Forthcoming, American Journal of Epidemiology.

TECHNICAL REPORT
R-530
August 2023

Perspective on ‘Harm’ in Personalized Medicine – An Alternative Perspective

Scott Mueller[1] and Judea Pearl[1]

[1]Cognitive Systems Laboratory, Department of Computer Science,
University of California, Los Angeles
[1]{scott, judea}@cs.ucla.edu

August 2023

Abstract

This commentary examines an article by Sarvet and Stensrud (SS), in which they discuss the concept of ‘harm’ and its application in medical practice. SS advocate for an intervention-based interpretation of harm, downplaying its counterfactual interpretation. We take issue with this stance. We show that the counterfactual approach is vital for effective decision-making policies and that neglecting it might lead to flawed decisions. In response to SS’s contention that “when the outcome is death and a counterfactual approach is used . . . more people will die,” we demonstrate how counterfactual reasoning can actually prevent deaths. Additionally, we highlight the advantages of counterfactual thinking in the fields of medical malpractice, legal reasoning, and general diagnoses. Relying solely on intervention-based analyses limits our ability to accurately represent reality and hinders productive discussions about evidence, assumptions, and consensus building.

1 Introduction

Sarvet and Stensrud (SS)’s article, “Perspective on ‘Harm’ in Personalized Medicine” [Sarvet and Stensrud, 2023], offers a comprehensive examination of the concept of ‘harm’ and how it is used and abused in the practice of medicine. The authors make clear that the concept of harm as intended by the Hippocratic Oath, and subsequently by generations of ethics-minded professionals, has two faces. The first, which the authors call *interventional*, concerns treatments that put an individual at a higher risk of an adverse outcome, usually death. The second, which the authors call *counterfactual*, concerns treatments that would kill an individual who would otherwise be alive. Operationally, interventional decisions are based on estimates of how two subpopulations, each resembling the individual in question, would react, each subject to a different treatment. In contrast, counterfactual-based decisions are based on estimates of how one subpopulation resembling the individual would react to one treatment, given its reaction to another. Since the two reactions cannot be observed simultaneously, joint reactions are reconstructed with the help of logic, in the same way that 3-dimensional objects are reconstructed from their 2-dimensional projections using the laws of optics.

The qualms that we have about the paper in question are directed toward the unequivocal, potentially exclusive, preference that the authors express towards the interventionist approach vis-à-vis the counterfactual alternative. Although the authors do not advocate a total banishment of the counterfactual method from consideration, they insinuate its unspoken exclusion across various sections of their work. In Section 6, for example, they write, "a serious problem [with counterfactuals] is that we have no direct evidence that these principal strata exist." They further warn us, "when policy-makers optimize *counterfactual* utilities, then, in general, more people will die." What we strongly object to is the conclusion implied by SS that a rational decision maker may well apply the interventional perspective to the exclusion of counterfactual considerations.

2 Metaphysical Trepidations

Let us start with the metaphysical trepidations that SS express vis-à-vis the very notion of counterfactual, as expressed by the phrase, "treatments that would kill an individual who would otherwise be alive." According to SS, we have no direct evidence that such treatments (or an individual) exist. Moreover, "when a *counterfactual* framework is deployed to determine social policies and regulations, it coerces conformity to an unverifiable metaphysics and a corresponding logic that deals in those terms." These trepidations are unjustified. In his commentary on [Dawid, 2000], Pearl likens counterfactual logic to the use of imaginary numbers in mathematics and engineering; they seem "nonexisting" or "metaphysical" in standard algebra, but can be rigorously defined in the algebra of pairs (i.e., complex plane) and, when properly understood, become indispensable even in the analysis of real quantities. Indeed, in contrast to SS's apocalyptic warning, our paper [Mueller and Pearl, 2023] demonstrates that counterfactual considerations are necessary in social policy and regulations, without which critical decisions could misfire. This is explicated in the following section.

3 Policy Decisions and Utility

In a recent paper Mueller and Pearl, 2023, we illustrate an example of a treatment that diminishes the death rate by 30 percentage points, from 80% to 50%, equally in both men and women. Yet, while interventional data makes men and women appear totally indistinguishable, counterfactual analysis reveals a significant disparity: The 30 percentage point reduction in women's deaths is comprised entirely of women who were cured by the treatment and would have died if left untreated, while the 30 percentage point reduction in men's deaths is split between 50% who are cured and 20% who are killed by the treatment–referring to patients who would have survived if untreated but died due to adverse reactions to the treatment.

These conclusions are not metaphysical but logically derivable from the available data (assuming that the treatment and outcomes are binary and that the system is deterministic[1], hence every individual must fall within one of the four possible response types, or principle strata ($S \in \{1, 2, 3, 4\}$) as defined by SS).

To show, in the example above, that counterfactual considerations are critical for policy making, let's assume that a post-mortem autopsy can identify the cause of death and that families of patients who die due to the treatment are likely to sue the hospital for negligence. If such lawsuits become public, they could severely undermine public trust in all medical services provided by hospitals. In such circumstances, it would be entirely ethical for decision-makers to suspend treating male patients, even though this suspension raises the risk of death from 50% to 80%.

In fact, it is doubtful whether such a treatment, once found to be fatal in some patients, would be approved by regulating agencies, despite its potential to reduce the average death rate by 30 percentage points. It is more probable that such a treatment would be declared unsafe and returned to the developer for re-evaluation or refinement. Regrettably, this capacity to cause death would not be exposed until autopsies are conducted, and lawsuits are initiated; the experimental data alone can only deduce that men and women are indistinguishable and that there's a 30% reduction in fatalities in each group. This is precisely where counterfactual analysis comes in; it can detect and quantify the harm caused by the treatment at the study itself, before any autopsies are performed.

To demonstrate how counterfactual analysis can reduce deaths compared with interventional analysis, let's consider the female patients in the same example. Upon learning that the treatment might be unsafe for some patients, a prudent policy would be to suspend treating all patients until the causes for fatal reactions are identified. This implies denying women the benefit of a treatment that could prevent 30% of them from death. By leveraging counterfactual analysis, on the other hand, and based solely on the study data, we can be assured that the treatment is entirely safe for women, thus sanctioning the treatment for women and potentially saving 30% of them.

In conclusion, we assert that counterfactual considerations can save lives on both fronts; it cautions against treating those who could be harmed (men, in

[1] The deterministic assumption was contested by Dawid [Dawid, 2000] and defended in [Pearl, 2000]. Dawid's contention emanates from the observation that the response of each individual may vary with unknown factors (e.g. time of day, previous history, patient's mood, etc) and cannot, therefore, be a deterministic function of the treatment. However, if we include those factors in the definition of a *unit*, determinism regains its legitimacy (barring quantum uncertainties).

this case) and provides the green light to treat those who are safe from harm (women, in this instance). Note that if there had been an alternative treatment that men would respond to as safely as women, our preference would definitely lean towards this second treatment over the one described in our example. Paradoxically, the interventional framework would judge the two treatments as equally efficacious, merely because the death counts for both treatments are the same.

Some may argue that probability of death should indeed be the only criterion they should abide by, regardless of whether the treatment was actually helpful. This attitude, in our opinion, is biased by the severity and cultural fear of death as the ultimate source of anxiety. Imagine a scenario where the outcome of interest is merely a headache, and that the treatment, which is generally efficacious, may escalate the agony twofold in some individuals. In such circumstances, many might choose the path of avoidance, despite the overall positive Average Treatment Effect (ATE).

4 Counterfactual Harm as Early Warning Sensor

An additional benefit to counterfactual harm analysis is its potential in uncovering the mechanism behind treatment effects. As shown in the male and female example, the detection of counterfactual harm can serve as an early warning signal for a hitherto unknown disorder in certain subpopulations. This can be used by medical researchers to launch a rigorous and systematic search aimed at discerning the mechanisms underlying these disorders, as well as discovering tangible biomarkers that distinctly categorize individuals afflicted by such disorders.

5 The But-for Criterion in Legal Reasoning

Our last objection concerns the legal ‘but-for’ criterion. In the language of counterfactuals, it is formulated as Probability of Necessity (PN):

$$\text{PN} = P(Y^{a=0} = 0|a = 1, Y = 1). \tag{1}$$

An injury, $Y = 1$, that was observed under action $a = 1$ would not have occurred, but for the action taken, $Y^{a=0} = 0$. In many legal cases, it is specifically formulated in terms of high probability (“more probable than not”) as opposed to complete certainty, or $P(\text{harm}) = 1$ as stated in SS’s paper. If the criterion for medical malpractice liability was, in fact, $P(\text{harm}) = 1$, the bar would be so high that no doctor would ever lose a case[2]. More seriously, SS do not distinguish between PN and Probability of Necessity and Sufficiency (PNS), and take

[2] A possible oversight: SS argue that merely checking the basic interventional probability $P(Y^a = 1|L = l) = 1$ negates the need for counterfactuals. However, $P(Y^{a=0} = 1|L = l) = 1$ indicates death when no treatment is given, so no one could be *harmed* by treatment since all

the latter as the ultimate criterion for assigning legal responsibility. We refer to the latter as $P(\text{benefit})$:

$$\text{PNS} = P(\text{benefit}) = P(Y^{a=0} = 0, Y^{a=1} = 1). \tag{2}$$

Although PN and PNS are related (see Eq. (9.5) in [Pearl, 2009] and Eqs. (35) and (36) in [Tian and Pearl, 2000]), they are distinct counterfactual probabilities measuring different types of causal relations. It is easy to come up with an example where PNS is low and PN is high, or vice-versa. For instance, suppose the vast majority of people with an illness do not benefit from a particular medicine. However, the few individuals who do benefit always choose to take the medicine, perhaps because they realize somehow that it is in their best interest. In this case, PNS, the proportion of benefiters, is close to 0, while $\text{PN} = 1$.

Additional legal nuances, such as Probability of Sufficiency (PS) and Probability of Actual Cause, have been formulated in the language of counterfactuals and given algorithmic embodiment Pearl, 2009. In general, the important and ubiquitous challenge of assessing Causes of Effects, the degree to which one event is responsible for a later event known to have occurred, cannot be articulated without the language of counterfactuals. It is no wonder that legal language is laden with counterfactual terms. In the absence of such language, the legal profession will go back to the days of Hammurabi when doctors' hands were chopped off [Becker et al., 2018].

6 Summary

We argue that counterfactual logic should not be purged from consideration of harm and benefit as implied by SS. First, it is pivotal for policy-making in medical practice. We demonstrated its role in minimizing deaths in certain subpopulations, its role in early detection of potential disorders in certain patients, and its critical role in legal reasoning. We end by warning that purging counterfactual thinking is dangerously misguided. As outlined in this paper and in [Pearl, 2000, 2022; Pearl and Mackenzie, 2018], shunning counterfactuals would prevent researchers from using the most natural communication language that science has invented for modeling reality, discussing evidence, communicating assumptions, and reaching consensus.

Acknowledgments

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].

would have died without it. Likewise, $P(Y^{a=1} = 1 | L = l) = 1$ signals death after treatment, but it does not inform about the proportion harmed, as some may have died regardless of treatment.

References

Becker, C. L., Specter, S., & Kline, T. R. (2018, January 25). Chapter 16: The supreme court and medical malpractice law. In *Chapter 16: The supreme court and medical malpractice law* (pp. 241–258). Penn State University Press. https://doi.org/10.1515/9780271081991-019

Dawid, A. P. (2000). Causal inference without counterfactuals [Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.2000.10474210]. *Journal of the American Statistical Association*, *95*(450), 407–424. https://doi.org/10.1080/01621459.2000.10474210

Mueller, S., & Pearl, J. (2023). Personalized decision making – a conceptual introduction [Publisher: De Gruyter]. *Journal of Causal Inference*, *11*(1). https://doi.org/10.1515/jci-2022-0050

Pearl, J. (2000). Causal inference without counterfactuals: Comment [Publisher: [American Statistical Association, Taylor & Francis, Ltd.]]. *Journal of the American Statistical Association*, *95*(450), 428–431. https://doi.org/10.2307/2669380

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pearl, J. (2022). Causation and decision: On dawid's "decision theoretic foundation of statistical causality" [Publisher: De Gruyter]. *Journal of Causal Inference*, *10*(1), 221–226. https://doi.org/10.1515/jci-2022-0046

Pearl, J., & Mackenzie, D. (2018, May 15). *The book of why: The new science of cause and effect* (1st edition). Basic Books.

Sarvet, A. L., & Stensrud, M. J. (2023). Perspective on 'harm' in personalized medicine. *American Journal of Epidemiology*, kwad162. https://doi.org/10.1093/aje/kwad162

Tian, J., & Pearl, J. (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, *28*(1), 287–313. https://doi.org/10.1023/A:1018912507879