Research Article

Scott Mueller* and Judea Pearl

Personalized decision making – A conceptual introduction

https://doi.org/10.1515/jci-2022-0050
received August 06, 2022; accepted February 16, 2023

Abstract: Personalized decision making targets the behavior of a specific individual, while population-based decision making concerns a subpopulation resembling that individual. This article clarifies the distinction between the two and explains why the former leads to more informed decisions. We further show that by combining experimental and observational studies, we can obtain valuable information about individual behavior and, consequently, improve decisions over those obtained from experimental studies alone. In particular, we show examples where such a combination discriminates between individuals who can benefit from a treatment and those who cannot – information that would not be revealed by experimental studies alone. We outline areas where this method could be of benefit to both policy makers and individuals involved.

Keywords: causality, individual treatment effect, conditional average treatment effect, PNS, monotonicity

MSC 2020: 62C99, 62D20

1 Introduction

The purpose of this article is to provide a conceptual understanding of the distinction between personalized and population-based decision making, and to demonstrate both the advantages of the former and how it could be achieved.

Formally, this distinction is captured in the following two causal effects. Personalized decision making optimizes the individual treatment effect (ITE):

$$\mathrm{ITE}(u) = Y(1, u) - Y(0, u), \tag{1}$$

where $Y(x, u)$ stands for the outcome that individual u would attain had decision $x \in \{1, 0\}$ been taken.[1] In contrast, population-based decision making optimizes the conditional average treatment effect (CATE)[2]:

[1] The formal definitions of ITE (or individual causal effect), based on structural causal models, are given in ref. [1, Section 3]. However, these definitions are not necessary for understanding our results; they apply regardless of the model used to derive those quantities. By a “unit,” we mean any entity (e.g., a patient, a customer, or an agricultural plot) whose behavior affects decisions. Beyond its measured features $C(u)$, u contains *all* characteristics of an individual u, measured and unmeasured, known and unknown, sufficiently detailed to make the response Y a deterministic function of the treatment.

[2] Theoretically, CATE can be viewed as a function of c, the vector of pretreatment characteristics observed on individual u, since CATE will be equal for two different u and u' such that $C(u) = C(u') = c$. However, to emphasize the distinction between u, the individual for whom a decision is contemplated, and the individuals u' in the study, we write $\mathrm{CATE}(u) = \mathrm{CATE}(C(u) = c)$. Clearly, an individual u with characteristics $C(u) = c$ obtains a unique CATE measure, given by equation (2).

*** Corresponding author: Scott Mueller,** Computer Science Department, University of California, Los Angeles, California, United States, e-mail: scott@cs.ucla.edu
Judea Pearl: Computer Science Department, University of California, Los Angeles, California, United States, e-mail: judea@cs.ucla.edu

$$\mathrm{CATE}(u) = E[Y(1, u') - Y(0, u')|C(u') = C(u)], \tag{2}$$

where $C(u)$ stands for a vector of pretreatment characteristics observed on individual u, and the average is taken over all units u' that share these characteristics.

We will show in this article that the two objective functions lead to different decision strategies and that, although $\mathrm{ITE}(u)$ is in general not identifiable, informative bounds on the probability distribution of $\mathrm{ITE}(u)$, for any given individual u, can nevertheless be obtained from aggregate data by combining experimental and observational studies.[3] Notably, these aggregate data allow us to improve decision making at the individual level. We will further demonstrate how these bounds can improve decisions that would otherwise be taken using $\mathrm{CATE}(u)$ alone as an objective function.

Note that, although both $\mathrm{ITE}(u)$ and $\mathrm{CATE}(u)$ are properties of an individual u and are defined in terms of counterfactual expressions, the former is in general nonidentifiable, while the latter is a *do*-expression, $\mathrm{CATE}(u) = E[Y|\mathrm{do}(X = 1), C(u)] - E[Y|\mathrm{do}(X = 0), C(u)]$ estimable directly from experimental data without invoking counterfactual assumptions. Note also that the words “individual” and “personalized” used in this article refer to the individual for whom a decision is contemplated and, although $\mathrm{ITE}(u)$ is the same for all individuals u' sharing measured characteristics $C(u)$ with u, decisions should vary from u to u' depending on their distinct, often unmeasured, personal utilities and beliefs. Our analysis aims to inform decision makers of the likely behavior of a randomly chosen individual in the population, so as to match decisions to the distinct preferences and beliefs of individuals u and u'.

This article is organized as follows. Section 2 demonstrates, using an extreme example, two rather surprising findings. First, that population data are capable of providing decisive information on individual response, and second, that nonexperimental data, usually discarded as bias-prone, can add information (regarding individual response) beyond that provided by a randomized controlled trial (RCT) alone. Section 3 generalizes these findings using a more realistic example and further demonstrates how critical decisions can be made using the information obtained and their ramifications both to the targeted individual and to a population-minded policy maker. Section 4 casts the findings of Section 3 in a numerical setting, allowing for a quantitative appreciation of the magnitudes involved. This analysis leads to actionable policies that guarantee risk-free benefits in certain populations. Section 4 explains the mathematics that gives rise to the results featured in Section 3, as well as the assumptions upon which they depend. Confounding in observational studies, typically problematic and in need of adjusting for, is shown to be helpful in narrowing the probabilities of benefit and harm. Section 5 discusses related topics such as monotonicity, number needed to treat (NNT), and probability of harm. Finally, Section 6 provides an annotated bibliography of the source papers, including related works.

For conceptual clarity, well-designed RCTs and observational studies are assumed throughout. As such, we consider RCTs as having 100% compliance and no selection bias or any other imperfections that often plague them (e.g., placebo effects). Similarly, observational studies are assumed to provide unbiased estimates of the statistical associations or conditional expectations they are designed to assess.

Trialists are usually suspicious of observational studies because the latter are either bias-prone or rely on subjective assumptions of “no confounding,” which are hardly testable. Such trepidations do not apply to our analysis for two reasons. First, our analysis makes no modeling assumptions whatsoever when interpreting observational studies, and second, we actually benefit from the presence of confounding in the observational studies.

2 Preliminary semi-qualitative example

Our target of analysis is an individual response to a given treatment, namely, how an individual would react if given treatment and if denied treatment. Since no individual can be subjected to both treatment and its

[3] Formally, we will derive bounds for the proportion of individuals having particular values for ITE (e.g., $P(\mathrm{ITE}(u) = t)$). When Y is binary, we provide bounds on $P(\mathrm{ITE}(u) = 1) = P(Y(1, u) > Y(0, u))$ (i.e., the proportion of individuals benefiting from treatment) and $P(\mathrm{ITE}(u) = -1) = P(Y(1, u) < P(Y(0, u)))$ (i.e., the proportion of individuals harmed by treatment). In this way, we supplement the information usually provided by the RCT, $\mathrm{ATE}(u)$, with two additional parameters that may be crucial for individual decision making.

denial, its response function must be inferred from population data, originating from one or several studies. Therefore, we are asking: To what degree can population data inform us about an individual response?

Before tackling this general question, we wish to address two conceptual hurdles. First, why should population data provide *any* information whatsoever on the individual response, and second, why should nonexperimental data add any information (regarding individual response) to what we can learn with an RCT alone? The next simple example will demonstrate both points.

We conduct an RCT and find no difference between treatment (drug) and control (placebo), say 10% in both treatment and control groups die, while the rest (90%) survive. This makes us conclude that the drug is ineffective, but also leaves us uncertain between (at least) two competing models:

- Model-1 – The drug has no effect whatsoever on any individual and
- Model-2 – The drug saves 10% of the population and kills another 10%.

From a policy maker viewpoint, the two models may be deemed equivalent; the drug has zero average effect on the target population. But from an individual viewpoint, the two models differ substantially in the sets of risks and opportunities they offer. According to Model-1, the drug is useless but safe. According to Model-2, however, the drug may be deemed dangerous by some and a life saver by others.[4]

To see how such attitudes may emerge, assume, for the sake of argument, that the drug also provides temporary pain relief. Model-1 would be deemed desirable and safe by all, whereas Model-2 will scare away those who do not urgently need the pain relief, while offering a glimpse of hope to those whose suffering has become unbearable and who would be ready to risk death for the chance (10%) of recovery (hoping, of course, they are among the lucky beneficiaries).

Another reason for diverse individual decisions in the face of Model-2 is individual beliefs. For example, a person may believe the drug will not be harmful to them even though it has a 10% probability of harm in the population at large. Maybe a family member took the drug and recovered. In that case, the drug certainly was not harmful to that family member. This person would choose to take the drug under our Model-2 (the drug saves 10% of the population and kills another 10%) and reject it under Model-1 because, assuming immunity, there are still factors of pain, expense, and discomfort to consider. A different person, whose cousin happened to die after taking the drug, may not have confidence in natural immunity and will choose to refuse the drug under Model-2.

It should now be clear that individuals can have unique preferences for how they weigh the probability of benefit and the probability of harm. They may additionally place weights on the probability of being doomed (death, regardless of taking the drug or not) and the probability of being immune (recovery, regardless of taking the drug or not). Li and Pearl [3] provide bounds, or point estimates when certain assumptions can be made, on a linear combination of the probabilities of benefit, harm, being doomed, and being immune. They demonstrate how the average treatment effect (ATE) is a suboptimal criterion for decision making in light of the weights pertaining to individual preferences. Mueller and Pearl [4] supply an example of this weighting of probability of benefit and probability of harm in assessing which Covid-19 patients are in greatest need of treatment.

This simple example will also allow us to illustrate the second theme of our article – the crucial role of observational studies. We will now show that supplementing the RCT with an observational study on the same population (conducted, *e.g.*, by an independent survey of patients who have the option of taking or avoiding the drug) would allow us to decide between the two models, totally changing our understanding of what risks await an individual taking the drug.

Consider an extreme case where the observational study shows 100% survival in both drug-choosing and drug-avoiding patients, as if each patient knew in advance where danger lies and managed to avoid it. Such a finding, though extreme and unlikely, immediately rules out Model-1 which claims no treatment effect on any individual. This is because the mere fact that patients succeed 100% of the time to avoid harm

[4] Greenland and Robins [2] tabulated four types of individuals with binary treatment and outcome: doomed, exposure causative, exposure preventative, and immune.

where harm does exist (revealed through the 10% death in the randomized trial) means that choice makes a difference, contrary to Model-1's claim that choice makes no difference.

The reader will surely see that the same argument applies when the probability of survival among option-having individuals is not precisely 100% but simply higher (or lower) than the probability of survival in the RCT. By using the RCT study alone, in contrast, we were unable to rule out Model-1, or even to distinguish Model-1 from Model-2.

We now present another edge case where Model-2, rather than Model-1, is ruled out as impossible. Assume the observational study informs us that all those who chose the drug died and all who avoided the drug survived. It seems that drug choosers were truly dumb, while drug avoiders knew precisely what was good for them. This is perfectly feasible, but it also tells us that no one can be *cured* by the drug, contrary to the assertion made by Model-2, that the drug cures 10% and kills 10%. To be cured, a person must survive if treated and die if not treated. But none of the drug choosers could have been cured, because they all died, and none of the drug avoiders could have been cured, because they all survived (they might have survived had they taken the drug, but then it would not have been the drug that cured them). Thus, Model-2 cannot explain these observational results and must be ruled out.

We should emphasize that although Model-2 tells an individual that she has a 10% chance of being killed by the drug and a 10% chance of being saved by the drug, she cannot know which outcome pertains to her specifically. All she knows is that 10% of people with her characteristics are killed by the drug and another 10% will be cured due to the drug. This is information that is suppressed by the RCT, which does not distinguish between Model-1 and Model-2.

Now that we have demonstrated conceptually how certain combinations of observational and experimental data can provide information on individual behavior that each study alone cannot, and we are ready to go to a more realistic motivating example that, based on theoretical bounds derived in ref. [5], establishes individual behavior for any combination of observational and experimental data[5] and, moreover, demonstrates critical decision making ramifications of the information obtained.

3 Motivating numerical example

The next example to be considered deals with the effect of a drug on two subpopulations, males and females. Unlike the extreme case considered in Section 2, the drug is found to be somewhat effective for both males and females and, in addition, deaths are found to occur in the observational study as well. We will demonstrate that, although men and women are totally indistinguishable in the RCT study, adding observational data proves men to react markedly different than women, calling for two different treatment policies in the two groups. Although a woman has a 28% chance of benefiting from the drug and no danger at all of being harmed by it, a man has a 49% chance of benefiting from it and as much as a 21% chance of dying because of it.

To cast the story in a realistic setting, we imagine the testing of a new drug, aimed to help patients suffering from a deadly disease. An RCT is conducted to evaluate the efficacy of the drug, and it is found to be 28% effective[6] in both males and females; in other words, $\text{CATE}(\text{male}) = \text{CATE}(\text{female}) = 0.28$. The drug is approved and, after a year of use, a follow-up randomized study is conducted yielding the same results, namely, CATE remained 0.28, and moreover, men and women remained totally indistinguishable in their responses, as shown in Table 1.

[5] The example we will work out happened to be identifiable due to particular combinations of data, though, in general, the data will not permit point estimates of individual causal effects and the bounds will not be as narrow. This motivating example is merely an extreme case concocted to explain the origin of the bound-narrowing effect.

[6] To simplify matters, we are treating each experimental study data as an ideal RCT, with 100% compliance and no selection bias or any other biases that often plague RCTs.

Table 1: Female vs male CATE

	Female survivals	Male survivals
do(drug)	489/1,000 (49%)	490/1,000 (49%)
do(no drug)	210/1,000 (21%)	210/1,000 (21%)
CATE	28%	28%

Let us focus on the second RCT (Table 1), since the first was used for drug approval only, and its findings are the same as the second. The RCT tells us that there was a 28% improvement, on average, in taking the drug compared to not taking the drug. This was the case among both females and males: $\text{CATE}(\text{female}) = \text{CATE}(\text{male}) = 0.28$, where do(drug) and do(no-drug) are the treatment and control arms, respectively, in the RCT. It thus appears reasonable to conclude that the drug has a net remedial effect on some patients and that every patient, be it male or female, should be advised to take the drug and benefit from its promise of increasing one's chances of recovery (by 28%).

At this point, the drug manufacturer ventured to find out to what degree people actually buy the approved drug, following its recommended usage. A market survey was conducted (observational study) and revealed[7] that only 70% of men and 70% of women actually chose to take the drug; problems with side effects and rumors of unexpected deaths may have caused the other 30% to avoid it. A careful examination of the observational study has further revealed substantial differences in survival rates of men and women who chose to use the drug (shown in Tables 2 and 3). The rate of recovery among drug-choosing men was exactly the same as that among the drug-avoiding men (70% for each), but the rate of recovery among drug-choosing women was 43% lower than among drug-avoiding women (0.27 vs 0.70, in Table 2). It appears as though many women who chose the drug were already in an advanced stage of the disease, which may account for their low recovery rate of 27%.

At this point, having data from both experimental and observational studies, we can estimate the probability $P(\text{ITE}(u) > 0|C(u) = c) = P(\text{benefit}|C(u) = c)$ for both a typical man and a typical woman. Quantitative analysis shows (Section 4) that, with the data given earlier, the drug affects men markedly differently from the way it affects women. Although a woman has a 28% chance of benefiting from the drug and no danger at all of being harmed by it, a man has a 49% chance of benefiting from it and as much as a 21% chance of dying because of it – a serious cause for concern. Note that based on the experimental data alone (Table 1), no difference at all can be noticed between men and women.

The ramifications of these findings on personal decision making are broad. First, they tell us that the drug is not as safe as the RCT would have us believe; it may cause death in a sizable fraction of patients. Second, they tell us that a woman is totally clear of such dangers and should have no hesitation to take the drug, unlike a man, who faces a decision; a 21% chance of being harmed by the drug is cause for concern. Physicians, likewise, should be aware of the risks involved before recommending the drug to a man. Third, the data tell policy makers what the overall societal benefit would be if the drug is administered to women only; 28% of the drug takers would survive who would otherwise die. Finally, knowing the relative sizes of the benefiting vs harmed subpopulations opens the door to finding the mechanisms responsible for the differences, as well as to identifying measurable markers that characterize those subpopulations.

For example,

- Our analysis has identified "sex" to be an important feature, separating those who are harmed from those saved by the drug. In the same way, we can leverage other measured features, say family history, a genetic marker, or a side-effect, and check whether they shrink the sizes of the susceptible subpopulations. The results would be a set of features that approximate responses at the individual level. Note

[7] As with the experimental studies, observational studies are assumed to provide unbiased estimates for the conditional probabilities involved. We note that observational studies provide an easier arena for obtaining representative samples of the target population, partly due to the ease of recruiting units and partly due to their noninvasive nature.

Table 2: Female survival and recovery data

		Survivals	Deaths	Total
Experimental	do(drug)	489 (49%)	511 (51%)	1,000 (50%)
	do(no drug)	210 (21%)	790 (79%)	1,000 (50%)
Observational	drug	378 (27%)	1,022 (73%)	1,400 (70%)
	no drug	420 (70%)	180 (30%)	600 (30%)

Table 3: Male survival and recovery data

		Survivals	Deaths	Total
Experimental	do(drug)	490 (49%)	510 (51%)	1,000 (50%)
	do(no drug)	210 (21%)	790 (79%)	1,000 (50%)
Observational	drug	980 (70%)	420 (30%)	1,400 (70%)
	no drug	420 (70%)	180 (30%)	600 (30%)

again that absent observational data and a calculus for combining them with the RCT data, we would not be able to identify such informative features. A feature like “sex” would be deemed irrelevant, since men and women were indistinguishable in our RCT studies.

- Our ability to identify relevant informative features as described earlier can be leveraged to amplify the potential benefits of the drug. For example, if we identify a marker that characterizes men who would die only if they take the drug and prevent those patients from taking the drug, the drug would cure 62% of male patients who would be allowed to use it. This is because we do not administer the drug to the 21% who would have been killed by the drug. Those patients will now survive, so a total of 70% of patients will be cured because of this combination of marker identification and drug administration. This unveils an enormous potential of the drug at hand, which was totally concealed by the 28% effectiveness estimated in the RCT studies.

4 How the results were obtained

For the purpose of analysis, let us denote y_t as recovery among the RCT treatment group and y_c as recovery among the RCT control group. The causal effects for treatment and control groups, $P(y_t|\text{Gender})$ and $P(y_c|\text{Gender})$, were the same,[8] no differences were noted between males and females.

In addition to the aforementioned RCT, we posited an observational study (survey) conducted on the same population. Let us denote $P(y|t, \text{Gender})$ and $P(y|c, \text{Gender})$ as recovery among the drug choosers and recovery among the drug avoiders, respectively.

With this notation at hand, our problem is to compute the probability of benefit

$$P(\text{benefit}) = P(y_t, y'_c) \tag{3}$$

from the following data sources: $P(y_t)$, $P(y_c)$, $P(y|t)$, $P(y|c)$, and $P(t)$. The first two denote the data obtained from the RCT, and the last three denote data obtained from the survey. Nonrecovery is represented by y', so

[8] $P(y_t|\text{female})$ was rounded up from 48.9 to 49%. The 0.001 difference between $P(y_t|\text{female})$ and $P(y_t|\text{male})$ was not necessary, but was constructed to allow for clean point estimates.

y_c' is nonrecovery among the RCT control group. Equation (3) should be interpreted as the probability that an individual would both recover if assigned to the RCT treatment arm and die if assigned to control.[9]

The results of the observational and experimental studies are not independent of each other since, barring selection bias, participants in the two studies are selected from the same overall population, ideally consisting of the eventual users of the drug. At the individual level, the connection between behaviors in the two studies relies on an assumption known as *consistency* [1,6],[10] asserting that an individual's response to treatment depends entirely on biological factors, unaffected by the settings in which treatment is taken.[11] In other words, the outcome of a person choosing the drug would be the same had this person been assigned to the treatment group in an RCT study. Similarly, if we observe someone avoiding the drug, their outcome is the same as if they were in the control group of our RCT.

In terms of our notation, consistency implies:

$$P(y_t|t) = P(y|t), P(y_c|c) = P(y|c). \tag{4}$$

In words, the probability that a drug chooser would recover in the treatment arm of the RCT, $P(y_t|t)$, is the same as the probability of recovery in the observational study, $P(y|t)$.

On the basis of this assumption, and leveraging both experimental and observational data, Tian and Pearl [5] derived the following tight bounds on the probability of benefit,[12] as defined in (3):

$$\max\left\{\begin{array}{c} 0, \\ P(y_t) - P(y_c), \\ P(y) - P(y_c), \\ P(y_t) - P(y) \end{array}\right\} \leqslant P(\text{benefit}) \leqslant \min\left\{\begin{array}{c} P(y_t), \\ P(y_c'), \\ P(t, y) + P(c, y'), \\ P(y_t) - P(y_c) + \\ P(t, y') + P(c, y) \end{array}\right\}. \tag{5}$$

Here, $P(y_c')$ stands for $1 - P(y_c)$, namely, the probability of death in the control group. The same bounds hold for any subpopulation, say males or females, if every term in (5) is conditioned on the appropriate class.

Applying these expressions to the female data from Table 2 gives the following bounds on $P(\text{benefit}|\text{female})$:

$$\begin{gathered} \max\{0, 0.279, 0.09, 0.189\} \leqslant P(\text{benefit}|\text{female}) \leqslant \min\{0.489, 0.79, 0.279, 1\}, \\ 0.279 \leqslant P(\text{benefit}|\text{female}) \leqslant 0.279. \end{gathered} \tag{6}$$

Similarly, for men, we obtain:

[9] Tian and Pearl [5] called $P(\text{benefit})$ "probability of necessity and sufficiency" (PNS). The relationship between PNS and ITE (1) is explicated in Section 6.

[10] Consistency is a property imposed at the individual level, often written as follows:

$$Y = X \cdot Y(1) + (1 - X) \cdot Y(0).$$

for binary X and Y. Rubin [7] considered consistency to be an assumption in SUTVA, which defines the potential outcome framework. Pearl [6] considered consistency to be a theorem of Structural Equation Models, a violation of which reflects imperfections (e.g., placebo effects) in RCT practices.

[11] In medical practices, clinical experts rarely rely on the assumption of biological equivalence. The very participation in a study tends to create fears and expectations that affect patients' response to treatment. Moreover, selection bias [8] is a major problem in clinical trials, since subjects are recruited by stringent health criteria and, unlike those in observational studies, they must undergo consent procedures. For these two reasons, RCT practitioners compare only patients that undergo the same recruitment procedure and, accordingly, report only the difference $P(y_t) - P(y_c)$ (for treatment see https://ucla.in/3Fv8rxL). More elaborate procedures [8] must be deployed to overcome both selection bias and placebo effects when experimental and observational studies are to be combined.

[12] New methods have expanded on these bounds to include nonbinary outcomes [9], though without incorporating observational data. Li and Pearl advanced the $P(\text{benefit})$ bounds to exploit observational data *and* work with nonbinary treatment and outcomes [10].

$$\begin{aligned} \max\{0, 0.28, 0.49, -0.21\} \leqslant P(\text{benefit}|\text{male}) &\leqslant \min\{0.49, 0.79, 0.58, 0.7\}, \\ 0.49 \leqslant P(\text{benefit}|\text{male}) &\leqslant 0.49. \end{aligned} \tag{7}$$

Thus, the bounds for both females and males, in (6) and (7), collapse to point estimates:

$$\begin{aligned} P(\text{benefit}|\text{female}) &= 0.279, \\ P(\text{benefit}|\text{male}) &= 0.49. \end{aligned}$$

We are not always so fortunate to have a complete set of observational and experimental data at our disposal. When some data are absent, we are allowed to discard arguments to max or min in (5) that depend on that data. For example, if we lack all experimental data, the only applicable lower bound in (5) is 0 and the only applicable upper bound is $P(t, y) + P(c, y')$:

$$0 \leqslant P(\text{benefit}) \leqslant P(t, y) + P(c, y'). \tag{8}$$

Applying these observational data only bounds to males and females yields:

$$\begin{aligned} 0 \leqslant P(\text{benefit}|\text{female}) &\leqslant 0.279, \\ 0 \leqslant P(\text{benefit}|\text{male}) &\leqslant 0.58. \end{aligned}$$

Naturally, these are far more loose than the point estimates when combined experimental and observational data are fully available. Let's similarly examine what can be computed with purely experimental data. Without observational data, only the first two arguments to max of the lower bound and min of the upper bound of $P(\text{benefit})$ in (5) are applicable:

$$\max\{0, P(y_t) - P(y_c)\} \leqslant P(\text{benefit}) \leqslant \min\{P(y_t), P(y'_c)\}. \tag{9}$$

Applying these bounds (using only experimental data) to males and females yields:

$$\begin{aligned} 0.279 \leqslant P(\text{benefit}|\text{female}) &\leqslant 0.489, \\ 0.28 \leqslant P(\text{benefit}|\text{male}) &\leqslant 0.49. \end{aligned}$$

Again, these are fairly loose bounds, especially when compared to the point estimates obtained with combined data. Notice that the overlap between the female bounds using observational data, $0 \leqslant P(\text{benefit}|\text{female}) \leqslant 0.279$, and the female bounds using experimental data, $0.279 \leqslant P(\text{benefit}|\text{female}) \leqslant 0.489$ is the point estimate $P(\text{benefit}|\text{female}) = 0.279$. The more comprehensive Tian–Pearl bounds formula (5) was not necessary. However, the intersection of the male bounds using observational data, $0 \leqslant P(\text{benefit}|\text{male}) \leqslant 0.58$, and the male bounds using experimental data, $0.28 \leqslant P(\text{benefit}|\text{male}) \leqslant 0.49$, does not provide us with narrower bounds. For males, the comprehensive Tian–Pearl bounds in (5) were necessary for narrower bounds (in this case, a point estimate).

Having seen this mechanism of combining observational and experimental data in (5) work so well, the reader may ask what is behind this? The intuition comes from the fact that observational data incorporate individuals' whims, and whims are proxies for hidden factors that may affect that individual's response to treatments. Such "confounding" factors are usually problematic in causal inference, since they lead to biased conclusions, sometimes completely reversing a treatment's effect [11]. Confounding then needs to be adjusted for. However, here confounding helps us, exposing the underlying mechanisms its associated whims and desires are a proxy for.

Finally, as noted in Section 3, knowing the relative sizes of the benefiting vs harmed subpopulations demands investment in finding mechanisms responsible for the differences as well as characterizations of those subpopulations. For example, women above a certain age might be affected differently by the drug, which could be detected by investigating how age affects the bounds on the individual response. Such characteristics can be potentially narrowed repeatedly until the drug's efficacy can be predicted for an individual with certainty or the underlying mechanisms of the drug can be fully understood.

None of this was possible with only the RCT. Yet, remarkably, an observational study, however sloppy and uncontrolled, provides a deeper perspective on a treatment's effectiveness. It incorporates individuals' whims and desires that govern behavior under free-choice settings. And, since such whims and desires are often proxies for factors that also affect outcomes and treatments (i.e., confounders), we gain additional insight hidden by RCTs.

5 Monotonicity, probability of harm, NNT, and other results

A natural question to ask at this point is, under what condition will RCT results constitute a point estimate for our target quantity, $P(\text{benefit})$? Pearl [12] has shown that this occurs under a condition called *monotonicity*, namely, when the treatment cannot harm any individual, formally

$$P(y_t', y_c) = P(\text{harm}) = 0.$$

This can be shown through a general relationship between $P(\text{harm})$, $P(\text{benefit})$, and ATE, which reads[13]:

$$P(\text{harm}) = P(\text{benefit}) - \text{ATE}. \tag{10}$$

Equation (10) can serve two purposes. First, it tells us immediately that under monotonicity (i.e., $P(\text{harm}) = 0$), $P(\text{benefit})$ coincides with ATE, or, in other words, ATE constitutes a point estimate of $P(\text{benefit})$. Second, it allows us to compute $P(\text{harm})$ from $P(\text{benefit})$ and ATE in cases where monotonicity does not hold, as was the case for men in the numeric example of Section 3.

For each of females and males, in the aforementioned example, their respective $P(\text{benefit})$ and ATE are known. Therefore, their probabilities of harm are known as well:

$$\begin{aligned} P(\text{harm}|\text{female}) &= P(\text{benefit}|\text{female}) - \text{CATE}(\text{female}) \\ &= 0.279 - 0.279 = 0, \\ P(\text{harm}|\text{male}) &= P(\text{benefit}|\text{male}) - \text{CATE}(\text{male}) \\ &= 0.49 - 0.28 = 0.21. \end{aligned}$$

Another concept that has become popular among trialists is NNT,[14] which is defined as follows: “The number of persons needed to be treated, on average, to prevent one more event (e.g., occurrence of a disease to be prevented, complication, adverse reaction, relapse)” [14]. Unfortunately, generations of trialists have failed to notice the counterfactual nature of the verb “prevent” and have estimated NNT as the inverse of ATE [15]:

$$\text{NNT} = \frac{1}{P(y_t) - P(y_c)} \tag{11}$$

instead of the inverse of $P(\text{benefit})$:

$$\text{NNT} = \frac{1}{P(\text{benefit})}. \tag{12}$$

Equation (11) has been used indiscriminately, including cases where treatment may cause harm to some individuals. In such cases, NNT should be estimated as bounds, specifically the inverse of equations (5), (8), and (9). For example, if only experimental data are available, equation (11) merely provides an upper bound:

$$\max\left\{\frac{1}{P(y_t)}, \frac{1}{P(y_c')}\right\} \leqslant \text{NNT} \leqslant \frac{1}{P(y_t) - P(y_c)}, \tag{13}$$

[13] Equation (10) can be obtained by expanding ATE, subtracting $P(y_c) = P(y_c, y_t) + P(y_c, y_t')$ from $P(y_t) = P(y_t, y_c) + P(y_t, y_c')$ to obtain $\text{ATE} = P(y_t, y_c') - P(y_t', y_c) = P(\text{benefit}) - P(\text{harm})$.

[14] NNT is not without controversy. Issues revolve around cases where confidence intervals for ATE include 0 rendering NNT undefined. If the ATE is in fact 0, then the explanatory benefit of NNT can easily obtain lost. Stovitz and Shrier show why baseline risk is important for medical decision making if NNT is relied upon [13].

and the lower bound is provided by equation (9). This assumes ATE is positive, otherwise the upper bound is infinite.

Given its ubiquity in interpreting experimental studies, a natural question to ask is whether monotonicity is testable. This question can be answered by examining the bounds on $P(\text{harm})$ and asking what conditions would guarantee an upper bound of 0. The bounds on the probability of harm are as follows:

$$\max\left\{\begin{array}{c} 0, \\ P(y_c) - P(y_t), \\ P(y) - P(y_t), \\ P(y_c) - P(y) \end{array}\right\} \leqslant P(\text{harm}) \leqslant \min\left\{\begin{array}{c} P(y_c), \\ P(y'_t), \\ P(t,y') + P(c,y), \\ P(y_c) - P(y_t) + \\ P(t,y) + P(c,y') \end{array}\right\}. \tag{14}$$

We see that, when $P(y_t) \leqslant P(y_c)$, the sufficient test demands that any of the following pathological conditions be true:

$$P(y_c) = 0 \text{ or} \tag{15}$$

$$P(y_t) = 1 \text{ or} \tag{16}$$

$$P(t,y') = P(c,y) = 0 \text{ or} \tag{17}$$

$$P(y_t) - P(y_c) = P(t,y) + P(c,y'). \tag{18}$$

The necessary test for monotonicity is more informative and is given in causality [1, p. 294]:

$$P(y_t) \geqslant P(y) \geqslant P(y_c). \tag{19}$$

This test is useful for two reasons. First, it can quickly eliminate the possibility of monotonicity by checking for a violation of (19). Second, such a violation indicates a high variability among individuals in the subpopulation considered, which, in turn, calls for a search for the mechanism responsible for the variability.

6 Annotated bibliography for related works

The following is a list of papers that analyze probabilities of causation and lead to the results reported above.

- Chapter 9 of causality [1] derives bounds on individual-level probabilities of causation and discusses their ramifications in legal settings. It also demonstrates how the bounds collapse to point estimates under certain combinations of observational and experimental data.
- Tian and Pearl [5] develop bounds on individual-level causation by combining data from experimental and observational studies. This includes probability of sufficiency, probability of necessity, and PNS. PNS is equivalent to $P(\text{benefit})$. $\text{PNS}(u) = P(\text{benefit}|u)$, the probability that individual $U = u$ survives if treated and does not survive if not treated, is related to $\text{ITE}(u)$ (1) via the following equation:

$$\text{PNS}(u) = P(\text{ITE}(u') > 0 | C(u') = C(u)). \tag{20}$$

In words, $\text{PNS}(u)$ equals the proportion of units u' sharing the characteristics of u that would positively benefit from the treatment. The reason is as follows. Recall that (for binary variables) $\text{ITE}(u)$ is 1 when the individual benefits from the treatment, $\text{ITE}(u)$ is 0 when the individual responds the same to either treatment, and $\text{ITE}(u)$ is –1 when the individual is harmed by treatment. Thus, for any given population, $\text{PNS} = P(\text{ITE}(u) > 0)$. Focusing on the subpopulation of individuals u' that share the characteristics of u, $C(u') = C(u)$, we obtain (20). In words, $\text{PNS}(u)$ is the fraction of indistinguishable individuals that would benefit from treatment. Note that although (2) can be estimated by controlled experiments over the

population $C(u') = C(u)$, (20) is defined counterfactually, and hence, it cannot be estimated solely by such experiments; it requires additional ingredients as described in the aforementioned text.
- Mueller and Pearl [4] provide an interactive visualization of individual-level causation, allowing readers to observe the dynamics of the bounds as one changes the available data.
- Li and Pearl [3] optimize societal benefit of selecting a unit u, when provided costs associated with the four different types of individuals: benefiting, harmed, always surviving, and doomed.
- Mueller et al. [16] take into account the causal graph to obtain narrower bounds on PNS. The hypothetical study in this article was able to calculate point estimates of PNS, but often the best we can obtain are bounds.
- Pearl [17] demonstrates how combining observational and experimental data can be informative for determining causes of effects (CoE), namely, assessing the probability PN that one event was a necessary cause of an observed outcome.
- Dawid and Musio [18] analyze CoE, defined by PN, the probability that a given intervention is a necessary cause for an observed outcome. Dawid and Musio further analyze whether bounds on PN can be narrowed with data on mediators.
- Both Bareinboim and Pearl [19] and Lee et al. [20] combine different heterogeneous data sources to estimate causal effects. These are population-level causal effects, in contrast to individual-level causal effects discussed in this article.

7 Conclusion

One of the least disputed mantra of causal inference is that we cannot access individual causal effects; we can observe an individual response to treatment or to no-treatment but never to both. Instead, most causal inference research has focused on the ATE, a population quantity that is estimable directly from RCTs, but provides no information on how individuals who respond one way under treatment would respond under an alternative. Our theoretical results show that we can go beyond ATE, to estimate (or bound) the entire distribution of individual causal effects. The bounds estimated can be quite narrow and allow us to make accurate personalized decisions. In other words, a randomly chosen individual would be able to assess how she would respond both to treatment and to its negation and, accordingly, decide on a course of action that best fits both her personal preferences and societal needs.

We showed that the key to getting accurate information on individual causal effects lies in combining observational and experimental data. While the mathematics of this combination was developed two decades ago [5,12], this article explains the mechanism behind it and demonstrates its implications in decision making situations. We have shown that observational data reveal individual idiosyncrasies that are masked in experimental settings. In other words, the unobserved confounding factors that usually affect both treatments and outcomes in observational studies are proxies for individual idiosyncrasies, often emanating from unique experience, beliefs, or desires that also govern ITEs.

Having established the mechanism of combining observational and experimental data, we have demonstrated the added value of observational studies in situations where experimental data are available. We have seen, for example, that gaining access to observational data may flip individual decisions from treatment to no-treatment and *vice versa*. From a policy-making viewpoint, information obtained from observational studies may reveal significant heterogeneity among subpopulations (e.g., males vs females) that are totally indistinguishable in experimental settings. Such information can identify which subpopulations are most susceptible to harmful treatment effects and thus assist in differential policy making. Finally, identifying susceptible subpopulations opens the possibility of searching for the mechanisms responsible for their differences as well as identifying new predictive markers associated with those differences.

We should emphasize at this point that, while these findings illuminate individual responses to treatment and may help individual decision making, they apply equally to all individuals sharing u's measured characteristics $C(u)$ (footnote 2). Conditioning on additional characteristics of the individual involved should provide, of course, additional person-specific information. However, such additions are accompanied with increased variance and must therefore be limited by the sample size available in each stratum. Our bounds are not subject to this limitation and takes full advantage of the large sample size usually available in observational studies. Therefore, we project that these methods provide the key for next-generation personalized decision making.

Acknowledgments: We appreciate the useful comments of Melisa Galvan.

Funding information: This research was supported in parts by grants from the National Science Foundation (#IIS-2106908 and #IIS-2231798), Office of Naval Research (#N00014-21-1-2351), and Toyota Research Institute of North America (#PO-000897).

Conflict of interest: Prof. Judea Pearl is one of the Editors in the Journal of Causal Inference but was not involved in the review process of this article.

References

[1] Pearl J. Causality. 2nd ed. New York: Cambridge University Press; 2009.

[2] Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol. 1986;15(3):413–9. doi: 10.1093/ije/15.3.413.

[3] Li A, Pearl J. Unit selection based on counterfactual logic. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. (IJCAI-19), International Joint Conferences on Artificial Intelligence Organization, 2019. p. 1793–9. https://doi.org/10.24963/ijcai.2019/248.

[4] Mueller S, Pearl J. Which patients are in greater need: a counterfactual analysis with reflections on COVID-19; 2020. https://ucla.in/39Ey8sU.

[5] Tian J, Pearl J. Probabilities of causation: bounds and identification. Ann Math Artif Intell. 2000;28(1–4):287–313. http://ftp.cs.ucla.edu/pub/stat_ser/r271-A.pdf.

[6] Pearl J. On the consistency rule in causal inference: an axiom, definition, assumption, or a theorem? Epidemiology. 2010;21(6):872–5. https://ftp.cs.ucla.edu/pub/stat_ser/r358-reprint.pdf.

[7] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educat Psychol. 1974;66(5):688–701.

[8] Bareinboim E, Tian J, Pearl J. Recovering from selection bias in causal and statistical inference. Proceedings of the Twenty-eighth AAAI Conference on Artificial Intelligence. 2014. p. 2410–6. http://ftp.cs.ucla.edu/pub/stat_ser/r425.pdf.

[9] Huang EJ, Fang EX, Hanley DF, Rosenblum M. Constructing a confidence interval for the fraction who benefit from treatment, using randomized trial data. Biometrics. 2019;75(4):1228–39. https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13101.

[10] Li A, Pearl J. Probabilities of causation with nonbinary treatment and effect. Technical Report R-516. Los Angeles, CA: University of California Los Angeles; 2022. http://ftp.cs.ucla.edu/pub/stat_ser/r516.pdf.

[11] Pearl J. Understanding Simpson's paradox. Amer Stat. 2014;68(1):8–13. http://ftp.cs.ucla.edu/pub/stat_ser/r414-reprint.pdf.

[12] Pearl J. Probabilities of causation: Three counterfactual interpretations and their identification. Synthese. 1999;121:93–149. https://ftp.cs.ucla.edu/pub/stat_ser/r260-reprint.pdf.

[13] Stovitz SD, Shrier I. Medical decision making and the importance of baseline risk. British J General Practice. 2013;63(616):e795–7. https://bjgp.org/content/63/616/e795.

[14] Porta M. Number needed to treat (NNT). Oxford University Press; 2016. https://www.oxfordreference.com/view/10.1093/acref/9780199976720.001.0001/acref-9780199976720-e-1327.

[15] Vancak V, Goldberg Y, Levine S. Systematic analysis of the number needed to treat. Stat Meth Med Res. 2020;29(9):2393–410. PMID: 31906795. https://doi.org/10.1177/0962280219890635.

[16] Mueller S, Li A, Pearl J. Causes of effects: learning individual responses from population data. 2022. http://ftp.cs.ucla.edu/pub/stat_ser/r505.pdf.

[17] Pearl J. Causes of effects and effects of causes. J Sociologic Meth Res. 2015;44(1):149–64. http://ftp.cs.ucla.edu/pub/stat_ser/r431-reprint.pdf.

[18] Dawid AP, Musio M. Effects of causes and causes of effects. Annu Rev Stat Appl. 2022;9:261–287. https://doi.org/10.1146/annurev-statistics-070121-061120.

[19] Bareinboim E, Pearl J. Causal inference and the data-fusion problem. Proc National Academy Sci. 2016;113(27):7345–52. https://www.pnas.org/doi/10.1073/pnas.1510507113.

[20] Lee S, Correa JD, Bareinboim E. General identifiability with arbitrary surrogate experiments. In: Proceedings of The 35th Uncertainty in Artificial Intelligence Conference. PMLR; 2020. p. 389–98. ISSN: 2640-3498. https://proceedings.mlr.press/v115/lee20b.html.