Title

On Causal Inferences for Personalized Medicine: How Hidden Causal Assumptions Led to Erroneous Causal Claims About the D-Value

Journal

The American Statistician, 74(3)

ISSN

0003-1305

Authors

Greenland, Sander
Fay, Michael P
Brittain, Erica H
et al.

Publication Date

2020-07-02

DOI

10.1080/00031305.2019.1575771

Peer reviewed

HHS Public Access

Author manuscript

Am Stat. Author manuscript; available in PMC 2021 January 22.

Published in final edited form as:

Am Stat. 2020 ; 74(3): 243–248. doi:10.1080/00031305.2019.1575771.

On Causal Inferences for Personalized Medicine: How Hidden Causal Assumptions Led to Erroneous Causal Claims About the D-Value

Sander Greenland[1], **Michael P. Fay**[2], **Erica H. Brittain**[2], **Joanna H. Shih**[3], **Dean A. Follmann**[2], **Erin E. Gabriel**[4], **James M. Robins**[5]

[1]Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA, U.S.A., lesdomes@ucla.edu

[2]Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda MD, U.S.A.

[3]Biometric Research Branch, National Cancer Institute, Rockville, MD, U.S.A.

[4]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden

[5]Department of Epidemiology and Department of Biostatistics, Harvard T. Chan School of Public Health, Boston, MA

Abstract

Personalized medicine asks if a new treatment will help a particular patient, rather than if it improves the average response in a population. Without a causal model to distinguish these questions, interpretational mistakes arise. These mistakes are seen in an article by Demidenko [2016] that recommends the "D-value," which is the probability that a randomly chosen person from the new-treatment group has a higher value for the outcome than a randomly chosen person from the control-treatment group. The abstract states "The D-value has a clear interpretation as the proportion of patients who get worse after the treatment" with similar assertions appearing later. We show these statements are incorrect because they require assumptions about the potential outcomes which are neither testable in randomized experiments nor plausible in general. The D-value will *not* equal the proportion of patients who get worse after treatment if (as expected) those outcomes are correlated. Independence of potential outcomes is unrealistic and eliminates *any* personalized treatment effects; with dependence, the D-value can even imply treatment is better than control *even though most patients are harmed by the treatment*. Thus, D-values are misleading for personalized medicine. To prevent misunderstandings, we advise incorporating causal models into basic statistics education.

Keywords

Causality; D-value; Effect size; Individualized treatment; Patient-centered outcomes; Personalized medicine; Probability of causation

1. CAUSAL MODELS FOR INDIVIDUAL EFFECTS

1.1. Background: The D-value is Not the Proportion of Patients Harmed by Treatment

Despite great advances in causal modeling over recent decades, it still seems largely unrecognized that intuitive causal interpretations of statistical parameters can be rendered fallacious via their dependence on hidden assumptions about causal structure. This problem is well illustrated with a recent article (Demidenko [2016]) which made a number of remarkable causal claims about a measure of association called the D-value.[1] Over the ensuing two years it became one of the most downloaded articles in *The American Statistician* – which is alarming in light of the fact that all the causal claims in the article are incorrect. Foremost, the article states in the abstract that "The D-value has a clear interpretation as the proportion of patients who get worse after the treatment." This statement is false under even the simplest plausible causal model, yet is repeated on p. 37. Related incorrect assertions appear throughout the article, e.g., on p. 36 "The D-value is for personalized medicine when the treatment is sought, not on a group, but on an individual level". Worse, under the quoted misinterpretations the D-value can indicate that the new treatment is better than the control (D<0.5) *even though the treatment harms a majority of its recipients* (Hand [1992], Fay et al. [2018]).

How were such profoundly erroneous claims justified? We will show that the claims can be derived by introducing a statistically nonidentified causal assumption, one which we regard as extremely implausible in every setting we can imagine. Because similar hidden assumptions appear to be behind other common misinterpretations of effect measures, and given the attention received by Demidenko [2016], we provide a detailed review of the core problem: failure to recognize when interpretations are based on strong and often implausible assumptions about the effect of treatment on outcome.

1.2. Experimental statistics and modern causal modeling theory

Scientists have been making causal inferences from experiments since the dawn of modern science, and formal causal models for the design and analysis of randomized experiments go back nearly a century (Neyman [1923], Welch [1937]). Since then, the topic of causal inference has become a major sub-branch of statistical theory with the growing recognition that causal claims require additional formal structure to link them deductively to statistical analyses. Starting with the earliest randomization-test literature, by far the most popular structures for this purpose have been in the form of potential outcomes, which have gradually begun to appear in textbooks, e.g., see recent examples by Morgan and Winship [2015], Imbens and Rubin [2015], VanderWeele [2015], Pearl et al. [2016], Rosenbaum [2017], and Hernán and Robins [2019]. Using this model to study effects of treatment, the traditional outcome or response variable is expanded into a list (vector) of potential outcomes that shows what the outcome would be for the same unit under different treatments; conditional independence ("ignorability") assumptions about treatment assignment are then used to deduce tests and estimates of causal effects, with the latter

[1]The D-value (possibly with tie adjustments) is also known as the Mann-Whitney parameter (Fay et al. [2018]), the probabilistic index (Thas et al. [2012]), the relative treatment effect (Brunner and Munzel [2000]), and for discrimination problems the concordance index or c-index (Harrell et al. [1996]), where it equals the area under the ROC curve (Demidenko [2016]).

defined as contrasts comparing the marginal distributions of the different potential outcomes in the same study group (e.g., all trial participants).

The resulting causal-inference theory is more complex than classical regression or ANOVA, for there is now a vector variable inserted where only one outcome variable appeared before. Beyond that however it is distinct from traditional mathematical statistics to the extent that the latter focuses strictly on deductions derived from assumptions about probability distributions for the data. It has long been known that even complete knowledge of those probability distributions cannot pinpoint the exact causal mechanism (the explanation or "story") generating the data; for example, two mutually inconsistent potential-outcome models arising from distinct mechanisms can lead to identical data distributions, even when treatment is randomized (Robins [1986, sec. 2A]; Robins and Greenland [1989ab]; Dawid [2000]; Pearl [2009, Section 11.1.1]).

Thus, although experimental designs can identify and contrast various marginal distributions for the component potential outcomes, without further assumptions they cannot identify their joint (vector) distribution. That limitation may be unsurprising: Once we apply a treatment to a group, we can only observe its outcome distribution under that treatment; the potential outcomes under the unreceived (counterfactual) treatments are now unobservable. Some critics have labelled this nonidentification a problem with potential-outcome models; others have responded that it is instead a valuable representation of an innate limitation of purely statistical studies of causation, e.g., contrast Dawid [2000] with its discussants.[2] As we will explain in detail, it is precisely these inherent limitations that were overlooked in Demidenko [2016].

To show how the quoted claims fail, we will use a basic potential-outcome model for treatment effects in which X_i is what the outcome of a randomly selected trial subject indexed by *i* would be if given the control treatment (e.g., placebo or standard of care) and Y_i is what the outcome of the same subject would be if given the new treatment (e.g., an experimental drug). In more common causal notation X_i is $Y_i(0)$ and Y_i is $Y_i(1)$, but we here use X_i and Y_i to better match Demidenko's notation. Let (X_i,Y_i) be the potential outcome vector for a random subject, i.e., the outcomes under control or new treatment for subject *i*. Randomization splits the subjects into two random samples from the total: A control sample in which only the control response X is observed and Y is missing; and a new-treatment sample in which only the new-treatment response Y is observed and X is missing. The randomization allows estimation (identification) of the marginal X and Y distributions by classical techniques (Rubin [1978]).

Nonetheless, as has been long known (e.g., see Dawid [2000] and discussants), treatment randomization[3] does not identify the distribution of the pairwise (joint) potential-outcome

[2]Indeed, one alternative to potential outcomes (Dawid [2015]) deals with the problem by using a reduced structure in which only marginal responses to treatments are defined. This causal model reproduces the same observable distributions as those derived from potential-outcome models, and so yields identical inferential implications and limitations of statistical experiments; it thus lacks sufficient structure to correctly represent concepts like "proportion harmed" and "probability of causation" which contrast potential outcomes within single individuals.

[3]This means any weaker condition such as ignorability (independence of treatment and potential outcomes across patients) will also not identify the (X_i,Y_i) distribution.

(X_i,Y_i) distribution. Specifically, the randomized-trial design allows identification of only the marginal distributions of X_i and Y_i. In particular, if higher outcome values are considered undesirable, the proportion of patients who would be harmed by the treatment is $\pi = \Pr(Y_i>X_i)$, which is a statement about the pairs (X_i,Y_i). It thus cannot be identified from the trial data alone because it depends on the joint (X_i,Y_i) distribution. Because no complete (X_i,Y_i) pair is observed, we cannot verify harm $(Y_i>X_i)$, benefit $(Y_i<X_i)$, or no effect $(Y_i=X_i)$ in any individual from the data alone; we only observe marginal averages over these individuals (Senn [2009]).

2. WHY THE D-VALUE IS NOT THE PROPORTION HARMED

2.1. D-value vs. Proportion Harmed Under a Basic Model

Demidenko [2016] introduced the D-value first under normality assumptions for X and Y, and then (in his Section 3.1) non-parametrically, referring to the D-value as both the sample estimate and the population parameter. We focus on the D-value parameter, first defining it generally, then examine assumptions and their consequences.

Suppose we have N subjects in the study; then the complete unobserved potential outcomes are the N pairs, (X_1,Y_1), …(X_N,Y_N). The treatment-assignment process picks n of the N subjects to get the control treatment (say subjects 2,3,5,8,…) and N-n subjects to get new treatment (say 1,4,6,7,9,…). Then we only observe (., Y_1), $(X_2,.)$, $(X_3,.)$, (.,Y_4),$(X_5,.)$, (.,Y_6), (.,Y_7), $(X_8,.)$, (.,Y_9), and so on, where the "." represents the missing part of the pair (Rubin [1978]). The D-value population parameter is $\delta = \Pr(Y_i > X_j)$, where Y_i is the outcome of a patient randomly drawn from the new-treatment group, and X_j is the outcome of a distinct patient independently and randomly drawn from the control group (cf. Section 3.1 of Demidenko [2016]). Note well, δ is defined using different indices *i* and *j* to indicate that the two random variables in this sampling *cannot* be from the same individual, since δ is derived from sampling two distinct groups.

When larger response denotes more harm, one might be tempted to shorten the description of δ to "the proportion harmed," leading to the incorrect causal interpretations in Demidenko [2016]. To see why, we must contrast the D-value, $\delta = \Pr(Y_i > X_j)$, to the proportion harmed, which is $\pi = \Pr(Y_i >X_i)$. The key difference overlooked in Demidenko [2016] is that (in sharp contrast to δ), the indices of X and Y in the proportion harmed π are by definition *identical*: Use of the same index *i* in π denotes that the two outcomes must come from the *same* individual. This follows common sense: To claim patient *i* was harmed by the new treatment is to claim that patient *i* would have done better on the control treatment, i.e., that $Y_i>X_i$.

To highlight the difference between δ and π, we revisit the example in Demidenko [2016] which compares placebo to a weight-loss drug in a randomized trial. The average ending weight is one pound less in the drug group than the placebo group; the corresponding D-value is 0.486, which the paper states is "the proportion of patients who get worse after the treatment." Suppose however that the 1-pound average loss arose because the diet drug produced some weight loss in *every* patient and on average that loss was one pound (e.g., as would happen if the treatment causes a one-pound weight loss relative to placebo in

everyone: $Y_i - X_i = -1$ for all i); then no one was or would be harmed by treatment ($\pi=0$). Thus the D-value of 0.486 cannot be the proportion harmed.

If the outcome were transient with no carry-over or time trend, and the trial could be repeated in the same patient (an N-of-1 crossover trial), we could see both Y_i and X_i, and thus see directly whether a patient was harmed ($Y_i>X_i$), unaffected ($Y_i=X_i$), or helped ($Y_i<X_i$) by the treatment. Otherwise, one must turn to a causal model to impute the missing potential outcome (X_i missing in the treatment group, Y_i missing in the control group) (Greenland et al. [2008]; Morgan and Winship [2015]; Imbens and Rubin [2015]; Hernán and Robins [2019]). The error in Demidenko [2016] can thus be attributed to not holding the individual subscript i constant when claiming to estimate the proportion harmed π or other "personalized" effects of treatment, and thus failing to recognize that the marginal distributions identified by randomization do not determine the joint distribution on which the proportion harmed π is defined. This oversight will invalidate any statistical claim about personalized effects based solely on the treatment-assignment mechanism (as opposed to assumptions about the mechanism connecting treatment to the outcome).

2.2 D-value with Normality Assumptions

Demidenko [2016, Section 3] assumed the marginal distributions were normal, but made no explicit assumptions about the bivariate distribution F_{xy}. Gadbury and Iyer [2000] assumed that the (X_i,Y_i) pairs were draws from a bivariate normal distribution; they noted that π is not identifiable from a two-sample randomized experiment because the correlation between X_i and Y_i is not identifiable (since as noted above for each individual only one of the potential outcomes is observed). Under the bivariate normality assumption, we get $\delta = \pi$ if and only if the correlation is 0 so that X_i and Y_i are independent. Although π is not identifiable, Gadbury and Iyer [2000] explored bounds on π under the bivariate normality assumption. Even without bivariate normality, if the outcomes of the same subject under different treatments are independent, $\Pr(X_i,Y_i) = \Pr(X_i)\Pr(Y_i)$; under randomization the latter product equals $\Pr(X_i)\Pr(Y_j)$, leading to $\delta = \pi$; however, this independence would mean a subject's control outcome has no predictive value at all for their treatment outcome. As we hope will become obvious to the reader, this is an untenable assumption.

It follows that any statement derived from confusing the D-value with the proportion harmed can be highly misleading, because strong positive correlations of potential outcomes should be expected: regardless of treatment, an individual will retain the same baseline outcome predictors, such as age, sex, baseline weight, their entire genome, their entire biome, and countless unmeasured predictors. Thus, a patient who would have ended with a higher outcome than other patients under control is also likely to end with a higher outcome under the new treatment, simply by virtue of having identical baseline predictors.

Not only can δ and π differ to a large extent, they can even differ in direction in the sense of $\delta<½<\pi$, and thus convey a very different impression of typical direction of effect, so that the new treatment could look better in terms of the D-value (i.e., $\delta < ½$) but could actually be worse for most patients (i.e., $\pi>½$). Hand (1992) showed how this could happen even if X and Y are marginally normal; specifically, with means μ_x and μ_y and the same variance, σ^2, for X and Y, then it is possible to have opposite directional effects, and these opposite effects

can occur whenever $\left|\frac{\mu_y - \mu_x}{\sigma}\right| < 1.35$, a result sometimes called "Hand's Paradox." Under these normality assumptions, the D-value is $\delta = \Phi\left(\frac{\mu_y - \mu_x}{\sqrt{2}\,\sigma}\right)$, where Φ is the standard normal distribution; it follows that the paradox (i.e., δ and π indicate opposite direction of effects) can occur whenever $\delta \in (0.170, 0.830)$ (Fay et al [2018]). So even if δ indicates that treatment is worse than control, we cannot rule out the possibility that treatment benefits more than half the population. Basically, with δ, large harms for the minority can outweigh small benefits for the majority (and vice versa).

2.3. D-value without Normality Assumptions

To illustrate how the D-value behaves without normality assumptions, we explore three simple discrete-outcome examples. Let F_{xy} be the bivariate distribution of (X_i,Y_i), and let F_x and F_y be the marginal distribution of X_i and Y_i, respectively. Figure 1 shows three bivariate distributions F_{xy}, all with the same marginal distributions, F_x and F_y. The marginal distribution for X has probability $\frac{1}{3}$ on each of the points: 2, 4, and 6, while the marginal distribution on Y has probability $\frac{1}{3}$ on each of the points: 1, 3, and 5. As before, assume larger values are worse, so parameter values less than ½ imply that the new treatment is better than the control. Since the marginal distributions are the same in all three panels in Figure 1, the δ values are the same: all have $\delta = \Pr(Y_i > X_j) = \frac{1}{3}$. Figure 1a shows the bivariate distribution when X and Y are independent; this is the case when $\pi = \Pr(Y_i > X_i) = \delta = \frac{1}{3}$. Figure 1b shows the bivariate distribution with a constant effect $Y_i - X_i = -1$ for all patients, for which case $\pi=0$ since everyone benefits from treatment (the effect is strictly downward monotonic); yet, under Demidenko's misinterpretation of δ, we would falsely infer that $\delta=\frac{1}{3}$ of patients would do worse under the new treatment than under the control. Finally, in Figure 1c illustrates Hand's paradox where $\pi=\frac{2}{3}>½$, so that the new treatment harms most patients, yet under Demidenko's misinterpretation of $\delta=\frac{1}{3}<½$, we would falsely infer that most patients are not harmed by the new treatment. Of course, Hand's paradox can occur with continuous data as well. To see this, imagine mixtures of bivariate normal distributions, with their centers at the points in Figure 1c.

As in the case where F_{xy} is bivariate normal, in the general case we can identify δ but not π using only the marginal distributions F_x and F_y (which are all that are identified by the randomization indicator alone, as in classical randomization tests and their approximations).[4] Nevertheless, we can get bounds on π. Fay et al. [2018] review bounds for π given the marginal distributions F_x and F_y under both non-parametric and semi-parametric assumptions.

For continuously distributed responses, $F_x=F_y$ implies $\delta=½$, but for discrete responses we can have $\delta<½$ when, unlike the examples of Figure 1, there is the possibility of ties (i.e., $\Pr(Y_i=X_j)>0$). We could adjust the D-value for ties using $\Pr(Y_i>X_j) + \Pr(Y_i=X_j)/2$, so that when $F_x=F_y$ then the adjusted D-value parameter equals ½ even with discrete responses with ties. We could similarly adjust π to $\Pr(Y_i>X_i) + \Pr(Y_i=X_i)/2$ but this adjustment would

[4]To see this note that δ is a functional $g_\delta(F_X,F_Y) = \Pr(Y_i > X_j)$ of the randomization-identified marginal distributions, whereas π is a functional $g_\pi(F_{XY}) = \Pr(Y_i > X_i)$ of the nonidentified bivariate distribution.

destroy its interpretation as the proportion harmed; for example, π would then equal ½ for a treatment with no harm or benefit, i.e., $\Pr(Y_i=X_i) = 1$, and would equal ¼ for a treatment that harmed no one and benefitted half the patients, i.e., $\Pr(Y_i>X_i)=0$, $\Pr(Y_i=X_i)=\Pr(Y_i<X_i)=½$.

3. Personalized Medicine and Statistical Inferences

In order for any measure to be used for personalized medicine, we need baseline covariates to describe different subgroups with different effects. Without subgroup analysis we can only approach individual effects via their group averages. Examples in simple two-sample randomized trials include the average causal effect $E(Y_i) – E(X_i) = E(Y_i - X_i)$, which shows how the usual mean difference is a simple average of individual effects, a property not shared by other measures; notably, odds ratios are not weighted averages of individual causal effects (Greenland et al. [1999]). In the next section we show how $\delta - ½$ can be reconstructed as an average individual effect, albeit not a simple one. Because these effects are averaged over individuals for each group, it is misleading to state “the D-value is for personalized medicine when the treatment is sought, not on a group, but on an individual level” (Demidenko [2016, p. 36]).

If we want to come closer to individual effects we must measure baseline covariates to allow us to “personalize” the effects. Given data sufficient in quantity, quality, and detail (as from large trials and pooling projects), personalized treatment decisions can be aided by estimating potential outcomes as functions of baseline covariates Z available in the trial data (Robins [1986, sec. 3C; 2004], Murphy [2003]). Even with such resources however we think it probable that Y and X will remain highly correlated within covariate levels, and thus may render the covariate-specific D-values far from the actual proportion harmed.

4. A Causal Interpretation of the D-value

To summarize to this point: We can interpret δ causally by viewing its departure from its null value of ½ as one measure of the change in the marginal outcome distribution produced by the new treatment relative to the control. Nonetheless, as we have seen this interpretation is not a patient-specific (“individualized”) causal effect because it averages over within-patient effects within each group, so the same average effect can come from very different distributions of individual effects. For example, $\delta=½$ if there is no individual effect ($Y_i=X_i$ for everyone, in which case $\pi=0$), but $\delta=½$ could also occur if *everyone* was affected (e.g., if half the patients had $Y_i - X_i=1$ and half had $Y_i - X_i=-1$, in which case $\pi=\delta= ½$).

To understand the limitations of δ within a causal model, we can imagine that the patients enrolled in the trial define a population of potential outcomes under control treatment (the X marginal distribution) and a population of potential outcomes under the new treatment (the Y marginal distribution). Then δ is the probability that a randomly selected potential outcome under the new treatment is larger than an *independently* selected potential outcome under the control treatment. In contrast, π refers to pairs (X_i,Y_i) of these potential outcomes matched by patient: π is the probability that a randomly selected *patient* has a potential

outcome under new treatment that is larger than that *same* patient's potential outcome under the control treatment.

While in general $\delta \neq \pi$, we may ask if there is at least some sort of precise causal interpretation for δ in terms of the change in outcome *distribution* produced by treatment. It turns out there is one for the null-centered version $\delta - 1/2$, albeit one much less straightforward than the causal interpretation for π because it depends on averaging explicit quantiles of the marginal distributions of X and Y under the different treatments. Specifically, Fay et al. [2018] derive $\delta - 1/2$ in terms of what they call the expected quantile difference (EQD) causal effect.[5] For simplicity we will only describe the EQD under the continuity assumption; see Fay et al. [2018] for the general tie-adjusted version.

We first define the randomized outcome quantile level for the total patient population in a 1:1 randomized experiment. Let the observed outcome be $W_i = R_iY_i + (1-R_i)X_i$, where R_i is the independent random (Bernoulli) treatment-assignment indicator with $P(R=1) = 1/2$. w_i represents randomized observation of one of the two potential outcomes Y and X for the i^{th} individual, which yields the i^{th} individual's observed response. The distribution of w_i is $G(w) = (F_x(w) + F_y(w))/2 = \Pr(W_i \leq w)$. With $G(w_i) = q_i$, the observed outcome w_i for patient i is then at the q_i^{th} quantile of G, and q_i is the quantile level of the observed outcome in the total patient population. The difference in quantile levels for the i^{th} individual due to treatment is $D_i = G(Y_i) - G(X_i)$ which Fay et al. [2018] called the quantile difference causal effect for the i^{th} individual. D_i represents an individual causal treatment effect insofar as it measures the change in the individual's ordinal outcome location in the total patient population produced by the new treatment. For example, if D_i is 0.10, then the quantile level for the i^{th} individual would increase by 0.10 on treatment (e.g., go from 0.16 on control to 0.26 on treatment). Fay et al [2018] showed that the expectation of those individual causal treatment effects is $\delta - 1/2$, i.e.,

$$\delta - {}^{1}/_{2} = \mathrm{E}(D_i) = \mathrm{E}(G(Y_i)) - \mathrm{E}(G(X_i)).$$

Since G is a function of the marginal distributions only, δ is identifiable. So a proper causal interpretation of $\delta - 1/2$ from a randomized experiment is the expectation of individual difference in quantile level causal effects. Although the causal interpretation of δ is less straightforward than that of π, the lack of identifiability of π limits its practical use, and so it may be unsurprising that most of the modern causal modeling literature has focused on models for treatment effects on marginal distributions.

5. Discussion

The difference between the D-value parameter δ and the proportion harmed π has long been known (e.g., see Hand [1992]) yet has been overlooked repeatedly. One reason for the oversight may be that in a randomized trial the *observed* outcomes on which δ is defined (X_j and Y_i from different patients j and i) *are* independent, while the potential-outcome pairs on

[5] which should not be confused with the quantile causal effect $F_y^{-1}(Y_i) - F_x^{-1}(X_i)$; see Xu et al [2018].

which π is defined (X_i and Y_i from the *same* patient i) are likely to be highly correlated. Confusion thus seems inevitable when (as in Demidenko [2016]) there is no underlying causal model that distinguishes the two cases. Once such a model is given, the mistake of confusing δ with π can be seen as a mistake of equating independent marginal sampling to joint-distribution sampling (sampling of patient-specific potential-outcome pairs).

This issue is closely related to other confusions of marginal and individual causal measures. Consider a binary outcome or failure-time outcome. Most textbooks of the past century as well as many legal and policy documents mistakenly equated the attributable fraction $\varphi = (\Pr(Y_i=1)-\Pr(X_i=1))/\Pr(Y_i=1)$ ("attributable risk" or excess fraction) to the probability of causation $\Pr(Y_i>X_i)/\Pr(Y_i=1) = \pi/\Pr(Y_i=1)$. But as above for δ versus π, the fraction φ is defined from the marginal distributions and is thus identified by treatment randomization, whereas π is not (Greenland and Robins [2000]). Without strong assumptions about the bivariate distribution of potential outcomes (X_i,Y_i) we can only say $\varphi \le \pi/\Pr(Y_i=1) \le 1$ (Robins and Greenland [1989ab]).

Demidenko [2016] repeats the well-known fact that a P-value does not represent an effect-size estimate, which certainly bears emphasis. Unfortunately, his article concludes that replacing the P-value with the D-value offers a partial solution to "the poor reproducibility of scientific experiments." We strongly disagree with that statement regardless of the chosen measure of effect. Any measure is vulnerable to data-driven or otherwise motivated selection effects by the analyst, including artefacts of subgroup analysis. Even with perfectly honest and valid reporting, in fields like psychology, social sciences, and medicine, effect sizes can and should be expected to vary dramatically across groups due to differences in the distribution of innumerable modifiers of the effect. It is for example often lamented that the effects of drugs seen in randomized trials do not predict well the effects seen in clinical practice – a fact unsurprising when one realizes that trials have numerous admission criteria that could alter effects (e.g., excluding reproductive-age women), but clinical prescribers can and do prescribe far beyond those limits and practice in a much less controlled environment.

Those estimation problems aside, there seems to be poor understanding that a valid P-value should vary dramatically across studies (Senn [2001, 2002], Gelman and Stern [2006], Boos and Stefanski [2011], Greenland [2019]). Consequently, "replication failures" are often simply an artefact of using an arbitrary and easily crossed threshold such as 0.05 to dichotomize results into "positive" and "negative" and from focusing on only one hypothesis and its P-value (Poole [1987], Rothman et al. [1999]. Senn [2001, 2002], Greenland [2017a], Amrhein et al. [2019]). Such dichotomania and nullism has been condemned in scores if not hundreds of sources, and is addressed simply by reporting precisely and discussing the different P-values that one gets from testing different parameter values, rather than focusing on where those P-values fall relative to some arbitrary if conventional cutoff (in this journal see recent articles by Wasserman and Lazar [2016], Greenland et al. [2016], and Greenland [2019]). Simply adopting a different measure of effect does nothing to address this core source of spurious "nonreproducibility."

The main goal of this note has been to highlight the misinterpretation of statistical parameters that arise when the parameter is not precisely defined by an explicit causal

model. This has been illustrated by claims that the D-value (δ) is the proportion harmed (π), when in fact those are two different parameters which may not even be close under realistic assumptions. It is difficult to see that the two parameters are different without using potential outcomes or similar causal formalisms. Although there remains disagreement about the relative merits of formal and informal approaches to causal inference (Greenland [2017b]), we strongly advise that basic causal models become part of elementary statistics education so that mistaken intuitive interpretations of effect estimates can be avoided and valid methods for effect prediction can be more widely deployed.

REFERENCES

Amrhein V, Trafimow D, and Greenland S (2019), "Inferential Statistics are Descriptive Statistics," The American Statistician, in press.

Boos DD and Stefanski LA (2011), "P-Value Precision and Reproducibility," The American Statistician, 65, 213–221. [PubMed: 22690019]

Brunner E, & Munzel U (2000), "The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation," Biometrical Journal, 42(1), 17–25.

Dawid AP (2000), "Causal inference without counterfactuals" (with discussion), Journal of the American Statistical Association, 95, 407–448.

Dawid AP (2015), "Statistical causality from a decision-theoretic perspective," Annual Review of Statistics and Its Application, 2(1):273–303, 2015.

Demidenko E (2016), "The p-value you can't buy," The American Statistician, 70(1), 33–38. [PubMed: 27226647]

Divine GW, Norton HJ, Baron AE, and Juarez-Colunga E (2018), "The Wilcoxon-Mann-Whitney procedure fails as a test of medians," The American Statistician, 72(3), 278–286.

Fay MP, Brittain EH, Shih JA, Follmann DA and Gabriel EE (2018), "Causal estimands and confidence intervals associated with Wilcoxon-Mann-Whitney tests in randomized experiments," Statistics in Medicine, 37, 2923–2937. [PubMed: 29774591]

Gelman A, and Stern H (2006). "The difference between 'significant' and 'not significant' is not itself statistically significant," The American Statistician, 60, 4, 328–331.

Greenland S (2017a), "The need for cognitive science in methodology," American Journal of Epidemiology, 186, 639–645, Open access at 10.1093/aje/kwx259. [PubMed: 28938712]

Greenland S (2017b), "For and against methodology: Some perspectives on recent causal and statistical inference debates," European Journal of Epidemiology, 32, 3–20. Open access at 10.1007/s10654-017-0230-6 [PubMed: 28220361]

Greenland S (2018), "Valid P-values behave exactly as they should: Some misleading criticisms of P-values and their resolution with S-values," The American Statistician, 72, in press.

Greenland S and Robins JM (2000), "Epidemiology, justice, and the probability of causation," Jurimetrics, 40, 321–340.

Greenland S, Rothman KJ, and Lash TL (2008), "Measures of effect and measures of association," Ch. 4 in Rothman KJ, Greenland S, and Lash TL (eds.), Modern Epidemiology, 3rd ed. Philadelphia: Lippincott-Wolters-Kluwer.rd

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN and Altman DG (2016), "Statistical Tests, P-values, Confidence Intervals, and Power: A Guide to Misinterpretations," The American Statistician, 70, Online Supplement 1 to ASA Statement on P-values, Open access at http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_file/utas_a_1154108_sm5368.pdf; reprinted in the *European Journal of Epidemiology*, 31, 337–350.

Hand DJ (1992), "On comparing two treatments," The American Statistician, 46(3):190–192.

Harrell FE, Lee KL, and Mark DB (1996), "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," Statistics in medicine, 15(4), 361–387. [PubMed: 8668867]

Hernán MA and Robins JM (2018), Causal Inference, Bosa Raton: Chapman & Hall/CRC, forthcoming.

Imbens G and Rubin DB (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction, New York, Cambridge University Press.

Morgan S and Winship C (2015), Counterfactuals and Causal Inference: Methods and Principles for Social Research, 2nd ed., New York, Cambridge University Press.nd

Murphy SA (2003), "Optimal dynamic treatment regimes," Journal of the Royal Statistical Society, Series B, 65, 331–355.

Pearl J (2009). Causality: Models, Reasoning, and Inference, 2nd ed. New York: Cambridge University Press.nd

Pearl J, Glymour M and Jewell NP (2016), Causal Inference in Statistics: A Primer, New York: Wiley.

Poole C (1987), "Beyond the confidence interval," American Journal of Public Health, 77, 195–199. [PubMed: 3799860]

Robins JM (2004), "Optimal structural nested models for optimal sequential decisions," Proceedings of the Second Seattle Symposium in Biostatistics, New York: Springer, 189–326.

Robins JM and Greenland S (1989a), "Estimability and estimation of excess and etiologic fractions," Statistics in Medicine, 8, 845–859. [PubMed: 2772444]

Robins JM and Greenland S (1989b), "The probability of causation under a stochastic model for individual risks," Biometrics, 45, 1125–1138. [PubMed: 2611320]

Rosenbaum PR (2017). Observation and Experiment: An Introduction to Causal Inference. Cambridge MA: Harvard University Press.

Rothman KJ, Johnson ES, Sugano DS (1999), "Is flutamide effective in patients with bilateral orchiectomy?," The Lancet, 353, 1184.

Rubin DB (1978), "Bayesian inference for causal effects: the role of randomization," Annals of Statistics, 6, 34–58.

Senn SJ (2001), "Two Cheers for P-Values," Journal of Epidemiology and Biostatistics, 6, 193–204. [PubMed: 11434499]

Senn SJ (2002), Letter to the Editor re: Goodman 1992, Statistics in Medicine, 21, 2437–2444. [PubMed: 12210627]

Senn S (2009), "Three things that every medical writer should know about statistics," The Journal of the European Medical Writers Association, 18, 159–162.

Thas O, Neve JD, Clement L, & Ottoy JP (2012), "Probabilistic index models," Journal of the Royal Statistical Society Series B, 74(4), 623–671.

Walker E and Nowacki AS (2010), "Understanding Equivalence and Noninferiority Testing," Journal of General Internal Medicine, 26, 192–196. [PubMed: 20857339]

Wasserstein RL, and Lazar NA (2016), "The ASA's statement on p-values: context, process, and purpose," The American Statistician, 70(2), 129–133.

Xu D, Daniels MJ, and Winterstein AG (2018), "A Bayesian nonparametric approach to causal inference on quantiles" Biometrics, 74(3), 986–996. [PubMed: 29478267]

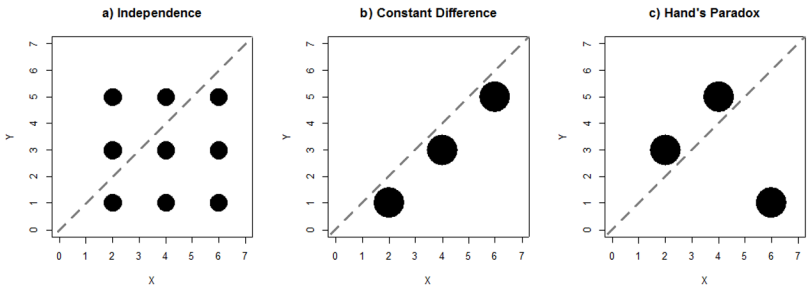

Figure 1:
Three bivariate distributions, all with the same marginal distributions: F_x has equal probabilities on 2,4, and 6, and F_y has equal probabilities on 1,3, and 5. Circle areas are proportional to the probabilities.