note that, as the sample size increases, the posterior distribution concentrates on 20.8%, since the parameter is identifiable in this setting

Table 1 Observed data for the vitamin A studies

Study	Treatment		Control	
	Survived	Total	Survived	Total
Aceh [25]	12,890	12,991	12,079	12,209
West Java [14]	5589	5775	5195	5445
Sarlahi [28]	14,335	14,487	13,933	14,143

It is suspected that vitamin A reduces childhood mortality by reducing the incidence, severity or duration of life-threatening diseases such as measles and diarrhoea [28]. As a *first approximation* to this process, we can borrow the same disjunctive model of the previous section. The variables now mean: (i) $Y = 1$ *survival*, and $Y = 0$ death during the trial; (ii) $H = 1$ *absence*, and $H = 0$ presence of severe measles; (iii) $X = 1$ participation in the treatment group (vitamin A supplementation), and $X = 0$ participation in the control group; finally, (iv) B summarizes biological factors that determine the response to treatment ($B = 1$ successful response, $B = 0$ otherwise). Here the monotonicity assumption states that vitamin A supplementation *does not* cause deaths. After presenting the results of our method, we discuss cases under which these assumptions may be violated, thus preventing one from inferring $Y_1 \perp\!\!\!\perp S \mid Y_0$.

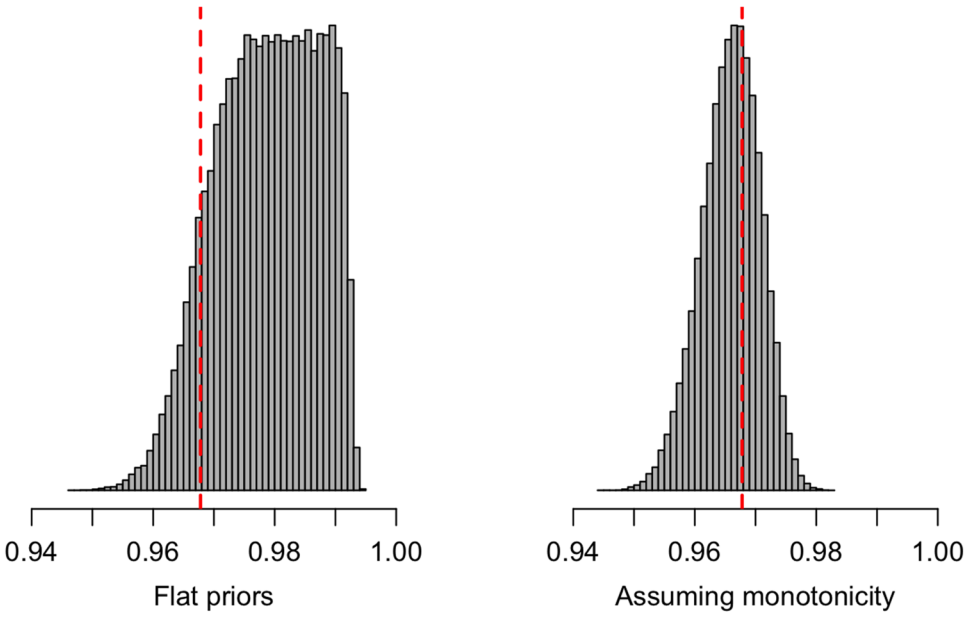


Fig. 7 Posterior of $P^{WJ}(Y_1 = 1)$ for the West Java trial, using data from the Aceh trial. Left: posterior of $P^{WJ}(Y_1 = 1)$ using "flat" priors. Right: posterior of $P^{WJ}(Y_1 = 1)$ assuming monotonicity. Red dashed lines show the observed value in the West Java trial, $\widehat{P}^{WJ}(Y_1 = 1)$

Our first task is to use the results of the Aceh trial (Π^A) to predict the effects of the West Java trial (Π^{WJ}). The estimates of the Aceh trial are $\widehat{P}^A(Y_1 = 1) = 0.992$ and $\widehat{P}^A(Y_0 = 1) = 0.989$; whereas the baseline risk in the Java trial is $\widehat{P}^{WJ}(Y_0 = 1) = 0.954$. As expected, note the large discrepancy of baseline risk in both trials, indicating the existence of structural differences in how mortality is determined, and thus forbidding a direct transport of $P^{WJ}(Y_1 = 1)$. Figure 7 shows the posterior distribution of $P^{WJ}(Y_1 = 1)$ using both a "flat" prior for all parameters (left), and a prior assuming monotonicity for the effect of vitamin A supplementation (right). In the first case, we obtain a 95% *credible interval* of 0.962 to 0.992 for $P^{WJ}(Y_1 = 1)$, in agreement with the asymptotic bounds of Theorem 1—this shows that, even without assuming monotonicity, the bounds are narrow enough to be consistent with a positive effect of vitamin A supplementation in West Java.[12] When assuming a monotonic effect of vitamin A, we obtain the posterior mean of 0.967 (95% CI 0.956–0.975). In both plots, a red dashed line indicates the actual value observed in the West Java trial, $\widehat{P}^{WJ}(Y_1 = 1) = 0.968$, which is consistent with the predictions of our method.

Our second task is to use the results of *both* the Aceh (Π^A) and West Java (Π^{WJ}) trials to predict the effects of the Sarlahi trial (Π^S). As per Theorems 2 and 3, in this setting we can identify the probabilities of sufficiency shared across regions, PS_{10} and PS_{01}, as well as the effect in Sarlahi, $P^S(Y_1 = 1)$, *without* assuming monotonicity. The posterior distributions of these three quantities are displayed in Fig. 8. The posterior mean for PS_{01} is 0.346 (95% CI 0.214–0.478), while the posterior mean for PS_{10} is 0.001 (95% CI 0.000–0.004). This suggests that, in the context of these trials, vitamin A supplementation is sufficient to prevent 21% to 48% of the deaths that would have otherwise occurred without supplementation, while it has no or little side-effects that are sufficient to cause the death of otherwise healthy subjects. Finally, we obtain the posterior mean of 0.989 (95% CI 0.987–0.991) for $P^S(Y_1 = 1)$, consistent with the actual value observed in the Sarlahi trial, $\widehat{P}^S(Y_1 = 1) = 0.989$.

Before moving to the conclusions, let us use this example to make some brief remarks about causal modeling in practice. Note that the working model in this section assumes the only factor causing deaths during the period of the trial can be summarized by H, consisting of diseases which, at least in principle, can be affected by the treatment (e.g, severe measles or diarrhoea). What happens, however, if we augment the model to allow for other causes of deaths unaffected by vitamin A supplementation? It can be shown that this new variable is a common cause of both potential responses, thus creating a colliding path and forbidding the conclusion that $Y_1 \perp\!\!\!\perp S \mid Y_0$.[13] This suggests caution when transporting these

[12] The 95% credible intervals for the risk difference and risk ratio are 0.008–0.04 and 1.009–1.042, respectively. Alternatively, if one prefers inferences on the bounds, we have 95% credible intervals of: 0.955–0.975 for the lower bound, 0.991–0.994 for the upper bound, and 0.002–0.020 for the lower bound of the risk difference (i.e, $P^{*L}_{11} - P^{*}_{01}$).

[13] Call these new causes C. The new structural equation for Y now reads $Y = (H \vee (X \wedge B)) \wedge \neg C$. This leads to $Y_0 = H \wedge \neg C$ and $Y_1 = Y_0 \vee (B \wedge \neg C)$. Note this creates the colliding path $S \rightarrow H \rightarrow Y_0 \leftarrow C \rightarrow Y_1$, thus forbidding the conclusion that $Y_1 \perp\!\!\!\perp S \mid Y_0$, even when there is no selection node pointing directly to C. For another illustration of when collider bias may arise, see the appendix.

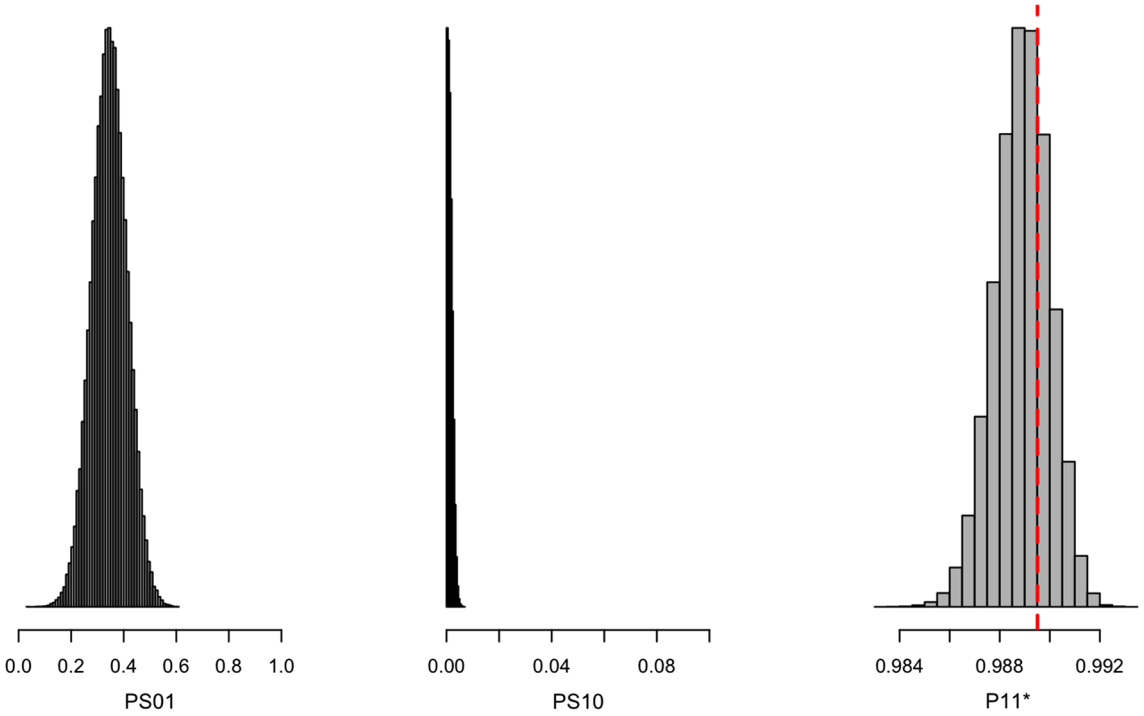


Fig. 8 From left to right, posterior of PS_{01}, PS_{10} and $P^S(Y_1 = 1)$ using data from *both* the Aceh and West Java trials [14, 25], and using "flat" priors for all parameters. Dashed red line indicates the observed value in the Sarlahi trial, $\widehat{P}^S(Y_1 = 1)$

results to populations where mortality due to diarrhoea or measles is not predominant.

More generally, while one may summarize the main "identification assumption" for the results in this paper in terms of the counterfactual independence $Y_1 \perp\!\!\!\perp S \mid Y_0$, note we did not commence the analysis by imposing this or any "identification assumption." Instead, we made an effort to explicate our understanding of the problem directly in a structural model, and the necessary counterfactual independence emerged naturally as a *logical consequence of the structure*. This is an important part of the process. If some of those modeling assumptions happen to be challenged, as they often are in practical settings (e.g, unobserved confounding between H and B), we should refrain from positing that $Y_1 \perp\!\!\!\perp S \mid Y_0$ and the model both warns us of possible threats, as well as helps us in finding alternative solutions.[14]

Conclusions

This paper showed how two apparently separate areas of causal inference research—the generalization of causal effects across populations [1, 10, 20] and the identification of "causes of effects" [16, 18, 19, 26]—can be merged for mutual benefit, unveiling important results in both areas.

The first lesson that emerges from this combined analysis is that certain functional constraints may entail the invariance of probabilities of causation across domains, which can then be used as instruments to license generalization. This may occur when the outcome is a product of several independent processes, only some of which are carriers of disparities, and when the outcome produced under the "no-treatment" condition is sufficient to block these sources of disparity. These functional constraints may enable the identification, or at least the bounding of the target effect in settings where non-parametric generalization is otherwise impossible.

A second lesson that surfaces from our investigation is that, whenever experimental data from multiple sites are available, these may lead to the point identification of probabilities of causation. These counterfactual probabilities can be the targets of investigations in public health, legal settings, and the production of explanations [13, 18, 19]. For example, drugs with a positive average treatment effect may still kill individuals who would have otherwise

[14] For example, a sensitivity analysis might still be possible, and one could investigate how big a departure from the original model assumptions would be necessary to invalidate the main conclusions. See, e.g., Cinelli et al. [3], Cinelli and Hazlett [2].

survived—being able to quantify the percentage of individuals that are saved or harmed by the treatment has important implications in many public health applications.

The development of tools for automating the types of analyses presented here, paralleling those available for non-parametric models, is a challenging topic for future work. As we have seen, determining the invariance of probabilities of causation requires additional constraints beyond the standard non-parametric model; some recent developments, such as algorithms for handling context-specific independencies for causal identification [27], may provide the initial steps towards this undertaking.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

An example with continuous treatment

Here we provide a simple example in which, although the treatment variable is continuous, the relevant dependencies among potential outcomes are still amenable to graphical representation. Suppose we have the same selection diagram as in Fig. 2b, but now let X, B, and H all be continuous variables. Next, consider the following functional specification for the structural equation of Y,

$$Y = I(H > 0) \vee I(X \times B > 0) \tag{4}$$

where $I(\cdot)$ denotes the indicator function. Now note from Equation 4 we can derive the potential outcomes $Y_0 = I(H > 0)$ for $x = 0$, and, $Y_x = I(H > 0) \vee I(xB > 0) = Y_0 \vee I(xB > 0)$, for $x \neq 0$. We can thus draw the same modified selection diagram as in Fig. 3, but now replacing Y_1 with Y_x, leading to the conclusion that $Y_x \perp S \mid Y_0$, for all $x \neq 0$.

Proofs

Bounds with a single source population

Here we show how to obtain the bounds of Theorem 1. To simplify notation, let $P_{ij} := P(Y_i = j)$, $P^*_{ij} := P^*(Y_i = j)$, $\text{PS}_{10} := P^*(Y_1 = 0|Y_0 = 1) = P(Y_1 = 0|Y_0 = 1)$ and $\text{PS}_{01} = P^*(Y_1 = 1|Y_0 = 0) = P(Y_1 = 1|Y_0 = 0)$. The target function to be optimized is P^*_{11}, which can be written as,

$$P^*_{11} = (1 - \text{PS}_{10})P^*_{01} + \text{PS}_{01}(1 - P^*_{01}) \tag{5}$$

Our goal is to pick PS_{10} and PS_{01} such that it maximizes (or minimizes) Eq. 5 subject to the following constraints: (i) PS_{10} and PS_{01} need to be between zero and one (since PS_{10} and PS_{01} need to be valid probabilities); and, (ii) PS_{10} and PS_{01} must conform to the observed results of the trial in the source domain, that is, $P_{11} = (1 - \text{PS}_{10})P_{01} + \text{PS}_{01}(1 - P_{01})$. Thus, our optimization problem is,

$$\begin{aligned} \max_{\text{PS}_{10},\text{PS}_{01}} \; & P^*_{11} = (1 - \text{PS}_{10})P^*_{01} + \text{PS}_{01}(1 - P^*_{01}) \\ \text{s.t.} \quad & P_{11} = (1 - \text{PS}_{10})P_{01} + \text{PS}_{01}(1 - P_{01}) \\ \text{and} \quad & 0 \leq \text{PS}_{10} \leq 1,\ 0 \leq \text{PS}_{01} \leq 1 \end{aligned}$$

To simplify the problem, we can use the equality constraint $P_{11} = (1 - \text{PS}_{10})P_{01} + \text{PS}_{01}(1 - P_{01})$ to eliminate one of the variables. For instance, writing PS_{10} in terms of PS_{01} gives us,

$$1 - \text{PS}_{10} = \frac{P_{11} - \text{PS}_{01}(1 - P_{01})}{P_{01}} \tag{6}$$

Which results in a new target function,

$$P^*_{11} = (1 - \text{PS}_{10})P^*_{01} + \text{PS}_{01}(1 - P^*_{01}) \tag{7}$$

$$= \left(\frac{P_{11} - \text{PS}_{01}(1 - P_{01})}{P_{01}}\right)P^*_{01} + \text{PS}_{01}(1 - P^*_{01}) \tag{8}$$

$$= \left(\frac{P_{11}}{P_{01}}\right)P^*_{01} + \left(\frac{P_{01} - P^*_{01}}{P_{01}}\right)\text{PS}_{01} \tag{9}$$

$$= RR \times P^*_{01} + \left(\frac{P_{01} - P^*_{01}}{P_{01}}\right)\text{PS}_{01} \tag{10}$$

where $RR = \frac{P_{11}}{P_{01}}$ is the causal *risk-ratio* in the trial of the source domain Π. Since $0 \leq (1 - \text{PS}_{10}) \leq 1$, the substitution also results in additional constraints on PS_{01},

$$\frac{P_{11} - P_{01}}{1 - P_{01}} \leq \text{PS}_{01} \leq \frac{P_{11}}{1 - P_{01}} \tag{11}$$

Thus, define the lower and upper bounds on PS_{01} as

$$\text{PS}^L_{01} = \max\left\{0, \frac{P_{11} - P_{01}}{1 - P_{01}}\right\}, \qquad \text{PS}^U_{01} = \min\left\{\frac{P_{11}}{1 - P_{01}}, 1\right\}$$

Our new maximization problem can be written as,

$$\max_{\text{PS}_{01}} \; RR \times P^*_{01} + \left(\frac{P_{01} - P^*_{01}}{P_{01}}\right)\text{PS}_{01} \qquad \text{s.t.} \quad \text{PS}^L_{01} \leq \text{PS}_{01} \leq \text{PS}^U_{01} \tag{12}$$

Since the target function is linear, the maximum occurs at the extreme points of PS_{01}. The same reasoning holds for the minimization problem. Thus, we have that,

$$P^{*L}_{11} \leq P^{*}_{11} \leq P^{*U}_{11}$$

where

$$P^{*L}_{11} = RR \times P^{*}_{01} + \min\left\{\left(\frac{P_{01} - P^{*}_{01}}{P_{01}}\right)\mathrm{PS}^{L}_{01}, \left(\frac{P_{01} - P^{*}_{01}}{P_{01}}\right)\mathrm{PS}^{U}_{01}\right\}$$

and

$$P^{*U}_{11} = RR \times P^{*}_{01} + \max\left\{\left(\frac{P_{01} - P^{*}_{01}}{P_{01}}\right)\mathrm{PS}^{L}_{01}, \left(\frac{P_{01} - P^{*}_{01}}{P_{01}}\right)\mathrm{PS}^{U}_{01}\right\}$$

Informativeness of the bounds

We now derive the width of the bounds for P^{*}_{11} for the case when the bounds for PS_{01} do not reach 0 nor 1 (this will happen when both $P_{11} > P_{01}$ and $P_{11} < 1 - P_{01}$). Define the width W of the bounds as the difference between the upper and lower bound of P^{*}_{11}, that is,

$$W = P^{*U}_{11} - P^{*L}_{11}$$

Expanding the terms we obtain,

$$W = P^{*U}_{11} - P^{*L}_{11} \tag{13}$$

$$= \left|\left(\frac{P_{01} - P^{*}_{01}}{P_{01}}\right)\mathrm{PS}^{U}_{01} - \left(\frac{P_{01} - P^{*}_{01}}{P_{01}}\right)\mathrm{PS}^{L}_{01}\right| \tag{14}$$

$$= \frac{|P_{01} - P^{*}_{01}|}{P_{01}} \times \left(PS^{U}_{01} - \mathrm{PS}^{L}_{01}\right) \tag{15}$$

$$= \frac{|P_{01} - P^{*}_{01}|}{P_{01}} \times \frac{P_{01}}{1 - P_{01}} \tag{16}$$

$$= \frac{|P_{01} - P^{*}_{01}|}{1 - P_{01}} \tag{17}$$

Thus, when the bounds for PS_{01} are "interior," the informativeness of the bounds depend only on P_{01} and P^{*}_{01}. Moreover, even if the bounds for PS_{01} are "wide," the bounds for P^{*}_{11} may be "narrow," provided the baseline risks of the source and target population are close enough.

Identification with multiple source domains

We now show how to obtain the identification results of Theorems 2 and 3. Consider two source populations Π^{a} and Π^{b}. Again, to simplify notation, let $P^{a}_{ij} := P^{a}(Y_i = j)$, $P^{b}_{ij} := P^{a}(Y_i = j)$, $\mathrm{PS}_{10} := P^{a}(Y_1 = 0|Y_0 = 1) = P^{b}(Y_1 = 0|Y_0 = 1) = P^{*}(Y_1 = 0|Y_0 = 1)$ and $\mathrm{PS}_{01} := P^{a}(Y_1 = 1|Y_0 = 0) = P^{b}(Y_1 = 1|Y_0 = 0) = P^{*}(Y_1 = 1|Y_0 = 0)$.

First note that PS_{10} and PS_{01} are identified from the experimental data in Π^{a} and Π^{b}. Using the law of total probability for P^{a}_{11} and P^{b}_{11} write,

$$P^{a}_{11} = (1 - \mathrm{PS}_{10}) \times P^{a}_{01} + \mathrm{PS}_{01} \times P^{a}_{00} \tag{18}$$

$$P^{b}_{11} = (1 - \mathrm{PS}_{10}) \times P^{b}_{01} + \mathrm{PS}_{01} \times P^{b}_{00} \tag{19}$$

We thus have a system of two equations and two unknowns,

$$\begin{bmatrix} P^{a}_{01} & P^{a}_{00} \\ P^{b}_{01} & P^{b}_{00} \end{bmatrix} \begin{bmatrix} (1 - \mathrm{PS}_{10}) \\ \mathrm{PS}_{01} \end{bmatrix} = \begin{bmatrix} P^{a}_{11} \\ P^{b}_{11} \end{bmatrix} \tag{20}$$

Yielding the solution,

$$\begin{bmatrix} (1 - \mathrm{PS}_{10}) \\ \mathrm{PS}_{01} \end{bmatrix} = \frac{1}{P^{a}_{01}P^{b}_{00} - P^{b}_{01}P^{a}_{00}} \times \begin{bmatrix} P^{b}_{00} & -P^{a}_{00} \\ -P^{b}_{01} & P^{a}_{01} \end{bmatrix} \begin{bmatrix} P^{a}_{11} \\ P^{b}_{11} \end{bmatrix} \tag{21}$$

Which amounts to:

$$\mathrm{PS}_{10} = 1 - \frac{P^{a}_{11}P^{b}_{00} - P^{b}_{11}P^{a}_{00}}{P^{a}_{01}P^{b}_{00} - P^{b}_{01}P^{a}_{00}} \tag{22}$$

$$\mathrm{PS}_{01} = \frac{P^{b}_{11}P^{a}_{01} - P^{a}_{11}P^{b}_{01}}{P^{a}_{01}P^{b}_{00} - P^{b}_{01}P^{a}_{00}} \tag{23}$$

All values of the RHS can be computed from the experimental data of Π^{a} and Π^{b}. Note that, since PS_{10} and PS_{01} must be between 0 and 1, not all solutions are valid. Therefore, two domains already entail some testable implications—if either PS_{10} and PS_{01} are not valid probabilities, this means that the assumption that the probabilities of sufficiency are invariant across domains is false. If we add a third or more source domains, it is easy to see that we will have three or more equations but still only two unknowns, and the system is thus over-identified.

Once in possession of PS_{10} and PS_{01}, we can transport the causal effect to the target population Π^{*} by appealing again to the law of total probability,

$$P^{*}_{11} = (1 - \mathrm{PS}_{10}) \times P^{*}_{01} + \mathrm{PS}_{01} \times P^{*}_{00} \tag{24}$$

$$= \frac{P^{a}_{11}P^{b}_{00} - P^{b}_{11}P^{a}_{00}}{P^{a}_{01}P^{b}_{00} - P^{b}_{01}P^{a}_{00}} \times P^{*}_{01} + \frac{P^{b}_{11}P^{a}_{01} - P^{a}_{11}P^{b}_{01}}{P^{a}_{01}P^{b}_{00} - P^{b}_{01}P^{a}_{00}} \times P^{*}_{00} \tag{25}$$

Finally, we note that all probabilities of causation, as discussed in [26], are also identifiable in this setting. First, consider the *probability of necessity and sufficiency*, $\mathrm{PNS} = P(Y_1 = 1, Y_0 = 0)$ for Π^{a}. Using the chain rule, PNS can be written as,

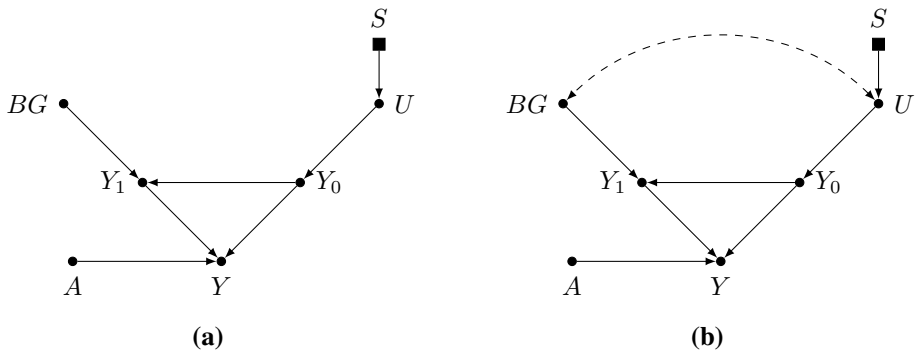


Fig. 9 Two selection diagrams compatible with the verbal description of Huitfeldt et al. [10, page 11]. Yet, model **a** implies $Y_1 \perp S \mid Y_0$, and model **b** implies the opposite; conditioning on Y_0 opens the colliding path $S \to U \leftrightarrow BG \to Y_1$

$$P^a(Y_1 = 1, Y_0 = 0) = P^a(Y_1 = 1 \mid Y_0 = 0)P^a(Y_0 = 0) \quad (26)$$

$$= \text{PS}_{01} \times P^a(Y_0 = 0) \quad (27)$$

Note PS_{01} was already identified, and $P^a(Y_0 = 0)$ is given by the trial data in Π^a, thus rendering PNS^a identifiable. Similar reasoning holds for Π^b.

For the *probability of necessity*, define $\text{PN}_{01} := P(Y_0 = 0 \mid Y_1 = 1)$. Due to the randomization of X, PN_{01} coincides with Tian and Pear's probability of necessity *during the trial* (not the observational PN), by the same argument we provide for PS in the main text. The final step is to note that,

$$P^a(Y_0 = 0 \mid Y_1 = 1) = \frac{P^a(Y_0 = 0, Y_1 = 1)}{P^a(Y_1 = 1)} = \frac{\text{PNS}^a}{P^a(Y_1 = 1)}$$

The numerator is simply the PNS, which we have already identified, and the denominator is given by the trial data in Π^a. Again, analogous argument can be given for Π^b.

Modeling functional constraints

To illustrate the usefulness of explicitly modeling functional constraints in a structural framework, we apply the same modeling strategy of the paper in an example described in Huitfeldt et al. [10, p. 11]:

> Consider a team of investigators who are interested in the effect of antibiotic treatment on mortality in patients with a specific bacterial infection (...) the investigators believe that the response to this antibiotic is completely determined by an unmeasured bacterial gene, such that only those who are infected with a bacterial strain with this gene respond to treatment. The prevalence of this bacterial gene is equal between populations, because the populations share the same bacterial ecosystem (...) if the investigators further believe that the gene for susceptibility reduces the mortality in the presence of antibiotics, but has no effect in the absence of antibiotics, they will conclude that G may be equal between populations.

Here the conclusion that G may be equal between populations is equivalent to claiming $Y_1 \perp S \mid Y_0$. But is the description above sufficient for substantiating this claim? Figure 9 shows two models compatible with the description, yet leading to two opposite conclusions.

Let the variable A represent the binary treatment (antibiotic), Y represent the binary outcome (mortality), BG stand for the presence or absence of the "bacterial gene" and finally let U be a binary variable that summarizes all other factors that may cause death ($Y = 1$). The description of the problem suggests the functional specification,

$$Y = U \wedge (\neg A \vee \neg BG) \quad (28)$$

showing the antibiotics and the bacterial gene both helping to *reduce* mortality ($\neg$ denotes the logical "not"). Equation 28 entails the potential outcomes $Y_0 = U$ and $Y_1 = U \wedge (\neg BG) = Y_0 \wedge (\neg BG)$, which are explicitly shown in both diagrams as dictated by the functional specification. Moreover, in both models the prevalence of the bacterial gene BG is equal between populations (i.e., $BG \perp S$). In the model of Fig. 9a, as in our previous analysis, we indeed conclude that $Y_1 \perp S \mid Y_0$, and that $P^*(Y_1)$ is transportable. However, in the model of Fig. 9b, there is an unmeasured confounder between BG and U.[15] Conditioning on Y_0 (a child of a collider) opens the colliding path $S \to U \leftrightarrow BG \to Y_1$, thus not licensing the independence $Y_1 \perp S \mid Y_0$.

[15] This could arise, for instance, as a result of population stratification.

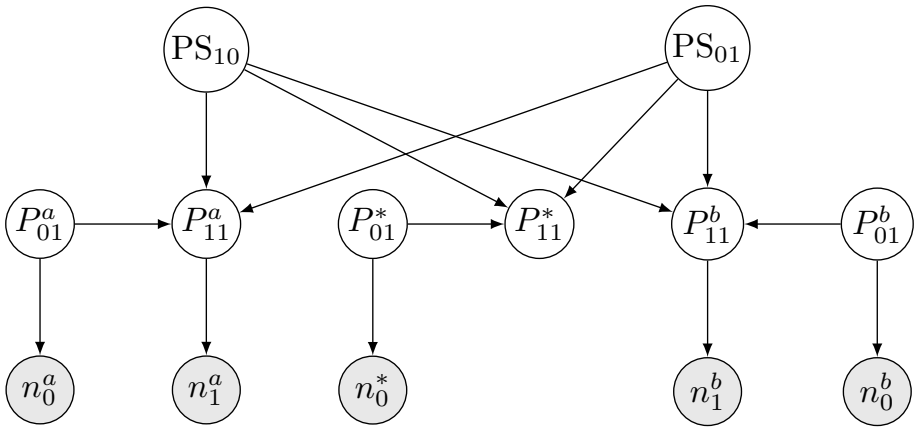


Fig. 10 Probabilistic graphical model with two source populations Π^a, Π^b and one target population Π^*. Gray nodes (n^a_0, n^a_1, n^*_0, n^b_0, n^b_1) denote observed variables. White notes denote latent parameters (P^a_{01}, P^a_{11}, PS_{10}, PS_{01}, P^*_{11}, P^*_{01}, P^b_{11}, P^b_{01}). Note that P^a_{11}, P^*_{11} and P^b_{11} share the parameters PS_{10} and PS_{01}, which are invariant across populations

Bayesian estimation

Multiple source domains

In this section we show how to extend the probabilistic graphical model of the section "A Bayesian approach to estimation" to two or more sources. Let us start with two source populations Π^a and Π^b, and one target domain Π^*. The observed data is now $\mathcal{D} = \{n^a_0, n^a_1, n^*_0, n^b_0, n^b_1\}$, all with binomial distributions:

$$n^a_0 \sim \text{Binomial}(N^a_0, P^a_{01}) \tag{29}$$

$$n^a_1 \sim \text{Binomial}(N^a_1, P^a_{11}) \tag{30}$$

$$n^*_0 \sim \text{Binomial}(N^*_0, P^*_{01}) \tag{31}$$

$$n^b_0 \sim \text{Binomial}(N^b_0, P^b_{01}) \tag{32}$$

$$n^b_1 \sim \text{Binomial}(N^b_1, P^b_{11}) \tag{33}$$

We also have the following deterministic relationships for P^a_{11}, P^b_{11} and P^*_{11}:

$$P^a_{11} = (1 - PS_{10})P^a_{01} + PS_{01}(1 - P^a_{01}) \tag{34}$$

$$P^b_{11} = (1 - PS_{10})P^b_{01} + PS_{01}(1 - P^b_{01}) \tag{35}$$

$$P^*_{11} = (1 - PS_{10})P^*_{01} + PS_{01}(1 - P^*_{01}) \tag{36}$$

The probabilistic graphical model for this case is shown in Fig. 10.

Thus, one needs to place priors on the parent nodes only, and then perform inference as before. The extension to more than two populations follows the same logic. It is worth noting that, as we have seen in the section "Building the structural model" with two or more source populations the model entails testable implications. Therefore, we advise researchers to check whether the data is compatible with the model [6].

Finally, similarly to the discussion in the section "A Bayesian approach to estimation" a simpler modeling alternative here is to place priors only on the parameters of the observed data directly, and make inferences using the posterior of the functionals of the observed data that identify the target quantities.

Replication code

R code to replicate the estimation examples using JAGS [22] and the package rjags [21] is provided in the online supplemental material.

References

1. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. Proc Natl Acad Sci. 2016;113(27):7345–52.
2. Cinelli C, Hazlett C. Making sense of sensitivity: extending omitted variable bias. J R Stat Soc Ser B (Stat Methodol). 2020;82:39–67.
3. Cinelli C, Kumor D, Chen B, Pearl J, Bareinboim E. Sensitivity analysis of linear structural causal models. In: International conference on machine learning; 2019.
4. Dahabreh IJ, Petito LC, Robertson SE, Hernán MA, Steingrimsson JA. Toward causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a new target population. Epidemiology. 2020;31(3):334–44.
5. Geiger D, Verma T, Pearl J. Identifying independence in Bayesian networks. Networks. 1990;20(5):507–34.
6. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. Boca Raton: CRC Press; 2013.
7. Gustafson P. Bayesian inference for partially identified models: exploring the limits of limited data, vol. 140. Boca Raton: CRC Press; 2015.
8. Hartman E, Grieve R, Ramsahai R, Sekhon JS. From SATE to PATT: combining experimental with observational studies to

estimate population treatment effects. J R Stat Soc Ser A (Stat Soc). 2015;10:1111.
9. Huitfeldt A. Effect heterogeneity and external validity in medicine; 2019. https://www.lesswrong.com/posts/wwbrvumMWhDfeo652/.
10. Huitfeldt A, Goldstein A, Swanson SA. The choice of effect measure for binary outcomes: introducing counterfactual outcome state transition parameters. Epidemiol Methods. 2018;7(1):20160014.
11. Huitfeldt A, Swanson SA, Stensrud MJ, Suzuki E. Effect heterogeneity and variable selection for standardizing causal effects to a target population. Eur J Epidemiol. 2019;34:1119–29.
12. Lu Y, Scharfstein DO, Brooks MM, Quach K, Kennedy EH. Causal inference for comprehensive cohort studies; 2019. arXiv:1910.03531.
13. Mueller S, Pearl J. Which patients are in greater need: a counterfactual analysis with reflections on covid-19. In: Causal analysis in theory and practice; 2020. https://ucla.in/39Ey8sU.
14. Muhilal PD, Idjradinata YR, Muherdiyantiningsih KD. Vitamin a-fortified monosodium glutamate and health, growth, and survival of children: a controlled field trial. Am J Clin Nutr. 1988;48(5):1271–6.
15. Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82(4):669–88.
16. Pearl J. Probabilities of causation: three counterfactual interpretations and their identification. Synthese. 1999;121(1–2):93–149.
17. Pearl J. Causality. Cambridge: Cambridge University Press; 2009.
18. Pearl J. Causes of effects and effects of causes. Sociol Methods Res. 2015;44(1):149–64.
19. Pearl J. Sufficient causes: on oxygen, matches, and fires. J Causal Inference. 2019;7(2):1–11.
20. Pearl J, Bareinboim E. External validity: from do-calculus to transportability across populations. Stat Sci. 2014;29(4):579–95.
21. Plummer M. rjags: Bayesian graphical models using MCMC. R package version. 2016;4(6).
22. Plummer M, et al. Jags: a program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd international workshop on distributed statistical computing, vol. 124. Vienna, Austria; 2003. p. 1–10.
23. Richardson TS, Evans RJ, Robins JM. Transparent parameterizations of models for potential outcomes. Bayesian Stat. 2011;9:569–610.
24. Silva R, Evans R. Causal inference through a witness protection program. J Mach Learn Res. 2016;17(1):1949–2001.
25. Sommer A, Djunaedi E, Loeden A, Tarwotjo I, West K, Tilden JR, Mele L, Group AS, et al. Impact of vitamin a supplementation on childhood mortality: a randomised controlled community trial. Lancet. 1986;327(8491):1169–73.
26. Tian J, Pearl J. Probabilities of causation: bounds and identification. Ann Math Artif Intell. 2000;28(1–4):287–313.
27. Tikka S, Hyttinen A, Karvanen J. Identifying causal effects via context-specific independence relations. In: Advances in neural information processing systems; 2019. p. 2800–10.
28. West KP Jr, Katz J, LeClerq SC, Pradhan E, Tielsch JM, Sommer A, Pokhrel R, Khatry S, Shrestha S, Pandey M. Efficacy of vitamin a in reducing preschool child mortality in Nepal. Lancet. 1991;338(8759):67–71.