

Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness

Sander Greenland · Mohammad Ali Mansournia

Received: 26 March 2014 / Accepted: 22 January 2015 / Published online: 17 February 2015
© Springer Science+Business Media Dordrecht 2015

Abstract We describe how ordinary interpretations of causal models and causal graphs fail to capture important distinctions among ignorable allocation mechanisms for subject selection or allocation. We illustrate these limitations in the case of random confounding and designs that prevent such confounding. In many experimental designs individual treatment allocations are dependent, and explicit population models are needed to show this dependency. In particular, certain designs impose unfaithful covariate-treatment distributions to prevent random confounding, yet ordinary causal graphs cannot discriminate between these unconfounded designs and confounded studies. Causal models for populations are better suited for displaying these phenomena than are individual-level models, because they allow representation of allocation dependencies as well as outcome dependencies across individuals. Nonetheless, even with this extension, ordinary graphical models still fail to capture distinctions between hypothetical superpopulations (sampling distributions) and observed populations (actual distributions), although potential-outcome models can be adapted to show these distinctions and their consequences.

Keywords Causal graphs · Confounding · Directed acyclic graphs · Ignorability · Inverse probability weighting · Unfaithfulness

Introduction

Potential-outcome (counterfactual) and graphical causal models are now standard tools for analysis of study designs and data. Expositions can be found in modern textbooks [1–3]; in most applications we see, however, the causal models refer to individuals within an implicit population, while the graphs are not specific about whether they refer to causation within individuals or within populations. We describe how such treatments miss important distinctions between confounded and unconfounded study designs, and thus may lead some users into erroneous procedures and conclusions, especially when random variation is important or when faithfulness assumptions may be violated. The problems regarding random confounding have long been recognized in various forms [4–12], but few sources have address the converse problems involving unfaithfulness [13].

Faithfulness assumptions formalize the idea that no perfect cancellation of effects or associations is occurring in the system or population being studied; these assumptions are often made implicitly and have been used explicitly in graphical causal modeling [14]. While they are sometimes defensible, there are natural settings in which there may be approximate unfaithfulness (that is, cancellations statistically indistinguishable from unfaithfulness), as with hazardous medical interventions that may save some patients but kill others. In these settings, the unfaithfulness is accidental or unstable in that it is not implied by the causal process generating the data; it is nonetheless real when it happens,

S. Greenland
Department of Epidemiology, UCLA School of Public Health,
University of California, Los Angeles, CA, USA

S. Greenland
Department of Statistics, UCLA College of Letters and Science,
University of California, Los Angeles, CA, USA

M. A. Mansournia (✉)
Department of Epidemiology and Biostatistics, School of Public
Health, Tehran University of Medical Sciences,
PO Box: 14155-6446, Tehran, Iran
e-mail: mansournia_ma@yahoo.com

and its possibility severely limits statistical procedures for causal inference [13].

There are also study designs that produce exact unfaithfulness, such as balanced matched-cohort studies, in which selection paths are used to cancel confounding paths [15]; in these settings unfaithfulness is a consequence of the data-generating process and thus a stable feature of the system. Thus, unlike some sources [2], we make a sharp distinction between faithfulness and stability: stable properties are taken to be those deducible from postulated laws or causal mechanisms, and these structural elements may induce unfaithful exact independencies [16].

To illustrate our points, we discuss how stratified (blocked) randomized-trial designs induce unfaithfulness as a cancellation of path-specific associations, and how this cancellation is not visible in an ordinary causal graph. Such examples show how ordinary causal graphs cannot always discriminate between unconfounded and confounded studies, and provide very incomplete information about the proper statistical analysis to employ. Full population modeling addresses some of these problems, and appears better suited for causal modeling in population sciences like epidemiology. Ordinary graphs can be used for this population modeling simply by reinterpreting their component variables (nodes) as population distributions, but this reinterpretation still does not allow them to exhibit design features crucial to proper statistical analysis.

Our development will proceed in detail within a simple model in which a causal mean difference is the target and the only source of statistical variation for analysis is treatment randomization, as in models for exact permutation tests of effects [17–20]. We first lay out our notation, then use that to review the relation of randomization to confounding in the potential-outcome framework, emphasizing properties that are not visible in ordinary causal graphs. We then expand the framework to clarify the properties. For a much more detailed yet general technical account of the relation of faithfulness to confounding and causal inference, see Robins et al. [13]; and for extensive accounts of the relation of graphs to causal models, including new tools to address limitations of ordinary causal graphs, see Richardson and Robins [21, 22].

Basic notation and definitions

We assume the reader is familiar with potential-outcome (counterfactual) models and causal directed acyclic graphs (cDAGs) [1–3, 23–25]. Figure 1 provides examples. Especially important is that two variables in a graph are *connected* if there is an open path between them (d-connected), otherwise they are *disconnected* or *separated*

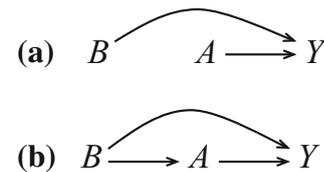


Fig. 1 Individual-level causal diagrams representing (a) simple randomization and (b) B -stratified randomization

(d-separated). A DAG model asserts that disconnected variables are independent, such as A and B in Fig. 1a. Nonetheless, a DAG model does *not* assert that connected variables are dependent (associated), a fact that we have seen overlooked in some discussions. These discussions assume that connected variables are always associated, which is called a *faithfulness* assumption. Faithfulness is not one of the basic DAG assumptions used to determine whether an effect is identified (estimable) given confounding and selection bias, although it does have important statistical implications [13], some of which we will describe.

Suppose we observe a target cohort composed of n individuals indexed by $i = 1, \dots, n$, each of whom will be allocated by some mechanism to receive one of two study treatments as designated by the individual indicators $A_i = 1$ or 0 . The full-cohort allocation is the list (A_1, \dots, A_n) . “Unconditional” will refer to the pre-allocation state in which the A_i values are still undetermined and thus considers all possible lists of values (cohort allocations) allowed by the allocation protocol. “Conditional” will refer to the post-allocation state in which all A_i values have been assigned, and thus we know those values and their relation to other variables observed in the cohort. We will be concerned with accounting for unconditional dependencies among the A_i (“treatment interference” between subjects) and conditional independencies between the A_i and other variables, each of which may be produced by common experimental designs.

We assume that each individual has a pair of potential outcomes: Y_{1i} which is observed when $A_i = 1$, and Y_{0i} which is observed when $A_i = 0$; these are baseline (pre-treatment) variables, and treatment determines which one we observe [26]. The individual effect of having $A_i = 1$ as opposed to $A_i = 0$ is $Y_{1i} - Y_{0i}$, which cannot be observed, but the average of these differences can be estimated. As usual for noncontagious-disease modeling, we will assume both Y_{1i} and Y_{0i} are independent across individuals (a “no-interference” assumption on the outcomes).

We define $Y_{Ai} = A_i Y_{1i} + (1 - A_i) Y_{0i}$, so that $Y_{Ai} = Y_{1i}$ among those with $A_i = 1$ and $Y_{Ai} = Y_{0i}$ among those with $A_i = 0$. Our examples will involve no censoring, nonadherence, measurement error, or other methodologic problems, so that Y_{Ai} will equal the patient’s observed outcome

(the consistency condition). Because the pair (Y_{1i}, Y_{0i}) is fixed at baseline, the only random variation in Y_{Ai} is from variation in treatment A_i ; once A_i is determined, so is Y_{Ai} . This model is deterministic but may be extended to stochastic outcomes by replacing (Y_{1i}, Y_{0i}) with a pair of potential parameters $(\theta_{1i}, \theta_{0i})$ for an individual outcome distribution [5].

We will drop the subscript i when describing an unspecified individual, and all summations and averages will be over all individuals in the cohort. We assume for now that the target population of inference is the total cohort, and that the causal effect of interest is the difference Δ in the population-average outcome $\text{Av}(Y_1) = \sum Y_{1i}/n$ with everyone given experimental treatment and the average $\text{Av}(Y_0) = \sum Y_{0i}/n$ with everyone given control treatment:

$$\Delta = \text{Av}(Y_1) - \text{Av}(Y_0);$$

Δ equals the average effect on population members, $\text{Av}(Y_1 - Y_0) = \sum(Y_{1i} - Y_{0i})/n$. The numbers in each treatment group are $n_1 = \sum A_i$ and $n_0 = \sum(1 - A_i) = n - n_1$. Although n , n_1 and n_0 are usually treated in exposure modeling (propensity scoring) as if they are completely random, more often they are wholly or partially fixed by design.

The treatment-specific averages can be written as total-cohort averages of the potential outcomes, each weighted by their corresponding received-treatment indicator:

$$\begin{aligned} \text{Av}(Y_A|A = 1) &= \text{Av}(Y_1|A = 1) = \sum A_i Y_{1i}/n_1 \text{ and} \\ \text{Av}(Y_A|A = 0) &= \text{Av}(Y_0|A = 0) = \sum (1 - A_i) Y_{0i}/n_0. \end{aligned}$$

We will define D as the observed difference,

$$\begin{aligned} D &= \text{Av}(Y_A|A = 1) - \text{Av}(Y_A|A = 0) \\ &= \text{Av}(Y_1|A = 1) - \text{Av}(Y_0|A = 0). \end{aligned}$$

When Y_1 and Y_0 are indicators, Δ and D are the total-cohort and unadjusted (“crude”) observed risk differences, respectively. B will represent a set of baseline (pre-treatment) predictors of the potential-outcome pair (Y_1, Y_0) .

Independence and unfaithfulness

A probability distribution is said to be *compatible* with a DAG if all variables that are dependent in the distribution are connected in the DAG. An analysis that assumes a particular DAG will be logically coherent only if it uses a distribution that is compatible with that DAG. A compatible distribution may however be *unfaithful* to the DAG if two variables independent in the distribution are connected in the DAG. Disconnections in an unfaithful DAG do not reveal all the independencies in the distribution. Some DAG expositions assume faithfulness from the start [14];

nonetheless, as we will discuss, this is not always a reasonable restriction. Use of an unfaithful distribution is logically coherent; if however unfaithfulness is allowed (i.e., faithfulness is not assumed), deductions from the DAG are strictly one-way: Absence of connection implies independence, but connection does *not* imply dependence (association). In particular, in a causal graph without a faithfulness assumption, a direct connection does not imply that a given effect measure such as Δ will be non-null.

Unfaithfulness can be depicted in causal models other than DAGs. We say *causal* unfaithfulness occurs when two variables are unconditionally independent despite one having an effect on the other [4]. A well-known example is confounding in which there is no association of A and Y because the association due to the effect of A on Y is cancelled by the confounding. Such cancellation between causal pathways requires an association between A and other causes of Y , and thus is precluded by Fig. 1a but allowed by (compatible with) Fig. 1b.

A limitation of DAGs is that unfaithfulness can occur even if there is no cancellation potential visible in the graph, as in Fig. 1a. As an example, suppose the outcome is binary, so that Y_1, Y_0 and hence Y_A equal either 1 or 0. Suppose also that $A = 1$ causes $Y = 1$ (equivalently, $A = 0$ prevents $Y = 1$) for some nonzero proportion p of the pairs, so that $Y_1 = 1$ and $Y_0 = 0$ for these pairs, and that $A = 1$ prevents $Y = 1$ (equivalently, $A = 0$ causes $Y = 1$) for an equal proportion p of pairs, so that $Y_1 = 0$ and $Y_0 = 1$ for these pairs; thus A profoundly affects the outcome of all individuals in either type of pair. Nonetheless, A and Y will appear unassociated (satisfy the weak null hypothesis of $\Delta = 0$) because the causal and preventive effects of A will cancel out of the average (counterbalance one another).

The unfaithful cancellation in this example will also be statistically indistinguishable from the sharp (strong) null hypothesis of no effect of treatment on any pair ($Y_1 = Y_0$ for all individuals), even if treatment is randomized [4, 19]; that is, the strong null hypothesis of “no effect on any individual” is not identifiable without some assumption that precludes perfect cancellation, which in graphical terms would correspond to a faithfulness assumption. To appreciate the extent this nonidentifiability, note that Δ may be 0 even if treatment changes *everyone’s* outcome: If $p = 1/2$, half the cohort will have $Y_1 = 1, Y_0 = 0$ while the other half will have $Y_1 = 0, Y_0 = 1$.

Of course, background information may lead us judge perfect cancellation as implausible. One may even judge that there is no cancellation at all, as encoded in a strong monotonicity assumption that $A = 1$ is never preventive ($Y_1 \geq Y_0$ for everyone), or that $A = 1$ is never causal ($Y_1 \leq Y_0$ for everyone). Nonetheless, without further assumptions or information, basic statistical procedures

cannot distinguish between any degree of perfect cancellation ($0 < p \leq 1/2$) and $Y_1 = Y_0$ for everyone (the strong null, for which $p = 0$), which is to say that failure to reject the strong null will also entail failure to reject perfect cancellation [19].

Ignorability and unconfoundedness

Independence of treatment A and the potential-outcome pair (Y_1, Y_0) is often called strong ignorability, complete (or full) exchangeability, or unconfoundedness in the pre-allocation distribution or treatment assignment, in which A is random [3, 26, 27]. Figure 1a illustrates this independence via the absence of any nondirected open path from A to Y . An immediate consequence of this condition is that *on average* (in expectation) over all possible allocations, the observed outcome difference between the groups equals the target effect:

$$\begin{aligned} E[D] &= E[\text{Av}(Y_1|A=1) - \text{Av}(Y_0|A=0)] \\ &= E[\text{Av}(Y_1|A=1)] - E[\text{Av}(Y_0|A=0)] \\ &= \text{Av}(Y_1) - \text{Av}(Y_0) = \Delta. \end{aligned} \quad (1)$$

This equation says that the observed mean difference D is an unbiased estimator of the target Δ , where $E[\]$ is the average over all possible treatment allocations for the cohort (i.e., $E[\]$ is the expectation taken over the unconditional treatment distribution for the cohort).

Unconditional unbiasedness ($E[D] = \Delta$) also follows from weaker conditions such as weak (or marginal) ignorability, in which A is only independent of Y_1 and Y_0 separately (marginally):

$$Y_1 \perp\!\!\!\perp A \text{ and } Y_0 \perp\!\!\!\perp A \quad (2)$$

In fact, $E[D] = \Delta$ follows from marginal mean (regression) independence of A from Y_1 and Y_0 :

$$E[\text{Av}(Y_1|A=1)] = E[\text{Av}(Y_1|A=0)] \text{ and} \quad (3a)$$

$$E[\text{Av}(Y_0|A=1)] = E[\text{Av}(Y_0|A=0)], \quad (3b)$$

which is even weaker than weak ignorability (1) if Y is not binary, and which does *not* require Fig. 1a under the usual graphical interpretations. That is, Eq. (3) and thus $E[D] = \Delta$ (unbiasedness) may hold even if B affects A , making the treatment distribution incompatible with Fig. 1a.

The sufficiency of Eq. (3a, 3b) for unbiased estimation of Δ illustrates a further limitation of common causal models: Neither Fig. 1a nor the usual ignorability (exchangeability, unconfoundedness) assumption used in much of the causal-modeling literature are necessary for unbiased estimation of Δ (although they are sufficient) [5].

In response, no-confounding can be defined in terms of unbiasedness for a particular effect measure rather than as a general property of the variable distributions [5], which can lead to parsimonious adjustments for confounding control. For example, a consequence of (3) is that if Y_1 and Y_0 are linear functions of B and the $A = 1$ and $A = 0$ subcohorts have the same means for B , then the unadjusted estimator D will be unbiased for Δ conditional on the observed allocation as well as unconditionally; thus mean matching of treated and untreated cohorts on B will be sufficient for control of confounding of D by B under a linear structural model. Closer matching or analytic adjustment for B will then only be necessary to improve precision or account for possible nonlinearities in the relation of B to Y_1 and Y_0 . This fact is a special case of more general robustness properties of matching on covariate scores (such as propensity scores) [28–30] and provides one rationale for the common practice of checking the success of matching by comparing covariate means (rather than entire distributions) across treatment groups [31]. The catch is that matching will usually alter the distribution of B in the total cohort, thus altering the distribution of $Y_1 - Y_0$ and hence its average Δ if (as one should expect) B includes modifiers of the differences $Y_1 - Y_0$ [32].

Confounding in randomized trials

Randomized trials are often treated as the “gold standard” for causal inference, because *on average* randomization balances covariates between treatment groups, even if those covariates are unobserved, and provides a known treatment distribution for statistical procedures. In causal-modeling terms, randomization enforces strong ignorability for all outcomes by cutting off effects of covariates on A , leading to Fig. 1a and unbiasedness of D for Δ ($E[D] = \Delta$) [2, 5]. Nonetheless, this unbiasedness is only unconditional, an average property over allocations. In contrast, “confounding” is often described informally as a distortion of estimates arising from baseline risk factors that are associated with but not affected by treatment, where the association refers to an imbalance of risk factors across the actual treatment groups [4–6]. Particular allocations produced by simple randomization usually exhibit some imbalance, leading to $D \neq \Delta$. Standard teaching is that if such a risk-factor imbalance is observed, it is incumbent upon the analyst to adjust for it [4, 5, 7, 8, 33–35].

The discrepancy $D - \Delta$ produced by random covariate imbalance has been called random (chance) confounding [4, 18], accidental bias [36–38], and allocation bias [9]. In the deterministic causal models underlying classical exact permutation tests, variation in A is the only source of

variation in D , and random imbalance is the *only* source of discrepancy between D and Δ [10, 17–20]. This imbalance is not visible in Fig. 1a, however; hence the figure gives the impression that covariate adjustment is unnecessary (using the back-door criterion for unbiased effect estimation [2]). This misimpression arises because the figure implies Eq. (1), which says that D is unconditionally unbiased ($E[D] = \Delta$), i.e., the average of D is over all possible treatment allocations is Δ .

If B is unmeasured, uncertainty about the discrepancy $D - \Delta$ is accounted for by standard statistical procedures. If however B is measured, our uncertainty about Δ can and should be reduced by using it in our estimator. In particular, suppose we see that B is imbalanced between treatment groups (i.e., A and B are associated in the observed data). To address this imbalance, we narrow our frequency calculations to allocations that yield only the observed association of A and B (which is to say we now condition our frequency calculations on the observed joint distribution of A and B , so we are more conditional than before). If A and B are associated (dependent) in this observed distribution, the unadjusted estimate D usually will be biased ($E[D] \neq \Delta$) over these new frequency calculations based on the observed AB distribution [4, 5, 11, 34]. This is called B -conditional bias in D , and is the random confounding due to B [4, 5, 11, 34]; it can be removed by any of the usual stratification or modeling methods to adjust for B , subject to any additional assumptions of those methods.

Under those methods and their assumptions, the resulting B -adjusted estimate D_B of Δ will (like D) be unconditionally unbiased under simple randomization, but (unlike D) will also be conditionally unbiased, that is, unbiased given the observed AB distribution (when frequency calculations are narrowed to conform to that distribution). Note carefully the difference between a B -conditional frequency evaluation and a B -conditional estimator D_B : The performance of any estimator including D as well as D_B can be evaluated both unconditionally and also under B -conditional frequencies which take the observed AB -distribution as unvarying. Nonetheless, an estimator like D that does not use B -information in its formula is not B -conditional.

Using classical linear regression analysis of a simple randomized trial under the additive individual-effect model $E(Y_a|B = b) = \alpha + \beta a + \gamma b$,

$\Delta = \beta$ and D_B is the ordinary least-squares estimate of β . If $\gamma \neq 0$, the conditional unbiasedness of D_B leads to a reduced variance relative to D in the unconditional calculations (in which the joint distribution of A and B is allowed to vary): D_B will have lower unconditional variance than D , since in this setting the random confounding of D by B represents a component of variance in D removed from

D_B by adjustment [34]. Under our simple causal model, the only component of variance remaining in D_B will be the random confounding due to baseline factors that predict Y_1 or Y_0 given B . With no such additional predictors and a deterministic effect model, $D_B = \Delta$, so D_B will have zero variance as well as zero bias. This ideal can be approached very closely in many basic physics experiments, but not in epidemiologic research.

Figure 1a and Eq. (1) do not account for random confounding and thus miss the effects of adjusting for B . This failing is in part because they portray only the ignorability constraint on the pre-allocation (unconditional) joint distribution of the baseline variables B , Y_1 , and Y_0 , encoded as independence of A from any covariate affecting Y (including B). Random confounding does not however arise from the failure of ignorability of the assignment mechanism, but is instead a conditional error arising from random departure of the observed allocation (the final data distribution of A , Y_1 , and Y_0) from the average allocation (the expected distribution of A , Y_1 , and Y_0).

Sample exchangeability versus randomization and ignorability

The blindness of DAGs and ignorability conditions to random imbalance and confounding is addressed by traditional recommendations to adjust for baseline prognostic factors to improve precision [7, 12], and to use estimated treatment probabilities even if the true treatment probabilities are known [26]. Modifications of causal graphs and ignorability conditions to encompass random confounding could be helpful for grounding these recommendations in causal models, and for preventing confusion between pre-allocation (unconditional) and post-allocation (conditional) properties that are often give the same label, such as “no confounding” or “exchangeability.”

One simple graphical device to represent a post-allocation association of A and B is to connect them by a dashed line, paralleling graphical moralization procedures for DAGs in which conditioning on a variable leads to connecting its parents [23, 39]. Other possibilities may arise using single-world intervention graphs, in which a particular allocation a for A appears in the graph, and its corresponding potential outcome Y_a replaces the composite observable outcome Y_A [21]. We will not pursue these graphical extensions, but will instead describe how to extend the underlying potential-outcome notation to exhibit random confounding and the impact of adjustment, as described in the previous section.

A condition that implies no confounding, random or otherwise, is no association of potential outcomes with the *observed* treatment allocation [4, 5] (as opposed to “no

association *expected* before treatment is allocated,” which is equivalent to ignorability). To be precise, define $\Pr(Y_0 = y|A = 1)$ as the realized (actual) proportion having $Y_0 = y$ in the subcohort with $A = 1$, and so on for other combinations of potential outcomes and treatment. Then the subcohorts with $A = 1$ and $A = 0$ are *completely exchangeable* for estimating the effect of the observed allocation on Y if for every y

$$\Pr(Y_0 = y|A = 1) = \Pr(Y_0 = y|A = 0) \text{ and} \tag{4a}$$

$$\Pr(Y_1 = y|A = 0) = \Pr(Y_1 = y|A = 1); \tag{4b}$$

they are partially exchangeable if only one of these equalities holds for every y [4, 5]. Equation (4a) says that the distribution of Y_0 in the $A = 0$ cohort can be exchanged (substituted) for the distribution of Y_0 in the $A = 1$ cohort. In parallel, (4b) says that the distribution of Y_1 in the $A = 1$ cohort can be exchanged for the distribution of Y_1 in the $A = 0$ cohort.

The problem with (4a) and (4b) is that we would never expect them to hold exactly in practice: Once allocation is completed, actual treatment will usually be associated with potential outcomes in the cohort (as revealed by associations of A with baseline risk factors), violating (4a) and (4b) and producing the allocation-conditional bias we call random confounding [4, 6, 33, 34]. If the value y of Y_0 is unique for some individual, (4a) cannot hold for that y since one side will be zero and the other side positive; in parallel, if the value y of Y_1 is unique for some individual, (4b) cannot hold for that y .

Randomization does not enforce (4a) and (4b) but instead only forces the weaker conditions of equality of the proportions expected over all possible treatment allocations,

$$E[\Pr(Y_0 = y|A = 1)] = E[\Pr(Y_0 = y|A = 0)] \tag{5a}$$

$$E[\Pr(Y_1 = y|A = 0)] = E[\Pr(Y_1 = y|A = 1)] \tag{5b}$$

Conditions (5a and 5b) are together equivalent to weak ignorability and thus imply there will be no bias in the unadjusted effect estimator D when the latter is averaged over all possible allocations (no unconditional bias). Nonetheless, they do not prohibit confounding in the observed allocation (random confounding) from violations of (4a and 4b), as when the risk factor B is associated with A in this allocation despite being disconnected from A in the causal graph (Fig. 1). Again, this random confounding by B is a bias in D seen when averaging over the more restricted allocations that reproduce the observed AB distribution (i.e., conditional averaging).

Despite this limitation, randomization does provide a large-sample approximation to (4a and 4b): As the sizes of the treatment groups increase, violations of (4a and 4b) become smaller in probability, and thus random

confounding also becomes smaller in probability. In this sense, randomization provides a large-sample approximation to perfect exchangeability (4a, 4b). Furthermore, the usual standard errors for D provide a large-sample measure of the unconditional variation in D produced by the random confounding.

The impact of adjustment, revisited

Note first that, for each treatment level a of A , the unadjusted estimate of $\Pr(Y_a = y)$ is $\Pr(Y_a = y|A = a)$; setting undefined terms to zero, the latter equals

$$\sum_b \Pr(Y_a = y|A = a, B = b) \Pr(B = b|A = a) \tag{6}$$

where the sum is over all observed values b of B . The component proportions $\Pr(Y_a = y|A = a, B = b)$ and $\Pr(B = b|A = a)$ in (6) vary over allocations and are uncorrelated with one another. Random confounding of D by B arises because $\Pr(B = b|A = 1)$ may differ randomly from $\Pr(B = b|A = 0)$, resulting in differences in weighting of the B strata in the unadjusted estimates $\Pr(Y_1 = y|A = 1)$ and $\Pr(Y_0 = y|A = 0)$. In contrast, the unsmoothed (“nonparametric”) estimate of $\Pr(Y_a = y)$ standardized to the distribution of B in the full cohort is

$$\sum_b \Pr(Y_a = y|A = a, B = b) \Pr(B = b) \tag{7}$$

whenever all terms are defined. The distribution of B in the full baseline cohort is not affected by the randomization of A . Thus, for each treatment level a of A , the proportions $\Pr(B = b)$ which serve as the weights in (7) are constant over allocations and always the same between treatment groups, eliminating random confounding. This observation may provide an intuition as to how removal of random confounding reduces the unconditional variance of adjusted estimators relative to the unadjusted estimator D .

The standardized estimate (7) can be recast as the unsmoothed inverse treatment-proportion weighted (IPW) estimate of $\Pr(Y_a = y)$ [40, 41],

$$\sum_b \Pr(Y_a = y, A = a, B = b) / \Pr(A = a|B = b) \tag{8}$$

whenever all terms are defined. This formulation leads to numerous modifications which further improve precision and confounding control [31, 40, 42]. In contrast, because the treatment probabilities under simple randomization are all 1/2, the inverse probability of treatment weighted estimate is

$$\sum_b \Pr(Y_a = y, A = a, B = b) / 1/2 \tag{9}$$

This formula simplifies to $2\Pr(Y_A = y, A = a)$ which involves no adjustment by B , and hence will be confounded randomly; in fact (9) will equal the unadjusted proportion $\Pr(Y_A = y|A = a)$ when the proportion n_1/n with $A = 1$ is fixed by design at $1/2$.

Despite the fact that (9) uses the actual (allocation) probabilities and (8) does not, (8) and its generalizations (in which $\Pr(A = a|B = b)$ is replaced by a smoothed proportion) are usually called inverse-probability-of-treatment weighted (IPTW) estimates [3, 40]. Nonetheless, the $\Pr(A = a|B = b)$ in (8) are not actual treatment-assignment probabilities, but are instead post-allocation probabilities of seeing $A = a$ for an individual randomly sampled from the subcohort with $B = b$. Not even this limited interpretation is precisely correct for smoothed proportions, although one might describe those proportions as the hypothetical assignment probabilities that generate conditionally unbiased estimates from the inverse-proportion formula (8) under the smoothing assumptions.

Effects of stratified randomization

Along with their shortcomings for simple randomization, neither Fig. 1a nor Eq. (1) is sufficient for representing allocation mechanisms or distributions that induce unfaithfulness, including trial designs in which random confounding is constrained by randomization within blocks (which reduces the variance of the observed difference D). This fact violates suggestions that faithfulness can be viewed as an assumption that conditional independence relations are due to causal structure rather than to accidents of parameter values [13, 14].

To illustrate the above points and build models that can capture these distinctions, suppose the cohort size n is even and consider the following forms of randomization with equal allocation to $A = 1$ and $A = 0$, so that the unconditional probability of $A = 1$ and the final proportion with $A = 1$ are $1/2$:

1. Unstratified fixed-size randomization: Draw a random sample without replacement of fixed size $n/2$ from the total n and assign them $A = 1$, then assign the remainder $A = 0$. This design allows all $\binom{n}{n/2}$ possible equal allocations (=924 when $n = 12$, in contrast to the $2^{12} = 4,096$ possible allocations if unequal allocations were also allowed).
2. Stratified (blocked) randomization with strata equal by design: Stratify the initial cohort on baseline factors B , select equal even numbers from each stratum, and use equal-allocation randomization within each stratum. A binary $B = 1,0$ yields two strata of size $n/2$, leading to

$\binom{n/2}{n/4}^2$ possible allocations (=400 when $n = 12$, with two strata of size 6).

3. Stratified (blocked) randomization using natural strata: Stratify the cohort based on baseline factors B , and use equal-allocation randomization within each stratum. The stratum totals may vary, and for a stratum with an odd total, one would have to randomly exclude an individual. With two strata of sizes k_1 and k_2 , this design allows $\binom{k_1}{k_1/2} \binom{k_2}{k_2/2}$ possible allocations (=420 when $n = 12$ with $k_1 = 4, k_2 = 8$).

In all three designs, the probability of $A = 1$ remains $1/2$ upon conditioning on B , so that A and B are marginally independent before allocation, as implied by Fig. 1a. Nonetheless, there are profound statistical and causal differences among them.

Most importantly, in (2) and (3) random confounding by B is prevented by stratification, because stratified randomization restricts allocations to those exhibiting no marginal AB association and thus reduces the set of allocations that must be considered in frequency calculations such as permutation tests. This reduction is reflected by the fact that, in the examples with $n = 12$, fewer than half of the unstratified allocations are allowable stratified allocations (i.e., allocations in which A and B are unassociated). With no marginal AB association, the usual unadjusted and B -adjusted estimates of Δ will be identical.

Note that B can mechanically (if indirectly) affect the individual assignments A in the stratified designs. To illustrate, consider a two-stratum design (2) in which individuals are enrolled sequentially in the order indicated by i . Once the allocator has filled the $A = 1$ quota of $n/4$ for stratum $B = 1$, further recruits with $B = 1$ must be given $A = 0$, whereas treatments for those with $B = 0$ will not be limited in any way by what has happened among recruits with $B = 1$. If we are given that $n/4$ previous recruits had $A = 1$ and $B = 1$, the probability of $A_i = 1$ drops from $1/2$ to zero when we are additionally given that $B_i = 1$, but remains $1/2$ if we are given instead that $B_i = 0$.

In such examples, we see a strong causal effect (via the allocator) of B on A in the stratified designs, which must be accounted for in classical exact permutation inference. Thus, these designs need an arrow from B to A , as in Fig. 1b, in which we see an open confounding path from A to Y via B ; yet Fig. 1a, which represents the unstratified design, has no confounding path A to Y . It is somewhat paradoxical that the stratified designs eliminate confounding via B entirely, yet Fig. 1b is identical to a graph representing an observational study with confounding by B . At the same time, the unstratified design allows random confounding of D by B , yet Fig. 1a is usually described as

representing no confounding by B . We consider these discrepancies to be a problem with ordinary causal graphs rather than with our conceptualization of confounding.

Stratified designs (2) and (3) provide examples of *design unfaithfulness*: By design, B has an effect on A that creates not only pre-allocation marginal independence of B and A but also forces that independence to hold in the final cohort allocation. In contrast, unstratified randomization enforces only pre-allocation independence. The difference between (2) and (3) is that (2) suggests a more complex underlying structure in which the cohort is selected from a larger population to induce a particular distribution for B as well as for A . We will not pursue this elaboration since our points apply without it.

Population models

To more precisely represent design effects, we consider causal models defined on the entire study cohort (this parallels survey-sampling theory, which considers distributions of entire samples rather than of individuals [43]), assuming outcomes are independent given treatment allocation. Denote the complete cohort column vectors of individual variables by bold type: $\mathbf{A} = (A_1, \dots, A_n)'$, $\mathbf{Y}_1 = (Y_{11}, \dots, Y_{1n})'$, $\mathbf{Y}_0 = (Y_{01}, \dots, Y_{0n})'$, $\mathbf{Y}_A = (Y_{A1}, \dots, Y_{An})'$, and $\mathbf{B} = (B_1, \dots, B_n)'$ (if the B_i are row vectors corresponding to individual covariate records, \mathbf{B} is the matrix with rows B_i). A particular treatment allocation for the full cohort is a vector $\mathbf{a} = (a_1, \dots, a_n)$ of possible values for A (ones and zeros in our example), while a particular covariate distribution for the cohort is a vector $\mathbf{b} = (b_1, \dots, b_n)$ of possible values for B . We write $\mathbf{A} = \mathbf{a}$ when the value of A_i is given by entry a_i of \mathbf{a} .

To recast earlier concepts in terms of population-level variables, let $\mathbf{1}$ and $\mathbf{0}$ denote vectors of n ones and n zeros, and recall that the inner (dot) product $\mathbf{x}'\mathbf{z}$ of two vectors is the sum of the product of their components, $\sum x_i z_i$, so that $\mathbf{1}'\mathbf{x} = \sum x_i$. Then

1. the number treated is $n_1 = \mathbf{1}'\mathbf{a} = \sum a_i$, and thus equal allocation ($n_1 = n_0 = n/2$) is the constraint $\mathbf{1}'\mathbf{a} = n/2$.
2. with binary B , $\mathbf{1}'\mathbf{b}$ is the number with $B = 1$ and stratified equal allocation is the pair of constraints $\mathbf{b}'\mathbf{a} = \mathbf{1}'\mathbf{b}/2$, $\mathbf{1}'\mathbf{a} = n/2$.
3. the random pre-allocation cohort outcome variable is $\mathbf{Y}_A = \mathbf{A}'\mathbf{Y}_1 + (\mathbf{1} - \mathbf{A})'\mathbf{Y}_0$.
4. the cohort outcome that would result from a specific allocation \mathbf{a} is $\mathbf{Y}_a = \mathbf{a}'\mathbf{Y}_1 + (\mathbf{1} - \mathbf{a})'\mathbf{Y}_0$.
5. the causal marginal mean difference is $\Delta = \mathbf{1}'\mathbf{Y}_1/n - \mathbf{1}'\mathbf{Y}_0/n$; and

6. the unadjusted estimator of Δ is $D = (\mathbf{A}'\mathbf{Y}_1)/n_1 - \{(\mathbf{1} - \mathbf{A})'\mathbf{Y}_0\}/n_0$.

Statistical inferences may be constructed by comparing the observed \mathbf{Y}_a to the distribution of \mathbf{Y}_A induced by the distribution of \mathbf{A} given hypothesized structures (constraints) on \mathbf{Y}_1 and \mathbf{Y}_0 [10, 20].

Figure 2 provides population-level analogs of the graphs in Fig. 1, and looks identical but for the substitution of bolded letters for italic letters. The purpose of emphasizing such a seemingly trivial notation change is to remind us that many statistically important structural features must play a role in any statistical inferences on effects, even if only a marginal mean difference Δ is the target. Absence of an arrow from \mathbf{A} to \mathbf{Y} represents the constraint that $\mathbf{Y}_a = \mathbf{Y}_{a^*}$ for all pairs of cohort allocations \mathbf{a}, \mathbf{a}^* (the sharp null hypothesis of no effect on anyone), which makes every treatment allocation result in the same observed population outcome (i.e., \mathbf{Y}_A does not vary with allocation). Presence of the arrow means that no such constraint is imposed by the graph, so that there *may* be a pair of distinct allocations \mathbf{a}, \mathbf{a}^* that would produce different outcomes: $\mathbf{Y}_a \neq \mathbf{Y}_{a^*}$. As before, however, it does *not* imply that a given measure of effect is non-null, e.g., Δ may be 0 even if treatment changes everyone's outcome.

Absence of an arrow from \mathbf{B} to \mathbf{A} in Fig. 2a applies to design (1) where \mathbf{B} plays no role in allocation. In contrast, under designs (2) and (3), \mathbf{B} can causally affect the \mathbf{A} entries, necessitating an arrow from \mathbf{B} to \mathbf{A} . This is easily seen under design (3), where \mathbf{B} can vary naturally; for example, with $n = 12$ and B binary (e.g., smoking, non-smoking) having 6 individuals with $B = 1$ (so $\mathbf{1}'\mathbf{b} = \sum b_i = 6$) allows $\binom{6}{3} \binom{6}{3} = 400$ possible values for \mathbf{A} , while having 4 individuals with $B = 1$ ($\mathbf{1}'\mathbf{b} = 4$) allows $\binom{4}{2} \binom{8}{4} = 420$ possible values for \mathbf{A} . The purpose of these design effects is to improve precision, but this benefit will be recognized only if the effects are accounted for by the analyst.

In designs (2) and (3), the relation of \mathbf{B} to \mathbf{A} has been constrained so that every allowable pair (\mathbf{b}, \mathbf{a}) exhibits no association between the \mathbf{b} and \mathbf{a} entries (i.e., cross-

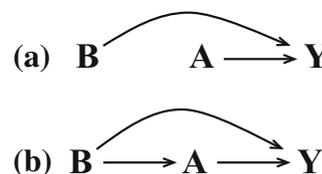


Fig. 2 Population-level causal diagrams representing (a) simple randomization and (b) B -stratified randomization

tabulation of observable b_i and a_i values would show perfect independence of A and B , with not even random associations). This independence is an example of unfaithfulness that is forced by the allocation algorithm. As with the individual graph (Fig. 1b), Fig. 2b fails to capture that unfaithful constraint; nonetheless, it does alert us that some aspect of the entire treatment distribution \mathbf{A} may depend on the entire covariate distribution \mathbf{B} .

In summary, balanced blocked designs require \mathbf{B} to have an arrow to \mathbf{A} , yet the design effect represented by that arrow leads to no marginal association of \mathbf{B} and \mathbf{A} —an example of unfaithfulness (independence despite graphical connectedness) that is stable. Balanced blocked designs are the experimental analog of balanced matched-cohort designs, in which the exposure \mathbf{A} and matching variable \mathbf{B} are independent in the matched subcohort selected from a larger cohort, despite being connected in the causal graph [15]. We note however that this unfaithful independence will usually be broken upon adjustment for unblocked or unmatched covariates, necessitating adjustment for all the variables [44].

Discussion

Our examples display limits of causal models that are ambiguous about the level or form of causal action, and address certain technical misunderstandings of the relation of DAG models to probability distributions and confounding. With allowance for small-sample issues, our observations about mean-difference measures also apply to collapsible ratio measures such as risk ratios and survival-time ratios. We caution however that several complications arise from the noncollapsibility of odds ratios and logistic regression coefficients: Stratification on balanced outcome predictors results in a change in the estimated causal parameter [5, 16, 45], which is often mistaken for confounding (although as in linear models such stratification can also improve the power of null tests [46, 47]; similar phenomena arise in probit and other binary-outcome models [48, 49]).

Misunderstandings of modern nonparametric DAGs may have been encouraged by traditional path diagrams for linear structural relations (LISRELs) and multivariate-normal models, in which all effects are represented by constant additive (and hence monotonic) effects on individuals. In these highly constrained models, unfaithfulness to a single arrow cannot occur, and thus certain design effects cannot be modeled. Causal DAGs do not preclude such effects, but do not distinguish designs that prevent random confounding from those that do not. This limitation arises because the DAG component in these models represents the joint distribution of the variables, rather than a

sample realization (put another way, the DAGs represent “infinite superpopulations” free of random variation).

Failure to recognize that causal DAGs do not incorporate randomly generated effects but do allow unfaithfulness (necessitating arrows even when conventional measures of association are null) can lead to erroneous interpretations and inferences. We thus think it important to emphasize these limitations of causal DAGs:

- (a) Epidemiologic biases can arise from baseline covariates graphically separated (disconnected) from the study exposure or treatment; in particular, it is important to understand that graphical criteria for confounding control [2, 25] apply only to non-random (stable or structural) confounding, and do not account for random confounding.
- (b) A covariate may have important effects on the distribution of exposure or treatment (e.g., by producing dependencies across individuals), altering variances even if it does not affect the marginal mean parameters or other distributional summaries; consequently, design effects must be accounted for in statistical procedures, and the arrow from the covariate to the variable should be included to alert the user to these effects.

Population-level variables do not address these graphical limitations directly, but do connect graphs to population potential-outcome models which can exhibit dependencies among individual variables. Population models may include not only outcome dependencies (as with contagious diseases), but also treatment dependencies, such as design constraints that prevent random confounding by allowing potential confounders to affect selection or assignment. We thus suggest that, for experimental as well as for observational research, description and teaching of causal models should be generalized to represent explicit population and distributions, and should explicitly exclude faithfulness from its basic assumptions [1].

Acknowledgments The authors are grateful to the reviewers and Katherine Hoggatt for comments leading to clarification of several key points.

References

1. Glymour MM, Greenland S. Causal diagrams. In: Rothman KJ, Greenland S, Lash T, editors. *Modern epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008. p. 183–209.
2. Pearl J. *Causality*. 2nd ed. New York: Cambridge University Press; 2009.
3. Hernán MA, Robins JM. *Causal inference*. London: Chapman & Hall/CRC; 2015.
4. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol*. 1986;15:413–9.
5. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci*. 1999;14:29–46.

6. Greenland S, Robins JM. Identifiability, exchangeability and confounding revisited. *Epidemiol Perspect Innov.* 2009;6:4.
7. Rothman KJ. Epidemiologic methods in clinical trials. *Cancer.* 1977;39:1771–5.
8. Greenland S, Neutra RR. Control of confounding in the assessment of medical technology. *Int J Epidemiol.* 1980;9:361–7.
9. Matthews JNS. An introduction to randomised controlled clinical trials. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2006.
10. Robins JM. Confidence intervals for causal parameters. *Stat Med.* 1988;7:773–85.
11. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc.* 1987;82:387–94.
12. Senn SJ. Testing for baseline balance in clinical trials. *Stat Med.* 1994;13:1715–26.
13. Robins JM, Scheines R, Spirtes P, Wasserman L. Uniform consistency in causal inference. *Biometrika.* 2003;90:491–515.
14. Spirtes P, Glymour C, Scheines R. Causation, prediction, and search. 2nd ed. Cambridge: MIT Press; 2001.
15. Mansournia MA, Hernán MA, Greenland S. Matched designs and causal diagrams. *Int J Epidemiol.* 2013;42:860–9.
16. Mansournia MA, Greenland S. The relation of collapsibility and confounding to faithfulness and stability. *Epidemiology.* 2015;26: in press.
17. Cox DR, Hinkley DV. Theoretical statistics. New York: Chapman and Hall; 1974.
18. Greenland S. Randomization, statistics, and causal inference. *Epidemiology.* 1990;1:421–9.
19. Greenland S. On the logical justification of conditional tests for two-by-two contingency tables. *Am Stat.* 1991;45:248–51.
20. Rosenbaum PR. Observational studies. 2nd ed. New York: Springer; 2002.
21. Richardson TS, Robins JM. Single world intervention graphs: a primer. Second UAI Workshop on Causal Structure Learning, Bellevue, Washington, 2013.
22. Robins JM, Richardson TS. Alternative graphical causal models and the identification of direct effects. In: Shrouf P, Keyes K, Ornstein K, editors. Causality and psychopathology: finding the determinants of disorders and their cures. Oxford: Oxford University Press; 2011. p. 1–52.
23. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology.* 1999;10:37–48.
24. Greenland S, Pearl J. Causal diagrams. In: Boslaugh S, editor. Encyclopedia of epidemiology. Thousand Oaks: Sage Publications; 2008. p. 149–56.
25. Pearl J. Causal diagrams for empirical research. *Biometrika.* 1995;82:669–710.
26. Rubin DB. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics.* 1991;47:1213–34.
27. Rosenbaum P, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70:41–55.
28. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal.* 2007;15:199–236.
29. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci.* 2010;25:1–21.
30. Waernbaum I. Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Stat Med.* 2012;31:1572–81.
31. Imai K, Ratkovic M. Covariate balancing propensity score. *J R Stat Soc Ser B.* 2014;76:243–63.
32. Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, Robins JM. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol.* 2006;163:262–70.
33. Fisher RA. The design of experiments. Edinburgh: Oliver and Boyd; 1935.
34. Robins JM, Morgenstern H. The foundations of confounding in epidemiology. *Comp Math Appl.* 1987;14:869–916.
35. Senn S. Seven myths of randomisation in clinical trials. *Stat Med.* 2013;32:1439–50.
36. Rosenberger WF, Lachin JM. Randomization in clinical trials: theory and practice. New York: Wiley; 2002.
37. Friedman LM, Furberg CD, DeMets DL. Fundamentals of clinical trials. 4th ed. New York: Springer; 2010.
38. Chow SC, Liu JP. Design and analysis of clinical trials. 2nd ed. Hoboken: Wiley; 2004.
39. Lauritzen SL, Dawid AP, Larsen BN, Leimar HG. Independence properties of directed Markov fields. *Networks.* 1990;20: 491–505.
40. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11:550–60.
41. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology.* 2003;14:680–6.
42. Kang JD, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Stat Sci.* 2007;22: 523–80.
43. Cochran WG. Sampling techniques. 3rd ed. New York: Wiley; 1977.
44. Sjölander A, Greenland S. Ignoring the matching variables in cohort studies—When is it valid and why? *Stat Med.* 2013; 32:4696–708.
45. Greenland S, Pearl J. Adjustments and their consequences—collapsibility analysis using graphical models. *Int Stat Rev.* 2011;79:401–26.
46. Gail MH. Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In: Moolgavkar SH, Prentice RL, editors. Modern statistical methods in chronic disease epidemiology. New York: Wiley; 1986. p. 3–18.
47. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression. *Int Stat Rev.* 1991;59:227–40.
48. Neuhaus J, Jewell NP. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika.* 1993;80:807–15.
49. Robinson LD, Dorroh JR, Lien D, Tiku ML. The effects of covariate adjustment in generalized linear models. *Commun Stat Theory Methods.* 1998;27:1653–75.