

Beyond Marginal Recovery: Weighing Benefit and Harm at the Bedside and in Policy

Shanshan Luo

Department of Applied Statistics, Beijing Technology and Business University

Martina Scauda

Statistical Laboratory, University of Cambridge

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

Abstract

We thank [Dawid and Senn \(2026\)](#) for clarifying why marginal interventional probabilities suffice when the objective is to maximize average recovery. We argue that the debate over counterfactuals in personalized decision-making persists because it conflates two levels of decision-making: individual bedside consultations and population policymaking. In a bedside consultation, a patient may be averse to treatment-caused harm relative to standard care, and the clinician’s role is to combine marginal study evidence with consultation-specific context to calibrate a subjective belief about joint potential outcomes. At the population level, a policymaker relies on marginal recovery rates conditional on measured study covariates, leaving counterfactual harm only partially identified. Using a three-arm extension of [Dawid and Senn \(2026\)](#), we characterize when marginal evidence suffices to uniquely determine an optimal treatment under harm aversion, when it yields a partial order that eliminates inferior options while leaving non-dominated treatments for clinical consultation, and how population harm budgets induce priority rules via lexical leapfrogging.

Keywords: individualized treatment rules, counterfactual belief, probability of harm, partial identification, shared decision-making, robust policy

1. Introduction

We thank [Dawid and Senn \(2026\)](#) for a lucid and incisive critique of counterfactual approaches to personalized decision-making. Against the proposal of [Mueller and Pearl \(2023\)](#) that treatment choice should be guided by the counterfactual probabilities of benefit (PB) and treatment-caused harm (PH), [Dawid and Senn \(2026\)](#) show that when the objective is to maximize the probability of recovery, marginal interventional probabilities suffice. Furthermore, they demonstrate that what [Mueller and Pearl \(2023\)](#) extract from combining experimental and observational data to narrow the bounds on PB and PH is simply the conditional outcome distribution given a latent “preferred treatment” covariate X^* , which can directly improve decisions within classical decision theory when observed prior to treatment ([Stensrud et al., 2024](#)). At the same time, [Dawid and Senn \(2026, Sections 4.5 and 5.4\)](#) and [Sarvet and Stensrud \(2025\)](#) argue that decision rules balancing PB against PH—such

as requiring $PB > \kappa PH$ under *primum non nocere* (“first, do no harm”)—are misguided because they reduce the overall recovery rate in a population.

We suggest that this controversy persists because the literature conflates two levels of decision-making—the *individual consultation* versus *population policymaking*—and two corresponding interpretations of probability:

1. In a bedside consultation (Section 3), a patient choosing between new drugs and standard care may be averse to treatment-caused harm (Mueller and Pearl, 2023; Ben-Michael et al., 2024) and hold *subjective beliefs* about how their own potential outcomes are coupled. The clinician’s role is to help calibrate those beliefs—drawing on marginal evidence from similar populations, uncodified patient-specific observations, and clinical experience—and to guide the patient through the benefit–harm trade-off toward an informed decision (not necessarily through a formal utility calculation).
2. A policymaker or regulator (Section 4), by contrast, designs guidelines for a heterogeneous population based on *population frequencies* estimated from clinical studies. Depending on the regulatory goal, the policymaker may optimize a simple utility that depends only on marginal distributions (such as expected recovery in the population), or adopt more sophisticated criteria that incorporate partial identification bounds on individual harm (Kallus, 2022; Ben-Michael et al., 2024; Wu et al., 2025; Scauda et al., 2026).

The remainder of this commentary is organized as follows. In Section 2, we extend the binary setup of Dawid and Senn (2026) to a choice between two new drugs and standard care, and formalize an asymmetric benefit–harm utility indexed by a harm-aversion parameter $\kappa \geq 0$. Section 3 uses this setting to illustrate an individual consultation in which a doctor helps a patient calibrate a subjective joint belief against empirical recovery rates before evaluating treatment choice across κ . Section 4 turns to the population perspective, using sharp Fréchet bounds to characterize the partial order of treatments across κ and connecting this dominance relation to harm-constrained policy learning and lexical leapfrogging. Section 5 concludes with implications for precision medicine and AI-assisted clinical decision-making at both the bedside and the policy level.

2. Weighing Benefit and Harm

We follow the notation of Dawid and Senn (2026), extending their binary treatment to three options: two new drugs and standard care. Let $X \in \{0, 1, 2\}$ denote the treatment received ($X = 0$ is standard care) and let $Y \in \{0, 1\}$ indicate recovery at a specified follow-up time ($Y = 1$ for recovery). Write $Y(x)$ for the potential outcome under treatment x and $X \leftarrow x$ for an intervention assigning x , assuming well-defined regimens, consistency ($Y = Y(X)$), and no interference across individuals.

To connect population evidence with an individual consultation, it is useful to distinguish two layers of pre-decision information available at the point of care:

1. *Measured study covariates* (L): Codified patient characteristics (such as age, sex, or standard biomarkers) that are recorded both for the target patient ($L = l$) and in existing randomized clinical trials or observational studies.

2. *Consultation-specific context (C)*: Additional clinical features observed during a consultation that are omitted from existing study databases. These could include structured variables (such as previous reactions to related medications, rare comorbidities, or the patient’s or clinician’s initial treatment preference X^*) as well as unstructured or uncodified information (such as biological mechanisms of action, clinical notes, emerging post-approval safety signals, or local care-delivery capabilities).

Empirical studies only identify quantities conditional on the measured covariates $L = l$. Under randomization or conditional exchangeability ($Y(x) \perp\!\!\!\perp X \mid L$) and positivity ($\Pr(X = x \mid L = l) > 0$), study data identify the marginal recovery probabilities

$$p_x(l) = \Pr\{Y(x) = 1 \mid L = l\} = \Pr(Y = 1 \mid L = l, X \leftarrow x), \quad x \in \{0, 1, 2\}.$$

For a policymaker, $p_x(l)$ represents an empirical population frequency among individuals with $L = l$; in a bedside consultation, $p_x(l)$ provides an empirical benchmark that the clinician and patient combine with further consultation-specific context to form their subjective beliefs and decisions.

Comparing any drug $x \in \{0, 1, 2\}$ with standard care 0, the individual treatment effect $\text{ITE}_x = Y(x) - Y(0)$ can take three values: +1 if drug x benefits the patient (recovery under x , but not under standard care), −1 if it causes harm (failure to recover under x when standard care would have succeeded), and 0 if the outcome is unchanged. The conditional average treatment effect, $\tau_x(l) = \text{CATE}_x(l) = p_x(l) - p_0(l)$ (with $\tau_0(l) = 0$), is simply the expected value of ITE_x given $L = l$. Ranking treatments by $\tau_x(l)$ is equivalent to ranking their recovery probabilities $p_x(l)$. However, $\tau_x(l) = \text{PB}_x(l) - \text{PH}_x(l)$ only captures the net difference between the probabilities of counterfactual benefit and harm (Mueller and Pearl, 2023; Dawid and Senn, 2026),

$$\text{PB}_x(l) = \Pr\{Y(x) = 1, Y(0) = 0 \mid L = l\}, \quad \text{PH}_x(l) = \Pr\{Y(x) = 0, Y(0) = 1 \mid L = l\},$$

which separately depend on the joint distribution of $(Y(x), Y(0))$ rather than the marginals alone.

When a decision maker distinguishes treatment-caused harm from untreated non-recovery, we can assign utility 1 to benefit, $-\kappa$ to harm, and 0 to no change, where $\kappa \geq 0$ is a harm-aversion parameter. This yields the expected benefit minus harm score (Richens et al., 2022; Ben-Michael et al., 2024)

$$V_\kappa(x \mid l) = \text{PB}_x(l) - \kappa \text{PH}_x(l) = \tau_x(l) - (\kappa - 1) \text{PH}_x(l), \quad x \in \{0, 1, 2\}, \quad (1)$$

with $V_\kappa(0 \mid l) = 0$. Because benefit and harm of each option are anchored to the common reference $X = 0$, $V_\kappa(x \mid l)$ assigns a scalar score to each treatment and induces a transitive ranking across multiple alternatives, avoiding the intransitivity that can arise from pairwise counterfactual-harm comparisons (Sawant and Stensrud, 2026). At $\kappa = 1$, benefit and harm receive equal weight and the score reduces to $\text{CATE}_x(l)$. Values $\kappa > 1$ reflect harm aversion relative to CATE (for example, $\kappa = 2$ penalizes harm twice as heavily as benefit), whereas $0 \leq \kappa < 1$ places less weight on harm. Varying κ thus provides a natural sensitivity analysis around the CATE rule.

3. Individual Decisions with Counterfactual Beliefs

Evidence-based medicine integrates external study evidence with individual clinical expertise and the patient’s own values (Sackett et al., 1996), a process operationalized in shared decision-making consultations (Elwyn et al., 2017). Consider a patient with recorded study covariates $L = l_1$, for whom study evidence gives marginal recovery probabilities

$$p_0(l_1) = 30\%, \quad p_1(l_1) = 60\%, \quad p_2(l_1) = 70\%.$$

Treatment 0 is standard supportive care, Treatment 1 is a milder, well-tolerated drug with moderate efficacy (60%), and Treatment 2 is a potent drug with high efficacy against the disease (70%) but notable toxicity.

During the consultation, the doctor observes consultation-specific context $C = c_1$ —for example, an uncomplicated baseline presentation that, in the doctor’s clinical experience, already characterizes about 85% of study patients with $L = l_1$. Focusing on this favourable presentation, the patient arrives expecting a 90% chance of recovery under Treatment 2. Even without counterfactuals, the doctor can explain that this expectation is incompatible with the study evidence ($0.85 \times 90\% = 76.5\% > 70\%$). Because $C = c_1$ represents the typical presentation in the study cohort, the doctor and patient agree after discussion to adopt the study benchmark (30%, 60%, 70%) as their calibrated working marginals given ($C = c_1, L = l_1$). Up to this point, no counterfactual dependence has been invoked.

With these working marginals, both active drugs outperform standard care and Treatment 2 has the highest recovery rate (70% vs. 60%). However, because recovery requires both resolving the disease and avoiding severe side effects, the patient worries that toxicity from Treatment 2 could prevent recovery when standard care or the milder Treatment 1 would have succeeded. Marginal recovery rates alone cannot resolve this concern because they do not determine how the responder groups overlap. Suppose clinical experience suggests that the well-tolerated Treatment 1 acts nearly monotonically—rarely harming patients who would recover under standard care—whereas the more potent Treatment 2 cures many more otherwise non-recovering patients at the cost of occasional toxicity, leading to the illustrative joint belief in Table 1.

Table 1 implies $(PB_1, PH_1) = (35\%, 5\%)$ and $(PB_2, PH_2) = (55\%, 15\%)$ (indeed, once the working marginals are fixed, evaluating (1) requires only the two harm probabilities PH_1 and PH_2 rather than all eight joint cells): Treatment 2 offers greater counterfactual benefit than Treatment 1, but three times the risk of treatment-caused harm relative to standard care (where the 10% excess harm corresponds to row (1, 1, 0), in which both Treatment 0 and Treatment 1 would have succeeded). The resulting benefit–harm scores (1),

$$V_\kappa(1 \mid c_1, l_1) = 0.35 - 0.05\kappa, \quad V_\kappa(2 \mid c_1, l_1) = 0.55 - 0.15\kappa, \quad V_\kappa(0 \mid c_1, l_1) = 0,$$

show that the patient prefers Treatment 2 for $0 \leq \kappa < 2$, switches to the safer Treatment 1 for $2 < \kappa < 7$ (including $\kappa = 3$), and prefers standard care 0 for $\kappa > 7$. Unlike the initial calibration of marginal recovery rates, this benefit–harm trade-off depends on a counterfactual joint belief.

$(Y(0), Y(1), Y(2))$	What would happen	$\Pr\{Y(0), Y(1), Y(2) \mid C = c_1, L = l_1\}$
(0, 0, 0)	No recovery with any treatment	15%
(0, 0, 1)	Recovery with Treatment 2 only	20%
(0, 1, 0)	Recovery with Treatment 1 only	0%
(0, 1, 1)	Recovery with Treatment 1 or 2, but not 0	35%
(1, 0, 0)	Recovery with Treatment 0 only	5%
(1, 0, 1)	Recovery with Treatment 0 or 2, but not 1	0%
(1, 1, 0)	Recovery with Treatment 0 or 1, but not 2	10%
(1, 1, 1)	Recovery with all three treatments	15%

Table 1: An illustrative joint belief conditional on the recorded covariates $L = l_1$ and consultation-specific context $C = c_1$, with working marginals (30%, 60%, 70%).

4. Policy Decisions Based on Population Evidence

4.1 Why counterfactual harm matters to a policymaker

Unlike a clinician and patient in a bedside consultation, a policymaker or guideline panel faces two additional hurdles. First, the policymaker cannot observe consultation-specific context C or rely on a subjective belief about how potential outcomes are coupled (as in Table 1); population guidelines must instead be justified using the marginal recovery rates $p_x(l) = \Pr\{Y(x) = 1 \mid L = l\}$ identified by clinical studies, leaving the joint distribution of $(Y(0), Y(1), Y(2))$ only partially identified. Second, whereas a bedside consultation tailors the decision to a single patient’s harm weight κ , a population guideline applies to many individuals whose attitudes toward harm vary and are not usually observed by the policymaker.

So why should a policymaker account for unrealized counterfactual outcomes that can never be directly observed? As Dawid and Senn (2026, Sections 4.5 and 5.4) observe, penalizing counterfactual harm under the clinical maxim *primum non nocere* (Mueller and Pearl, 2023; Ben-Michael et al., 2024) can reduce the average recovery rate relative to a CATE-maximizing rule. Nevertheless, two practical considerations motivate incorporating counterfactual harm into population guidelines:

1. *Aligning guidelines with public preferences:* Individuals affected by a policy often care asymmetrically about treatment-caused harm versus untreated non-recovery ($\kappa \neq 1$). A guideline based solely on CATE ($\kappa = 1$) may recommend an aggressive treatment that harm-averse patients would reject or fail to adhere to once informed of its potential to disrupt natural recovery.
2. *Downstream consequences of perceived harm:* Individuals who experience an adverse outcome after being switched to a new policy may suspect that they would have fared better under standard care or an alternative treatment. Even when this counterfactual is unobservable in the primary endpoint Y , perceived harm can trigger real

consequences elsewhere—such as prompting discontinuation of care, reducing treatment adherence, or eroding public trust in medical guidelines.

Because neither the true potential-outcome copula nor each individual’s exact harm weight κ is available at the population level, we view the policymaker’s primary role not as dictating a single best treatment on average, but as eliminating clearly inferior treatments while leaving non-dominated options open for clinical consultations (Manski, 2019).

4.2 Partial ordering of treatments under harm aversion

To formalize which treatments can be ruled out from marginal evidence alone, fix $L = l$ and consider $\kappa \geq 1$ to simplify the exposition. For each active drug $x \in \{1, 2\}$, the binary Fréchet bounds restrict the harm probability to (Dawid and Senn, 2026, Section 4.1.1)

$$L_x(l) := \max\{0, p_0(l) - p_x(l)\} \leq \text{PH}_x(l) \leq U_x(l) := \min\{p_0(l), 1 - p_x(l)\},$$

with $L_0(l) = U_0(l) = 0$. Because $V_\kappa(x \mid l) = \tau_x(l) - (\kappa - 1)\text{PH}_x(l)$ is non-increasing in $\text{PH}_x(l)$ for $\kappa \geq 1$, its sharp lower and upper bounds are obtained simply by plugging in $\text{PH}_x(l) = U_x(l)$ and $\text{PH}_x(l) = L_x(l)$, respectively (with $V_\kappa(0 \mid l) = 0$). Moreover, because benefit and harm are both defined relative to $Y(0)$, every pair $(\text{PH}_1(l), \text{PH}_2(l)) \in [L_1(l), U_1(l)] \times [L_2(l), U_2(l)]$ is jointly realizable (Appendix A). Thus treatment x *robustly dominates* x' across all compatible joint distributions, written $x \succ_\kappa x'$, if and only if the lower bound of $V_\kappa(x \mid l)$ exceeds the upper bound of $V_\kappa(x' \mid l)$.

Proposition 1 (Robust dominance under harm aversion) *Fix $L = l$ with $p_2(l) \geq p_1(l) \geq p_0(l)$ and consider $\kappa \geq 1$. For any $x, x' \in \{0, 1, 2\}$, $x \succ_\kappa x'$ requires $p_x(l) > p_{x'}(l)$ and holds if and only if*

$$1 \leq \kappa < 1 + \frac{p_x(l) - p_{x'}(l)}{U_x(l)}.$$

Proposition 1 has a transparent interpretation: x robustly dominates x' whenever the excess harm weight $\kappa - 1$ is smaller than the CATE advantage $p_x(l) - p_{x'}(l)$ divided by the worst-case harm $U_x(l)$ of treatment x (interpreting a positive numerator divided by zero as $+\infty$). A treatment x' is *eliminated* if $x \succ_\kappa x'$ for some alternative x , and x is *uniquely optimal* if it dominates all alternatives.

Table 2 contrasts the population dominance relation for $L = l_1$ with the individual decision in Section 3. As harm aversion increases, marginal evidence progresses from singling out the most efficacious drug, to eliminating only standard care while leaving the two active drugs for bedside comparison, and finally to leaving all three options open. Because the patient’s calibrated belief in Table 1 assigns much lower harm rates than the worst case, the individual consultation resolves the choice between the active drugs in the intermediate regime ($2 < \kappa < 7/3$, with a tie at $\kappa = 2$) and continues to favour the milder Treatment 1 even after the population bounds cease to establish dominance ($7/3 \leq \kappa < 7$).

4.3 Harm-constrained policy optimization

Instead of fixing a preference weight κ within a single stratum $L = l$, a policymaker can specify a population-wide harm budget $B > 0$ across strata, from which an effective harm threshold emerges as a “shadow price” (Wu et al., 2025; Scauda et al., 2026).

Range of κ	Partial order $\succ_\kappa$	Eliminated	Undominated set	Choice under Table 1
$[1, 4/3)$	$2 \succ_\kappa 1 \succ_\kappa 0$	$\{0, 1\}$	$\{2\}$	2
$[4/3, 2)$	$\{1, 2\} \succ_\kappa 0$	$\{0\}$	$\{1, 2\}$	2
$[2, 7/3)$	$2 \succ_\kappa 0$	$\{0\}$	$\{1, 2\}$	1 ($\{1, 2\}$ at $\kappa = 2$)
$[7/3, 7)$	$\emptyset$	$\emptyset$	$\{0, 1, 2\}$	1
$[7, \infty)$	$\emptyset$	$\emptyset$	$\{0, 1, 2\}$	0 ($\{0, 1\}$ at $\kappa = 7$)

Table 2: Population partial orders and undominated sets from marginals alone $((p_0, p_1, p_2) = (30\%, 60\%, 70\%), U_1 = U_2 = 30\%)$ versus the individual optimal choice under the joint belief in Table 1.

For a deterministic binary treatment rule $\pi(L) \in \{0, 1\}$ against standard care 0, [Wu et al. \(2025\)](#) maximize expected post-policy recovery subject to a cap on worst-case counterfactual harm:

$$\max_{\pi} \mathbb{E}\{\pi(L)Y(1) + (1 - \pi(L))Y(0)\} \quad \text{subject to} \quad \sup_{\gamma \in \Gamma} \mathbb{E}_{\gamma}\{\pi(L)\text{PH}_1(L)\} \leq B,$$

where Γ is either the Fréchet class (which places no constraint on the dependence between potential outcomes beyond matching the marginals) or a subclass bounding $\text{Corr}(Y(0), Y(1) \mid L)$ from below. Assuming $V_{\kappa^*}(1 \mid L)$ has a continuous distribution so that the boundary $\{l: V_{\kappa^*}(1 \mid l) = 0\}$ has probability zero, Lagrangian duality yields the threshold rule $\pi^*(l) = \mathbb{I}\{V_{\kappa^*}(1 \mid l) > 0\}$ evaluated at the worst-case harm rate in Γ . Here $\kappa^* \geq 1$ plays the same role as the patient’s harm-aversion weight κ in Section 3, except that the excess harm weight $\kappa^* - 1 \geq 0$ now emerges as the shadow price (Lagrange multiplier) of the population budget B rather than an individual preference. Under the Fréchet class, this rule assigns treatment 1 precisely when $1 \succ_{\kappa^*} 0$ in Proposition 1.

[Scauda et al. \(2026\)](#) generalize this framework in several directions: they allow multi-valued treatments, general outcomes and harm measures, policy transitions from an incumbent treatment X to a new assignment X' (measuring harm relative to the treatment X each patient would otherwise receive), and randomized transition probabilities $\pi(x' \mid x, l) = \Pr(X' = x' \mid X = x, L = l) \in [0, 1]$. When the marginal outcome distributions given $L = l$ are ordered by stochastic dominance (which holds automatically for binary outcomes) and Γ is a compatible ambiguity set (such as the Fréchet class or pairwise odds-ratio lower bounds), [Scauda et al. \(2026\)](#) show that the optimal transition rule has a *lexical leapfrogging* structure:

1. *Leapfrogging*: Any transition out of a suboptimal incumbent treatment $x \neq x^*(l)$ skips intermediate options and moves patients directly to the CATE-optimal treatment $x^*(l) = \arg \max_{x'} p_{x'}(l)$.
2. *Lexical ordering*: Subgroups (x, l) are transitioned in decreasing order of their priority score $\Lambda(x, l)$ —defined as the CATE gain $p_{x^*(l)}(l) - p_x(l)$ divided by the worst-case probability of harm from switching $x \rightarrow x^*(l)$ —which does not depend on the status-quo distribution of X .

Stratum (p_0, p_1, p_2)	Switch	CATE		Fréchet class			Corr ≥ 0 (OR ≥ 1)		
		Gain	Rank	Max harm	$\Lambda(x, l)$	Rank	Max harm	$\Lambda(x, l)$	Rank
$l_1: (30\%, 60\%, 70\%)$	$0 \rightarrow 2$	40%	1	30%	4/3	3	9%	40/9	2
	$1 \rightarrow 2$	10%	4	30%	1/3	4	18%	5/9	4
$l_2: (60\%, 70\%, 90\%)$	$0 \rightarrow 2$	30%	2	10%	3	1	6%	5	1
	$1 \rightarrow 2$	20%	3	10%	2	2	7%	20/7	3

Table 3: Lexical leapfrogging across two strata $L \in \{l_1, l_2\}$ with $x^*(l) = 2$. Intermediate switches $0 \rightarrow 1$ are bypassed by leapfrogging. Worst-case harm for $x \rightarrow 2$ is $\min\{p_x(l), 1 - p_2(l)\}$ under the Fréchet class and $p_x(l)\{1 - p_2(l)\}$ under non-negative correlation.

Table 3 illustrates this structure across our stratum $L = l_1$ and a second stratum $L = l_2$ with higher baseline recovery (60%, 70%, 90%). By leapfrogging, standard-care patients in both strata skip Treatment 1 and transition directly to Treatment 2. Within l_1 , where $1 - p_2(l_1) \leq p_0(l_1)$ so worst-case harm for Treatment 2 is 30% relative to both Treatment 0 and Treatment 1, adding 1 to the Fréchet priority scores gives $1 + \Lambda(0, l_1) = 7/3$ and $1 + \Lambda(1, l_1) = 4/3$, which match the exact κ cutoffs in Table 2 below which $2 \succ_\kappa 0$ and $2 \succ_\kappa 1$ hold. Across strata, however, CATE, the Fréchet class, and non-negative correlation produce three distinct priority orderings: both harm-aware criteria rank $(0, l_2)$ first because $p_2(l_2) = 90\%$ caps worst-case harm at 10%, while assuming non-negative correlation flips $(0, l_1)$ ahead of $(1, l_2)$ because low baseline recovery in $(0, l_1)$ ($p_0 = 30\%$) reduces worst-case harm from 30% to 9%.

5. Discussion

Much of the statistical literature on precision medicine has focused on learning optimal individualized treatment rules from covariates L measured in randomized trials or observational studies, a task that typically reduces to estimating $\text{CATE}_x(l)$ or its sign (Kosorok and Laber, 2019). As Dawid and Senn (2026) rightly emphasize, marginal interventional probabilities suffice for this objective. In clinical practice, however, several candidate treatments may have close $\text{CATE}_x(l)$ yet very different safety profiles, and clinicians routinely observe rich consultation-specific context C that is absent from study databases. This commentary has explored how counterfactual considerations help navigate such intricate scenarios at both the bedside and the policy level.

These considerations are increasingly relevant as artificial intelligence (AI)—broadly understood as automated prediction and decision-support algorithms—enters clinical practice (Goh et al., 2025). At the bedside, an algorithm trained on codified study covariates L can estimate marginal recovery rates $p_x(l)$, while large language models may further assist clinicians in synthesizing medical histories and mechanistic evidence to assess consultation-specific context C . Yet when marginal bounds leave multiple treatments undominated (Table 2), an automated system should not dictate a single CATE-maximizing choice. Framing the remaining uncertainty in terms of counterfactual benefit and harm helps clinicians ex-

plain the trade-off behind a recommendation and allows patients to weigh efficacy against the risk of treatment-caused harm. Moreover, because a clinician’s initial treatment preference X^* may carry prognostic information beyond L (Stensrud et al., 2024), an AI system should record the clinician’s intention X^* and the patient’s harm aversion κ before presenting algorithmic advice, helping document independent clinical judgement and mitigate anchoring (World Health Organization, 2021).

At the population level, automated decision rules deployed at scale inevitably condition on a finite set of measured covariates L . Even when a new treatment improves average recovery given $L = l$, a CATE-optimal policy may inadvertently harm a small, unidentified, yet real subgroup of patients who would have recovered under standard care (Kallus, 2022). By incorporating partial identification bounds on counterfactual harm—either to eliminate only robustly dominated treatments (Section 4.2) or to cap worst-case population harm (Section 4.3)—policymakers can hedge against such unintended harm while preserving room for clinical judgement at the point of care (Manski, 2019).

Appendix A. Proof of Proposition 1

Proof Fix $L = l$ with $p_2(l) \geq p_1(l) \geq p_0(l)$ and $\kappa \geq 1$. For each active drug $x \in \{1, 2\}$, a proposed harm probability $h_x(l) = \text{PH}_x(l)$ determines the conditional joint distribution of $(Y(x), Y(0))$:

	$Y(0) = 0$	$Y(0) = 1$
$Y(x) = 0$	$1 - p_x(l) - h_x(l)$	$h_x(l)$
$Y(x) = 1$	$p_x(l) - p_0(l) + h_x(l)$	$p_0(l) - h_x(l)$

Because $p_x(l) \geq p_0(l)$, we have $L_x(l) = 0$, so the entries above are non-negative if and only if $0 \leq h_x(l) \leq U_x(l)$. Since both tables share the same marginal distribution of $Y(0)$, drawing $Y(0)$ first and then drawing $Y(1)$ and $Y(2)$ independently given $Y(0)$ and $L = l$ realizes every pair in $[0, U_1(l)] \times [0, U_2(l)]$. Setting $U_0(l) = 0$ and $\tau_0(l) = 0$, the score $V_\kappa(x | l) = \tau_x(l) - (\kappa - 1)\text{PH}_x(l)$ for $x \in \{0, 1, 2\}$ and $\kappa \geq 1$ attains its sharp lower bound $\tau_x(l) - (\kappa - 1)U_x(l)$ at $\text{PH}_x(l) = U_x(l)$ and its sharp upper bound $\tau_x(l)$ at $\text{PH}_x(l) = 0$. Consequently, $x \succ_\kappa x'$ holds if and only if

$$\tau_x(l) - (\kappa - 1)U_x(l) > \tau_{x'}(l) \iff p_x(l) - p_{x'}(l) > (\kappa - 1)U_x(l),$$

which requires $p_x(l) > p_{x'}(l)$ (since $(\kappa - 1)U_x(l) \geq 0$) and holds if and only if $1 \leq \kappa < 1 + \{p_x(l) - p_{x'}(l)\}/U_x(l)$. $\blacksquare$

References

- Eli Ben-Michael, Kosuke Imai, and Zhichao Jiang. Policy learning with asymmetric counterfactual utilities. *Journal of the American Statistical Association*, 119(548):3045–3058, 2024.
- A. Philip Dawid and Stephen Senn. Personalised decision-making without counterfactuals. *Observational Studies*, 2026. To appear.

- Glyn Elwyn, Marie Anne Durand, Julia Song, Johanna Aarts, Paul J. Barr, Zackary Berger, Nan Cochran, Dominick Frosch, Dariusz Galasiński, Pål Gulbrandsen, Paul K. J. Han, Martin Härter, Paul Kinnersley, Amy Lloyd, Manish Mishra, Lilisbeth Perestelo-Perez, Isabelle Scholl, Kounosuke Tomori, Lyndal Trevena, Holly O. Witteman, and Trudy van der Weijden. A three-talk model for shared decision making: multistage consultation process. *BMJ*, 359:j4891, 2017.
- Ethan Goh, Robert J. Gallo, Eric Strong, Yingjie Weng, Hannah Kerman, Jason A. Freed, Joséphine A. Cool, Zahir Kanjee, Kathleen P. Lane, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Jason Hom, Jonathan H. Chen, and Adam Rodman. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nature Medicine*, 31(4):1233–1238, 2025.
- Nathan Kallus. What’s the harm? Sharp bounds on the fraction negatively affected by treatment. In *Advances in Neural Information Processing Systems*, volume 35, pages 15996–16009, 2022.
- Michael R. Kosorok and Eric B. Laber. Precision medicine. *Annual Review of Statistics and Its Application*, 6(1):263–286, 2019. doi: 10.1146/annurev-statistics-030718-105251. URL <http://dx.doi.org/10.1146/annurev-statistics-030718-105251>.
- Charles F. Manski. *Patient Care under Uncertainty*. Princeton University Press, Princeton, NJ, 2019.
- Scott Mueller and Judea Pearl. Personalized decision making – a conceptual introduction. *Journal of Causal Inference*, 11(1):20220050, 2023.
- Jonathan Richens, Rory Beard, and Daniel H. Thompson. Counterfactual harm. In *Advances in Neural Information Processing Systems*, volume 35, pages 36350–36365, 2022.
- David L. Sackett, William M. C. Rosenberg, J. A. Muir Gray, R. Brian Haynes, and W. Scott Richardson. Evidence based medicine: What it is and what it isn’t. *BMJ*, 312(7023): 71–72, 1996. doi: 10.1136/bmj.312.7023.71.
- Aaron L. Sarvet and Mats J. Stensrud. Perspectives on “harm” in personalized medicine. *American Journal of Epidemiology*, 194(6):1743–1748, 2025.
- Amit N. Sawant and Mats J. Stensrud. Counterfactual harm: A counter-argument. *American Journal of Epidemiology*, 195(8):2230–2236, 2026.
- Martina Scauda, Tobias Freidling, and Qingyuan Zhao. Counterfactual optimization of policy interventions: Lexical ordering and leapfrogging. arXiv preprint arXiv:2608.20505, 2026.
- Mats J. Stensrud, Julien D. Laurendeau, and Aaron L. Sarvet. Optimal regimes for algorithm-assisted human decision-making. *Biometrika*, 111(4):1089–1108, 2024.
- World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*. World Health Organization, Geneva, 2021. ISBN 9789240029200.

Peng Wu, Qing Jiang, Shanshan Luo, and Zhi Geng. Safe individualized treatment rules with controllable harm rates. arXiv preprint arXiv:2505.05308, 2025.