The Probability of Tiered Benefit: Partial Identification with Robust and Stable Inference

Johan de Aguas JOHANMD@MATH.UIO.NO
Department of Mathematics, University of Oslo

Sebastian Krumscheid SEBASTIAN.KRUMSCHEID@KIT.EDU
Scientific Computing Center, Karlsruhe Institute of Technology

Johan Pensar JOHANPEN@MATH.UIO.NO
Department of Mathematics, University of Oslo

Guido Biele GUIDO.BIELE@FHI.NO
Dpt. Child Health & Development, Norwegian Institute of Public Health

Editors: Biwei Huang and Mathias Drton

Abstract

We define the *probability of tiered benefit* in scenarios with a binary exposure and an outcome that is either categorical with $K \geq 2$ ordered tiers or continuous partitioned by $K - 1$ fixed thresholds into disjoint intervals. Similarly to other pure counterfactual queries, this parameter is not g-identifiable without additional assumptions. We demonstrate that strong monotonicity does not suffice for point identification when $K \geq 3$ and provide sharp bounds both with and without such constraint. Inference and uncertainty quantification for these bounds are challenging tasks due to potential nonregularity induced by ambiguities in the underlying individualized optimization problems. Such ambiguities can arise from immunities or null treatment effects in subpopulations with positive probability, affecting the lower bound estimate and hindering conservative inference. To address these issues, we extend the available *stabilized one-step correction* (S1S) procedure by incorporating stratum-specific stabilizing matrices. Through simulations, we illustrate the benefits of this approach over existing alternatives. We apply our method to estimate bounds on the probabilities of tiered benefit and harm from pharmacological treatment for ADHD upon academic achievement, employing observational data from diagnosed Norwegian schoolchildren. Our findings indicate that while girls and children with low prior test performance could have moderate chances of both benefit and harm from treatment, a clear-cut recommendation remains uncertain across all strata.

Keywords: counterfactuals, benefit, identification, bounds, inference

1. Introduction

Counterfactual inference involves probabilistic assessments of events for which certain causal antecedents or consequences are contrary to facts (Balke and Pearl, 1994a,b). An exemplar scenario is determining *the probability that a given student would have passed a math qualification test had they taken ADHD medication, given that they did not actually take it and failed the test*. The scope of application of counterfactual queries spans across diverse domains such as epidemiology, economics, legal deliberation, psychology, and AI (Pearl, 2009). Notably, it finds high utility in explainability of machine learning (Beckers, 2022), event attribution (Hannart et al., 2016), impact assessment (Possebom and Riva, 2022), fairness analysis (Plečko and Bareinboim, 2024), and personalized decision-making (Mueller and Pearl, 2023a), among other tasks.

An important counterfactual parameter, in the context of a binary exposure and a binary outcome, is the *probability of necessity and sufficiency* (PNS). This parameter quantifies the proportion of *compliers* in the population, referring to units who would have benefited *if and only if* they had been treated (Pearl, 1999). Typically, this parameter is not g-identifiable, meaning it cannot be determined solely from any unconstrained combination of causal graphs, observational data and experiments (Robins and Greenland, 1989; Pearl, 1999). While point identification may be achievable under a monotonicity assumption (Balke and Pearl, 1997), such a constraint is often unrealistic, as it presupposes the absence of unintended effects from the exposure. Several generalizations of the PNS have been proposed for nonbinary categorical, continuous, and vector-valued exposures and outcomes (Li and Pearl, 2024a; Kawakami et al., 2024). Yet, these have mainly focused on settings with unordered categorical outcomes or on queries that are identifiable under monotonicity.

When g-identification is unfeasible, an alternative approach is *partial identification*. Bounds can always be computed from population-level observational and experimental distributions, or solely from the former under the assumption of conditional ignorability (Tian and Pearl, 2000). Yet, conducting inference and uncertainty quantification for estimates of these bounds is challenging due to potential sources of nonregularity or lack of smoothness in the involved functionals. This difficulty stems from impossibility results indicating that if the target functional is not *pathwise differentiable* at the true distribution, there exist no sequence of *regular* and *locally unbiased* estimators. Furthermore, correction procedures may not fully eliminate bias and could cause the variance to diverge (Hirano and Porter, 2012). Lack of regularity and differentiability not only affect the validity of asymptotic inference but also compromise the interpretation of uncertainty quantification under the bootstrap and the Bayesian frameworks (Dümbgen, 1993; Fang and Santos, 2019; Kitagawa et al., 2020).

Settings: The studied scenario involves a categorical pre-exposure covariate $X \in \mathcal{X}$ defining strata from the population, a binary point exposure $A \in \{0, 1\}$, and an ordered categorical outcome $\tilde{Y}$ with $K \geq 2$ tiers $C = \{C_k\}_{k=1}^{K}$. We also consider the case of a continuous outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}$, with support partitioned into K disjoint tiers given by intervals with fixed thresholds $c = \{c_k\}_{k=1}^{K-1}$. We define the *x-specific probability of tiered benefit* $\mathrm{PB}(x)$ as the probability that an individual in stratum $x \in \mathcal{X}$ would attain any outcome tier under no treatment and, counterfactually, a higher outcome tier under treatment. Sharp bounds for $\mathrm{PB}(x)$ follow organically from the Fréchet inequalities (Fréchet, 1951). We examine the problem of developing semiparametric estimators for these bounds that are consistent, doubly-robust, and with stable inference in potentially nonregular settings. Our goal is to construct valid *uncertainty regions* for $\mathrm{PB}(x)$, accounting for both systemic and aleatoric uncertainties.

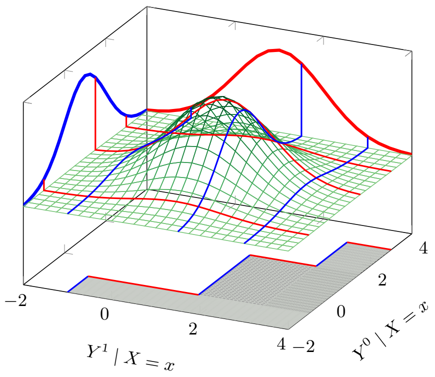

Figure 1: A setting with a continuous outcome and $K = 4, c_1 = -1.0, c_2 = 1.5, c_3 = 3.0$. Here, $\mathrm{PB}(x)$ is the volume under the joint PDF of potential outcomes $(Y^0, Y^1)|X = x$ enclosed above the benefit region (gray area).

Applied motivation: Educational and clinical fields are prime domains of application for the *probability of tiered benefit*, as quantitative measurements (e.g. test scores, blood pressure) are often aggregated into ordered tiers or assessed against fixed thresholds for tasks such as preliminary diagnosis, evaluation, categorization, and summarization.

For instance, in the Norwegian educational system, compulsory national tests assessing skills in numeracy, reading, and English are administered during school grades 5, 8, and 9. The Norwegian Directorate for Education has established fixed thresholds that group the test scores into *mastery tiers*, ranging from tier 1 (poorest) to tier 5 (highest). Considering the population of Norwegian schoolchildren diagnosed with ADHD and the context of treatment with stimulant medication, one can ask: *What is the probability that a child from a given subpopulation would achieve a higher test mastery tier under treatment than their corresponding counterfactual mastery tier under no treatment?* This is a pure counterfactual query, different from an experimentalist's investigation of the sign and magnitude of the *conditional average treatment effect* (CATE). Although these two approaches generally lead to similar conclusions regarding decision-making (Hedden, 2023), there is currently an illuminating debate on their contrasts, limitations, and bioethical considerations for personalized medicine. We refer the reader to Sarvet and Stensrud (2023, 2024) and Mueller and Pearl (2023a,b) for this debate.

Main contribution: We offer contributions to the problems of identification, partial identification, and doubly-robust, stable inference for the proposed *probability of tiered benefit*. This novel parameter generalizes the PNS to cases where the outcome is either ordered categorical or continuous with a fixed partition of its support. Unlike recent generalizations by Vlontzos et al. (2023), Li et al. (2023), and Kawakami et al. (2024), our focus is not restricted to queries where strong monotonicity ensures point identification. In fact, we show that this constraint does not suffice for identification when $K \geq 3$. We derive sharp bounds and introduce methods for estimation and inference for these bounds. To address uncertainty quantification issues induced by potential nonregularity, we adapt the *stabilized one-step correction* (S1S) approach by Luedtke and van der Laan (2018) with two extensions. First, our target estimand is a bivariate array containing the bounds, so we employ stabilizing 2×2 matrices rather than scalar weights, and leverage a martingale structure that yields a limiting bivariate Gaussian distribution for a linear transformation of the estimator's asymptotic bias. Second, we evaluate additive corrections using the next out-of-sample unit within each stratum, ensuring valid stratum-specific inference while employing pooled data to estimate population-level nuisance parameters efficiently.

Related literature: The *probabilities of causation* were formalized in counterfactual terms by Pearl (1999) for scenarios with binary exposure and outcome. It was shown that point identification can be achieved under a monotonicity assumption. Without this functional constraint, sharp bounds are derived using a combination of observational and experimental data (Tian and Pearl, 2000). Covariate information has been leveraged to define stratum-specific queries and to narrow the bounds through a marginalization step (Mueller et al., 2022; Li and Pearl, 2022). Partial identification has been extended to generalizations of the probabilities of causation in scenarios with nonbinary categorical variables (Li and Pearl, 2024a,b). More recently, a further generalization was proposed for continuous and vector-valued variables, where identification remains achievable under monotonicity constraints (Kawakami et al., 2024). Bounds have also been established for the limiting case of the distribution of *individualized treatment effects* (ITE) (Fan and Park, 2010), which is not g-identifiable due to the *fundamental problem of causal inference*, as well as for some of its functionals (Firpo and Ridder, 2010; Russell, 2021).

Robust inference for set-valued or partially identified parameters is an active area of research in statistics and econometrics. Given that bounds are often determined by extrema functionals of distributions, the estimation and inference tasks often target transformations of value functions from individualized optimization problems. Complete conditions for *pathwise differentiability* of such functionals have been characterized, enabling the development of *regular asymptotically linear* (RAL) estimators in a broad class of distributions referred to as *nonexceptional laws* (Robins, 2004; Luedtke and van der Laan, 2016). It has been shown that ambiguities and nonunique solutions in such optimization problems can result in lack of pathwise differentiability, making it impossible to obtain regular and locally unbiased estimators (Hirano and Porter, 2012). Nondifferentiability also poses challenges for other inference frameworks, causing various bootstrap-based methods to become inconsistent (Dümbgen, 1993; Fang and Santos, 2019), and making Bayesian credible intervals fail to be asymptotically equivalent to confidence intervals (Kitagawa et al., 2020).

Various methodological approaches have been proposed to conduct robust semiparametric inference in putative nonregular settings, including: *(i)* targeting thresholded or smoothed surrogates (Chakraborty et al., 2010), *(ii)* employing data-adaptive or cross-validated surrogates (Bibaut and van der Laan, 2017; van der Laan et al., 2018), and *(iii)* applying stabilizing weights to sequences of one-step corrections (Luedtke and van der Laan, 2016, 2018). Other strategies stem from alternative statistical optimality criteria, such as follows: *(iv)* median-bias correction for intersection bounds (Chernozhukov et al., 2013; Possebom and Riva, 2022), and *(v)* solutions using asymptotic minimax regret (Hirano and Porter, 2009; Song, 2014; Ponomarev, 2022; d'Adamo, 2023). Recently, the general problem has also been framed in terms of robust inference for conservative bounds (Balakrishnan et al., 2023) and Kantorovich dual bounds (Ji et al., 2023) from the optimal transport literature. In this context, nonregularities could be potentially addressed via post-hoc feasibility procedures.

2. The probability of tiered benefit and its identification

Let $X \in \mathcal{X}$ represent a categorical pre-exposure variable indicating different strata within a population, with $\min_{x \in \mathcal{X}} \mathbb{P}(X = x) > 0$.

Let $A \in \{0, 1\}$ denote a binary point exposure, where $A = 1$ signifies *treated* and $A = 0$ signifies *not treated*, and $\min_{a \in \{0,1\}} \mathbb{P}(A = a \mid X = x) > 0$ for all $x \in \mathcal{X}$.

Let Y be the outcome variable and Y^a denote the *potential outcome* under the intervention $\mathrm{do}(A = a)$. This is the value that Y assumes when A is fixed to have value $a \in \{0, 1\}$ within a *structural causal model* (SCM) $\mathcal{M}$. The potential outcome Y^a is a random variable that varies based on unit-level characteristics (Pearl, 2009; Bareinboim et al., 2022), and adheres to the consistency axiom, meaning that if $A = a$ and $Y = y$, then $Y^a = y$ (Robins, 1989).

Definition 1 *Let $\tilde{Y} \in C = \{C_k\}_{k=1}^K$ be a categorical outcome with $K \geq 2$ tiers, and $\preceq$ be a total order on C, with $\succ$ defined accordingly. The x-specific probability of tiered benefit $\mathrm{PB}(x)$ is the joint probability of attaining an outcome tier under no treatment $\tilde{Y}^0$ and, counterfactually, a higher outcome tier under treatment $\tilde{Y}^1$, for stratum $x \in \mathcal{X}$. That is:*

$$\mathrm{PB}(x) := \mathbb{P}\left(\bigvee_{k=1}^{K-1} [\tilde{Y}^0 = C_k \wedge \tilde{Y}^1 \succ C_k] \mid X = x\right) = \sum_{k=1}^{K-1} \mathbb{P}(\tilde{Y}^0 = C_k, \tilde{Y}^1 \succ C_k \mid X = x). \quad (1)$$

Definition 2 *Let $Y \in \mathcal{Y} \subseteq \mathbb{R}$ be an absolutely continuous outcome, with higher values being more desirable. Let c be a partition of the outcome support with fixed thresholds:* $\inf \mathcal{Y} = c_0 < c_1 < \cdots < c_{K-1} < c_K = \sup \mathcal{Y}$, *thereby producing $K \geq 2$ tiers given by disjoint intervals $I_k = (c_{k-1}, c_k]$, $k \in [K] := \{1, \cdots, K\}$. The x-specific probability of tiered benefit $\mathrm{PB}_c(x)$ is the joint probability* *of attaining an outcome value under no treatment Y^0 at any given tier and, counterfactually, an outcome value under treatment Y^1 at any higher tier, for stratum $x \in \mathcal{X}$. That is:*

$$\mathrm{PB}_c(x) := \mathbb{P}\left(\bigvee_{k=1}^{K-1} [Y^0 \in I_k \wedge Y^1 > c_k] \mid X = x\right) = \sum_{k=1}^{K-1} \mathbb{P}(Y^0 \in I_k, Y^1 > c_k \mid X = x). \quad (2)$$

The representation as sum of joint probabilities arises from the mutual exclusion of the counterfactual events involved. Without loss of generality, we will center on the given formulation within the context of a continuous outcome (definition 2).

The probability of tiered benefit can be computed from a fully specified SCM following a three-step procedure termed *abduction-action-prediction* (Pearl, 2009; Saha and Garain, 2022), or via an augmented graphical model known as a *twin network* (Balke and Pearl, 1994b). Such computations are sensitive to the specification of functional relationships in the SCM. Thus, in cases of *epistemic uncertainty* regarding the causal mechanisms, $\mathrm{PB}_c(\cdot)$ cannot be directly computed from the other components of the SCM (Pearl, 1999). Consequently, the probability of tiered benefit is not g-identifiable from any unconstrained combination of inputs involving the causal graph, observational data and subsidiary experiments (Robins and Greenland, 1989; Oberst and Sontag, 2019).

When $K = 2$, a single threshold c_1 separates the tiers. In this context, $\mathrm{PB}_c(x)$ returns the x-specific PNS (Pearl, 1999; Mueller et al., 2022), given by $\mathbb{P}(\tilde{Y}^0 = 0, \tilde{Y}^1 = 1 \mid X = x) = \mathbb{E}[(1 - \tilde{Y}^0)\tilde{Y}^1 \mid X = x]$, with $\tilde{Y}^a := \mathbb{I}(Y^a > c_1)$. Although it is not g-identifiable, it can be determined under a functional constraint known as *strong monotonicity*.

Assumption 1 (Strong monotonicity) *The potential outcome Y^a is strongly monotonic at stratum $X = x$, i.e. $\mathbb{P}(Y^1 - Y^0 \geq 0 \mid X = x) \in \{0, 1\}$.*

This assumption implies that treatment either universally benefits or universally harms individuals within the stratum. This is sufficient for identification of $\mathrm{PNS}(x)$, yielding the value $\mathrm{PNS}(x) = \Delta_a \mathbb{P}(Y^a > c_1 \mid X = x)$ for a nonharmful exposure (Tian and Pearl, 2000), where Δ_a is the difference operator relative to the binary term a.

For the general $K \geq 2$ case with a nonharmful exposure, strong monotonicity ensures that $\mathbb{P}(Y^0 > c_k, Y^1 \in I_k \mid X = x) = 0,\ \forall k \in [K-1]$ (Vlontzos et al., 2023). In other words, the *x-specific probability of tiered harm* is zero (Mueller and Pearl, 2023c), where this is defined analogously as $\mathrm{PH}_c(x) := \sum_{k=1}^{K-1} \mathbb{P}(Y^0 > c_k, Y^1 \in I_k \mid X = x)$.

Proposition 3 *For $K \geq 3$, strong monotonicity is not sufficient for point identification of the probability of tiered benefit.*

Justification of proposition 3 lies in the underdetermined nature of the linear system of counterfactual probabilities $\mathbb{P}(Y^0 \in I_a, Y^1 \in I_b \mid X = x)$, where $a, b \in [K]$. This system involves $N_K := (K-1)K/2$ null entries induced by strong monotonicity; $M_K := K(K+1)/2$ unknowns; and $L_K := 2K - 1$ constraints. The latter accounts for the $K - 1$ linearly independent marginal interventional probabilities for each arm, as well as the joint constraint that all cells sum to one. When $K = 2$, one has that $M_2 = L_2 = 3$, which allows for a unique solution for the system and hence for $\mathrm{PB}_c(x) = \mathrm{PNS}(x)$. Yet, for $K \geq 3$, one has that $M_K > L_K$, which prevents a unique solution.

One consequence is that, when $K \geq 3$, additional functional constraints on the causal mechanisms beyond monotonicity are necessary to identify $\mathrm{PB}_c(x)$ (Cinelli and Pearl, 2021). Yet, such constraints are hard to derive from domain expertise, and incorrect assumptions can lead to counterintuitive results (Oberst and Sontag, 2019; Vlontzos et al., 2023). As an alternative, we provide bounds $\Lambda(x) \leq \mathrm{PB}_c(x) \leq \Upsilon(x)$ both with and without the strong monotonicity constraint. We then propose a semiparametric estimation strategy for these bounds, enabling doubly-robust inference and valid asymptotic uncertainty quantification at both nonexceptional and potentially exceptional laws.

3. Partial identification

Let $W \in \mathcal{W}$ be a vector of pre-exposure variables such that the following two assumptions hold:

Assumption 2 (Conditional ignorability) $Y^a \perp\!\!\!\perp A \mid W, X, \ \forall a \in \{0,1\}$.

In the SCM framework, this assumption is implied by *backdoor admissibility* of $\{W, X\}$. That is, if $\mathcal{G}$ is a causal graph on the system's endogenous variables $\mathcal{V}$ containing $\{W, X, A, Y\}$, and $\mathcal{G}[\underline{\mathrm{A}}]$ is the mutilated graph removing the arrows coming out of A, then the d-separation statement $Y \perp\!\!\!\perp_d A \mid W, X$ in $\mathcal{G}[\underline{\mathrm{A}}]$ implies conditional ignorability (Pearl, 2009).

Assumption 3 (Propensity score positivity) *Let $\pi(w,x) := \mathbb{P}(A = 1 \mid W = w, X = x)$ be the propensity score, then $\pi(w,x) \in (0,1)$, $\forall (w,x) \in \mathcal{W} \times \mathcal{X}$ with $p_W(w \mid X = x)\,\mathbb{P}(X = x) > 0$.*

Let $S_k(w,x,a)$ and $R_k(w,x,a)$ be, respectively, the conditional probability of surpassing threshold c_k and the conditional probability of being in tier I_k under treatment arm $A = a$ given $(W = w, X = x)$. That is:

$$S_k(w,x,a) := \mathbb{P}(Y > c_k \mid W = w, X = x, A = a), \tag{3}$$

$$R_k(w,x,a) := \mathbb{P}(Y \in I_k \mid W = w, X = x, A = a) = S_{k-1}(w,x,a) - S_k(w,x,a). \tag{4}$$

Proposition 4 *For $K \geq 3$, and under strong monotonicity (assumption 1) with a nonharmful treatment, conditional ignorability (assumption 2), and propensity score positivity (assumption 3), sharp bounds for $\mathrm{PB}_c(x)$, with $x \in \mathcal{X}$, are given by:*

$$\mathrm{PB}_c(x) \geq \mathbb{E}_{W|X=x}\left[S_1(W,x,1) - R_K(W,x,0) - \sum_{k=2}^{K-1} \min\{R_k(W,x,0);\ R_k(W,x,1)\}\right], \tag{5}$$

$$\mathrm{PB}_c(x) \leq \mathbb{E}_{W|X=x}\left[S_1(W,x,1) - R_K(W,x,0) - \sum_{k=2}^{K-1} \max\{0;\ R_k(W,x,0) + R_k(W,x,1) - 1\}\right] \tag{6}$$

A detailed derivation is provided in appendix A.1.

In certain domains of application, strong monotonicity is an unrealistic assumption, as unintended effects from treatment may occur for some groups of units with positive probability. Therefore, we also provide a set of bounds not relying on strong monotonicity.

Proposition 5 *For $K \geq 2$, and under conditional ignorability (assumption 2), and propensity score positivity (assumption 3), sharp bounds for $\mathrm{PB}_c(x)$, with $x \in \mathcal{X}$, are given by:*

$$\mathrm{PB}_c(x) \geq \Lambda(x) := \sum_{k=1}^{K-1} \mathbb{E}_{W|X=x} \max\{0;\ R_k(W,x,0) + S_k(W,x,1) - 1\}, \tag{7}$$

$$\mathrm{PB}_c(x) \leq \Upsilon(x) := \sum_{k=1}^{K-1} \mathbb{E}_{W|X=x} \min\{R_k(W,x,0);\ S_k(W,x,1)\}. \tag{8}$$

A detailed derivation is provided in appendix A.2.

We now focus on estimation and inference for the stratum-specific bounds given by proposition 5. Thus, the target estimand is a bivariate array containing these bounds:

$$\Psi[P^*](x) := (\Lambda(x);\ \Upsilon(x))^\top \quad \text{for } x \in \mathcal{X}. \tag{9}$$

Here, $\Psi : \mathfrak{P} \times \mathcal{X} \to [0,1]^2$ is a functional that takes a distribution P in a semiparametric model $\mathfrak{P}$, along with a stratum $x \in \mathcal{X}$, and returns the value of the bounds for $\mathrm{PB}_c(x)$. We denote with P^* the true joint distribution of the data $O = (W, X, A, Y)$.

4. Estimation and inference at a nonexceptional law

Certain distributions known as *exceptional laws* may carry ambiguities that lead to nonregularity in the inference problem (Robins, 2004). Specifically, if for any $k \in [K-1]$, the two arguments of either the $\max$ or $\min$ operators in equations (7) and (8) are equal under P^* with positive probability, Ψ becomes nondifferentiable at P^*. In such cases, the limiting distribution of the plug-in estimator is nonstandard, and no regular and locally unbiased estimator for $\Psi[P^*]$ exists (Hirano and Porter, 2012). Consequently, standard methods —such as asymptotic, Bayesian, and most bootstrap-based approaches— fail to yield valid uncertainty quantification (Dümbgen, 1993; Fang and Santos, 2019; Kitagawa et al., 2020).

Proposition 6 *Let $x \in \mathcal{X}$ and $\mathcal{B}(x)$ be the set containing all $w \in \mathcal{W}$ such that for at least one $k \in [K-1]$ one has that either $R_k(w,x,0) + S_k(w,x,1) = 1$ or $R_k(w,x,0) - S_k(w,x,1) = 0$. Then, P^* is nonexceptional and $\Psi[\cdot](x)$ is pathwise differentiable at P^* if and only if $\mathcal{B}(x)$ has measure zero, i.e. $\int \mathbb{I}[w \in \mathcal{B}(x)]\, \mathrm{d}P^*_W(w \mid X = x) = 0$.*

This proposition follows from the extended definitions of exceptional laws and pathwise differentiability by Luedtke and van der Laan (2016).

At a nonexceptional law, estimation and inference can be conducted via a plug-in procedure $\widehat{\Psi}_{\text{plug}}(x) := \Psi[\widehat{P}](x)$, where $\widehat{P}$ couples the empirical distribution of $W \mid X = x$ with an estimated semiparametric model for the conditional outcome distribution. For instance, the latter can be specified as a Bayesian model with Gaussian likelihood, conditional mean $\mu(W, X, A)$ and homoskedastic variance σ^2. This approach requires the correct specification of such a model for consistency and valid uncertainty quantification. Alternatively, we propose a semiparametric *doubly-robust* method that leverages the propensity score model as well. Under certain technical conditions, this approach provides a consistent estimator even if either the outcome model or the propensity score model is misspecified, as long as both are not incorrect simultaneously (Daniel, 2018).

Let $\{O_i\}_{i \in J}$ be a held-out dataset from the same population, then a doubly-robust estimator of the bounds can be constructed using *one-step corrections* (1S) of the plug-in estimator as follows:

$$\widehat{\Lambda}_{1S}(x) := \widehat{\Lambda}_{\text{plug}}(x) + \frac{1}{|J(x)|} \sum_{i \in J(x)} \sum_{k=1}^{K-1} \left(\widehat{D}_k^R(O_i) + \widehat{D}_k^S(O_i)\right) \cdot \lambda_k[\widehat{P}](O_i), \tag{10}$$

$$\widehat{\Upsilon}_{1S}(x) := \widehat{\Upsilon}_{\text{plug}}(x) + \frac{1}{|J(x)|} \sum_{i \in J(x)} \sum_{k=1}^{K-1} \left[\widehat{D}_k^R(O_i) - \left(\widehat{D}_k^R(O_i) - \widehat{D}_k^S(O_i)\right) \cdot \upsilon_k[\widehat{P}](O_i)\right], \tag{11}$$

where $J(x)$ denotes the subset of units within J for which $X = x$, and $\widehat{D}_k^R, \widehat{D}_k^S$ represent the estimated components of the (uncentered) *efficient influence functions* (EIF) for the expectations of R_k under no treatment and of S_k under treatment, respectively, and given by:

$$\widehat{D}_k^R(O_i) := \frac{1 - A_i}{1 - \widehat{\pi}(W_i, x)} \cdot \left(\mathbb{I}[Y_i \in I_k] - \widehat{R}_k(W_i, x, A_i)\right), \tag{12}$$

$$\widehat{D}_k^S(O_i) := \frac{A_i}{\widehat{\pi}(W_i, x)} \cdot \left(\mathbb{I}[Y_i > c_k] - \widehat{S}_k(W_i, x, A_i)\right). \tag{13}$$

Moreover, λ_k, υ_k are *individualized rules* $\lambda_k, \upsilon_k : \mathfrak{P} \times \mathcal{W} \times \mathcal{X} \to \{0, 1\}$ given by:

$$\lambda_k[P](w, x) := \mathbb{I}[R_k(w, x, 0) + S_k(w, x, 1) - 1 > 0], \tag{14}$$

$$\upsilon_k[P](w, x) := \mathbb{I}[R_k(w, x, 0) - S_k(w, x, 1) > 0]. \tag{15}$$

A detailed derivation is available in appendix A.3.

Intuitively, these rules pick the solution index of the individual-level optimization problems. For instance, $\lambda_k[P](w, x) = 1$ indicates that the second term in $\max\{0;\ R_k(w, x, 0) + S_k(w, x, 1) - 1\}$ is the maximum of the two inputs. They also determine the "derivatives" of the $\min$ and $\max$ operators. These rules are well-defined and nonambiguous, even when the two terms being compared equate.

The proposed 1S procedure is analogous to *double/debiased machine learning* (DML) (Chernozhukov et al., 2018). Alternatively, at a nonexceptional law, an asymptotically equivalent substitution estimator can also be constructed using *targeted minimum-loss estimation* (TMLE) (van der Laan and Rose, 2011, 2018). Confidence and uncertainty regions can be constructed using the EIF-based estimator for the asymptotic covariance of $\widehat{\Psi}_{1S}(x)$ (Daniel, 2018), given by:

$$\widehat{\Omega}_{1S}(x) := \frac{1}{|J(x)|} \widehat{\mathrm{Cov}} \left(\{\widehat{\Lambda}_i + \partial\widehat{\Lambda}_i;\ \widehat{\Upsilon}_i + \partial\widehat{\Upsilon}_i\}_{i \in J(x)}\right), \tag{16}$$

where $\widehat{\Lambda}_i$ and $\partial\widehat{\Lambda}_i$ denote the unit-level prediction from the plug-in estimator and its respective correction for the lower bound; with $\widehat{\Upsilon}_i$ and $\partial\widehat{\Upsilon}_i$ serving analogous roles for the upper bound.

5. Estimation and inference at a potentially exceptional law

Example 1: Consider the case of an exposure having zero conditional risk difference for surpassing the first threshold c_1 for a subpopulation $(w', x') \in \mathcal{W} \times \mathcal{X}$ with positive probability. In this scenario, the probability of being in the first tier is identical for both treated and untreated individuals in such a group, so $R_1(w', x', 0) = R_1(w', x', 1)$. Since $S_1(w', x', 1) - 1 = -R_1(w', x', 1)$, this produces a nonregular behavior in the first sum component $\max\{0; R_1(w', x', 0) - R_1(w', x', 1)\}$ of the lower bound.

Example 2: Suppose the exposure results in a zero conditional risk difference for surpassing the last inner threshold c_{K-1}, and that the probability of attaining either of the final two tiers, I_{K-1} and I_K, is the same under no treatment for some subpopulation (w', x') with positive probability. This leads to $R_{K-1}(w', x', 0) = R_K(w', x', 0) = S_{K-1}(w', x', 0)$, which creates ambiguity in the last sum term $\min\{S_{K-1}(w', x', 0); S_{K-1}(w', x', 1)\}$ and hence nonregularity in the upper bound.

The first example presents a more plausible concern, particularly when prior knowledge indicates that the treatment effect could be zero or minimal, as it only necessitates a null conditional risk difference for surpassing the first inner threshold. In this case, the lower bound is impacted, hindering conservative or pessimistic inference (Balakrishnan et al., 2023). However, nonregularities may also emerge from unforeseen factors, for instance, when subpopulations have varying immunity to the exposure or when tiers are designed to meet specific balance or fairness criteria.

As demonstrated by Luedtke and van der Laan (2016, 2018), parametric-rate inference and valid uncertainty quantification can still be achieved under certain technical conditions, even at potentially exceptional laws. Here, we adapt and extend the *stabilized one-step correction* (S1S) procedure with two key modifications. While the original approach is designed for population-level queries on the real line and utilizes scalar stabilizing weights, we generalize it to handle bivariate, stratum-specific queries, replacing scalar weights with *stabilizing matrices*. The adapted procedure is summarized in the following steps and detailed in algorithm 1.

(i) Permute the units in the dataset. Select an initial data batch size $l > 0$.

(ii) Learn nuisance parameters $\widehat{P}$ from the initial batch. Identify the stratum for the next unit $x' := X_{l+1}$. Compute the plug-in values $\widehat{\Lambda}(x')$ and $\widehat{\Upsilon}(x')$.

(iii) For the next unit and for each unit in the batch within the same stratum x', compute their one-step corrections as given by equations (10) and (11).

(iv) Calculate T: the inverse-root-square of the covariance matrix of the unit-level corrected predictions within the stratum x'.

(v) Correct the plug-in estimates of the bounds for stratum x' using the next unit's individual correction. Save it with T.

(vi) Add the next unit to the batch and repeat from (ii) until reaching the penultimate unit.

(vii) The final estimate is the matrix-weighted average of all corrected estimates, weighted by their corresponding stabilizing matrices T.

A detailed derivation with conditions for asymptotic convergence is available in appendix A.4.

This procedure requires a sample size big enough to guarantee $n(x) \geq 2$ for all strata, after discounting the initial batch. Uncertainty regions can then be constructed using Monte Carlo sampling from the asymptotic distribution. Let $\{(s_\Lambda, s_\Upsilon)\}_{h=1}^{H}$ represent samples from a bivariate Gaussian distribution with zero mean and covariance matrix $\widehat{\Omega}_{S1S}(x)$. Define $\widehat{s}(x)$ as the 97.5% quantile of $\max\{s_\Lambda, -s_\Upsilon\}$. This yields a potentially conservative uncertainty region for $\mathrm{PB}_c(x)$ that incorporates both aleatoric and systemic uncertainties, as follows:

$$\widehat{\Lambda}_{S1S}(x) - \widehat{s}(x) \leq \mathrm{PB}_c(x) \leq \widehat{\Upsilon}_{S1S}(x) + \widehat{s}(x), \tag{17}$$

giving the following statistical guarantee:

$$\liminf_{n\to\infty} \mathbb{P}\left(\widehat{\Lambda}_{S1S}(x) - \widehat{s}(x) \leq \mathrm{PB}_c(x) \leq \widehat{\Upsilon}_{S1S}(x) + \widehat{s}(x)\right) \geq 0.95. \tag{18}$$

6. Simulations

We perform a simulation study in a simplified setup inspired by example 1 in section 5, to evaluate the performance of various methods for conducting inference on the bounds of the probability of

Algorithm 1 Estimation and inference for $\mathrm{PB}_c(\cdot)$ via S1S with stabilizing matrices

```
Input: data $\{O_i\}_{i=1}^n$, outcome thresholds $c = \{c_k\}_{k=1}^{K-1}$, initial batch size $0 < l < n$,
set of learners (and super-learning schemes)
Output: estimates $\widehat{\Psi}_{S1S}(\cdot)$, and asymptotic covariance matrix $\widehat{\Omega}_{S1S}(\cdot)$
Permute data indices; and initialize $n(x) \leftarrow 0$, $m(x) \leftarrow (0,0)$, and $M(x) \leftarrow 0 \cdot \mathrm{Id}_2$, $\forall x \in \mathcal{X}$
for $j \in \{l, l+1, \dots, n-1\}$
    Let $\mathcal{D} \leftarrow [j]$; and learn nuisance parameters of $\widehat{P}$ (cond. outcome and propensity score) from $\mathcal{D}$
    Let $x' \leftarrow X_{j+1}$, $\widehat{\Lambda}(x') \leftarrow \Lambda[\widehat{P}](x')$ and $\widehat{\Upsilon}(x') \leftarrow \Upsilon[\widehat{P}](x')$
        $\widehat{\Lambda}_i \leftarrow \sum_{k=1}^{K-1} \max\left\{0;\ \widehat{R}_k(W_i, x, 0) + \widehat{S}_k(W_i, x, 1) - 1\right\}$,    $\forall i \in [j] : X_i = x'$
        $\widehat{\Upsilon}_i \leftarrow \sum_{k=1}^{K-1} \min\left\{\widehat{R}_k(W_i, x, 0);\ \widehat{S}_k(W_i, x, 1)\right\}$,    $\forall i \in [j] : X_i = x'$
        $\partial\widehat{\Lambda}_i \leftarrow \sum_{k=1}^{K-1} \left(\widehat{D}_k^R(O_i) + \widehat{D}_k^S(O_i)\right) \cdot \lambda_k[\widehat{P}; O_i]$,    $\forall i \in [j+1] : X_i = x'$
        $\partial\widehat{\Upsilon}_i \leftarrow \sum_{k=1}^{K-1} \left[\widehat{D}_k^R(O_i) - \left(\widehat{D}_k^R(O_i) - \widehat{D}_k^S(O_i)\right) \cdot \upsilon_k[\widehat{P}; O_i]\right]$, $\forall i \in [j+1] : X_i = x'$
        $T(x') \leftarrow \widehat{\mathrm{Cov}}\left(\{\widehat{\Lambda}_i + \partial\widehat{\Lambda}_i;\ \widehat{\Upsilon}_i + \partial\widehat{\Upsilon}_i\}_{i \in \mathcal{D} : X_i = x'}\right)^{-\frac{1}{2}}$
        $m(x') \leftarrow m(x') + T(x')\left[(\widehat{\Lambda}(x');\ \widehat{\Upsilon}(x'))^\top + (\partial\widehat{\Lambda}_{j+1};\ \partial\widehat{\Upsilon}_{j+1})^\top\right]$
        $M(x') \leftarrow M(x') + T(x')$ and $n(x') \leftarrow n(x') + 1$
end for
$\widehat{\Psi}_{S1S}(x) \leftarrow M(x)^{-1} m(x)$ and $\widehat{\Omega}_{S1S}(x) \leftarrow n(x) M^{-2}(x)$, $\forall x \in \mathcal{X}$
return $\widehat{\Psi}_s(\cdot)$ and $\widehat{\Omega}_{S1S}(\cdot)$
```

tiered benefit. Data are generated according to the following SCM:

$$\begin{aligned}
&W_1 \sim \mathrm{Unif}(-1,1),\\
&W_2 = \mathbb{I}\left[v^2 > 0.25\right], \quad X = \mathbb{I}[v > 0], \quad v \sim \mathrm{Unif}(-1,1),\\
&A = \mathbb{I}\left[0.5\,(W_1 + 2X - 1) + u_A > 0\right], \quad u_A \sim N(0,1),\\
&Y = (2A-1) + (W_1 + X) \cdot [1 + 0.5(2A-1)] - A \cdot W_2 \cdot (W_1 + X + 2) + u_Y,\ u_Y \sim N(0,4).
\end{aligned}$$

Here $\mathrm{Unif}(a,b)$ denotes the continuous uniform distribution with support $[a,b]$, and $N(\mu,\sigma^2)$ indicates the Gaussian distribution with mean μ and variance σ^2. In this setup, we set $K = 3$, with fixed thresholds at $c_1 = -1.42$ and $c_2 = 1.09$.

Since the outcome follows a homoscedastic Gaussian distribution, the conditional risk difference, $\Delta_a\mathbb{P}(Y > c \mid W_1, W_2, X, A = a)$, is zero whenever the CATE, $\Delta_a\mathbb{E}\left[Y \mid W_1, W_2, X, A = a\right]$, is zero, for any value c in the outcome's support. Consider the subpopulation with $W_2 = 1$, which has positive probability in both strata $X \in \{0,1\}$. For this group, the conditional mean outcome is $\mathbb{E}\left[Y \mid W_1, W_2 = 1, X, A = a\right] = 0.5\,(W_1 + X) - 1$, which is independent of a, implying a null CATE. This causes nonregularity in the sum component of the lower bound for the first threshold, impacting the joint inference of the bounds. In other words, the joint distribution P^* governing the SCM is exceptional due to the existence of a subpopulation, with positive probability, that is immune to treatment.

We benchmark four types of estimators: *(i)* the naïve plug-in estimator, *(ii)* the one-step corrected (1S) estimator from section 4, *(iii)* the stabilized one-step corrected (S1S) estimator detailed in section 5, and *(iv)* the one-step corrected estimator with *smooth surrogates*. We use a sample size of $n = 5\,000$ and, for the S1S procedure, we set an initial batch size of $l = 2\,000$. For the

Table 1: Performance of proposed procedures and other methods in simulation setup in relation to marginal coverage of confidence intervals for lower bound (Lo.) and upper bound (Up.) of the probability of tiered benefit, joint coverage (Jo.) of bounds, and MSE results.

	$X = 0$					$X = 1$				
	Coverage (%)			MSE $\times 10^3$		Coverage (%)			MSE $\times 10^3$	
Estim.	Lo.	Up.	Jo.	Lo.	Up.	Lo.	Up.	Jo.	Lo.	Up.
Plug-in estimator	0.5	59.0	0.0	0.925	0.392	40.5	82.0	37.0	0.305	0.472
One-step (1S) corrected	95.0	77.5	75.0	0.257	0.839	96.0	93.0	91.5	0.356	0.331
1S with GELU($h = .05$)	95.5	66.5	64.0	0.279	1.028	96.5	92.5	91.0	0.357	0.334
1S with GELU($h = .15$)	70.5	46.0	28.0	0.784	2.029	91.0	85.5	75.5	0.497	0.702
S1S with stab. matices	**95.5**	**95.5**	**92.5**	0.607	0.654	**97.0**	**98.0**	**95.5**	0.589	0.522

surrogates, we replace nondifferentiable terms with a smooth variant, i.e., $\max\{0, x\}$ is approximated by $\mathrm{GELU}(x, h) := x \cdot \Phi(x/h)$, where Φ denotes the Gaussian CDF and $h > 0$ is a smoothing parameter (Lee, 2023). The outcome and propensity score models for all estimators are fit using *multivariate adaptive regression splines* and logistic regression. The true values of $\mathrm{PB}_c(\cdot)$ and its bounds are 0.30 with range 0.16–0.69 for $X = 0$, and 0.37 with range 0.25–0.66 for $X = 1$. Results, averaged over 200 iterations, are presented in table 1.

The 1S procedure consistently outperforms the naïve plug-in estimator across all metrics, especially in enhancing coverage for the lower bound. However, joint coverage remains below the nominal 95% in both strata. The performance of the smooth surrogate estimator is highly sensitive to the hyperparameter h: at $h = 0.05$, it performs similarly to the 1S procedure, but at $h = 0.15$, most performance metrics worsen. The S1S estimator improves joint coverage across both strata, particularly for stratum $X = 0$, though it is slightly conservative in its coverage for stratum $X = 1$ and exhibits higher MSE compared to the 1S estimator. However, coverage improvement comes at a significantly higher computational cost, as it requires re-learning all nuisance parameters with each new unit added to the data batch.

7. Application case

Attention-deficit/hyperactivity disorder (ADHD) is a neurodevelopmental disorder marked by persistent inattention and impulsivity, negatively impacting social, academic, and occupational functioning (World Health Organization, 2022). Affecting approximately 2–6% of children and adolescents globally, ADHD is among the most common mental health conditions in youth (Cortese et al., 2023). Individuals with ADHD often experience negative outcomes, including reduced quality of life and academic underachievement (Faraone et al., 2021). While stimulant medication effectively alleviates symptoms (Cortese et al., 2018), its impact on school-related outcomes is modest, with benefits not always translating to improved learning or higher standardized test scores (Storebø et al., 2015; Pelham et al., 2022).

In the Norwegian educational system, compulsory tests in numeracy, reading, and English are administered during school grades 5, 8, and 9. Since 2014, test scores have been categorized by the Norwegian Directorate for Education into *mastery tiers*, from tier 1 (poorest) to tier 5 (highest). On average, 12% of all schoolchildren do not achieve tier 2, and 34% do not reach tier 3; these

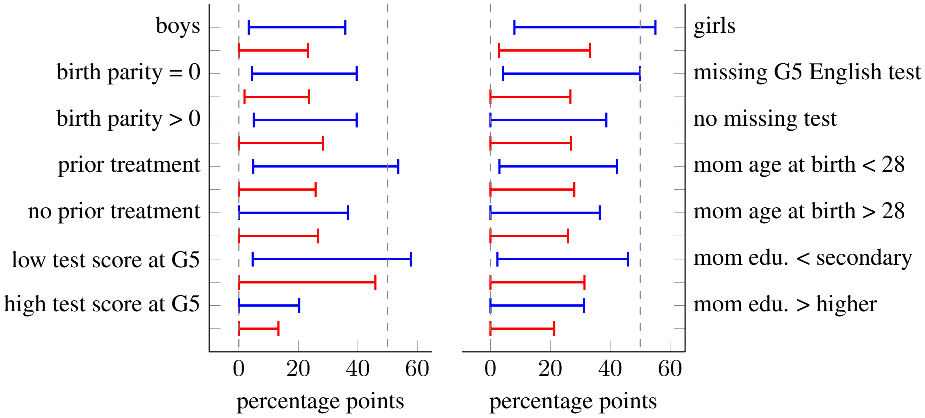

Figure 2: Uncertainty regions for the probability of tiered benefit (in blue) and for the probability of tiered harm (in red) for different strata of the population of Norwegian schoolchildren diagnosed with ADHD, relative to stimulant medication treatment and the mastery tiers for the numeracy national test at eighth school grade. Vertical broken lines indicate 0% and 50%, and *G5* indicates fifth school grade.

figures rise to 26% and 52%, respectively, for those with ADHD. Besides, methylphenidate, the most common ADHD medication, can cause adverse side effects such as sleep disturbances and appetite loss (Graham and Coghill, 2008), potentially affecting test performance and calling into question the applicability of the monotonicity assumption in this context.

We apply the developed methods to estimate the probability of tiered benefit and harm from pharmacological treatment for ADHD upon academic achievement, using the official thresholds among mastery tiers 1, 2, and 3 in the Norwegian numeracy test at eighth grade. As prior evidence suggests a small effect of stimulant medication on academic achievement, applying the stabilized one-step correction (S1S) procedure from section 5 is well-suited to avoid regularity issues arising from potential null treatment effects in some subpopulations. This analysis uses observational registry data from 9 352 individuals and includes a comprehensive set of pre-exposure variables, covering school, family, and child-level information, medical history, sociodemographics, and parental characteristics. Estimated uncertainty regions by the S1S procedure are shown in figure 2, with additional details on data sources and processing steps provided in appendix B.

These results can be translated into recommendations using the decision-making rules under partial identification proposed by Cui (2021) and the utility functions defined by Li and Pearl (2022). From a *pessimistic* perspective, the probability of benefit is uniformly small across all strata, peaking at 8% for girls —meaning only 8% of girls would achieve a higher mastery tier with treatment than they would without it—. Meanwhile, the probability of being harmed by stimulant medication can be as high as 33% for girls, and even higher, reaching 46% for children who already had low numeracy scores in fifth grade. In contrast, from an *optimistic* viewpoint, the probability of harm is nearly zero across all strata. The groups with the greatest chances of benefiting from medication are children with low numeracy scores in fifth grade (58%), girls (55%), and those who received prior ADHD treatment (54%). From a strict *opportunistic* standpoint, a treatment recommendation cannot be made, as in no group does the minimum benefit exceed the maximum harm.

The causal interpretation of these results hinges on several graphical and statistical assumptions, some of which are untestable. The potential impacts of violations of some of these assumptions could, in principle, be explored through sensitivity analysis (Díaz et al., 2018) and extended partial identification techniques (Ding and VanderWeele, 2016; VanderWeele and Ding, 2017).

8. Discussion

The proposed probability of tiered benefit extends the probability of necessity and sufficiency to new contexts. Though application-agnostic, it can be especially relevant in education and clinical sciences, where outcomes are often tiered and benchmarked against fixed thresholds for purposes like summarization, diagnosis, and intervention evaluation. While partial identification strategies, even when sharp, do not inherently guarantee informative bounds, they provide a more transparent and robust approach compared to imposing strict functional assumptions that may be hard to justify and verify.

The proposed 1S and S1S estimators yield doubly-robust inferential results that outperform the plug-in estimator for bound functionals, even in nonregular scenarios. These approaches could be extended to other bounds in causal inference, including assumption-free bounds for causal effects (Pearl, 1999), instrumental variables (Balke and Pearl, 1997), sensitivity analyses (Díaz et al., 2018), and latent confounding (Ding and VanderWeele, 2016; VanderWeele and Ding, 2017).

The S1S procedure provides more reliable results under exceptional distributions, which can emerge from immunities or null treatment effects in certain subpopulations. However, it has some limitations. Firstly, it is computationally intensive. Suppose flexible machine learning models and schemes are implemented, and the running time of the 1S estimator is $\mathcal{O}(n \log n)$ for a large sample size n relative to the number of features. In this case, the running time of S1S estimator, with unit-batch increases, would be $\mathcal{O}(n^2 \log n)$, which becomes prohibitively expensive for large n. Secondly, because it is not a substitution estimator, it may produce estimates outside the parameter space. To address this, computational efficiency could be improved by integrating and adapting methods from the *online learning* paradigm (van der Laan and Lendle, 2014) along with TMLE-based solutions for nondifferentiable parameters (van der Laan et al., 2018). We aim to explore these in future work.

Acknowledgements: We thank the organizing committee of the Causal Inference Workshop at UAI 2024 in Barcelona for the opportunity to present an early version of this work and receive valuable feedback.

Author contributions: Johan de Aguas conceptualized the theoretical and methodological research questions, designed and implemented the statistical and algorithmic tools, and drafted the manuscript. Sebastian Krumscheid and Johan Pensar provided feedback on theorems and other mathematical results, and revised the manuscript. Guido Biele proposed the study design, preprocessed the data, provided insights on the application case, and revised the manuscript.

Funding information: This study was supported by the Research Council of Norway (project number 301081, PI: Guido Biele). Johan Pensar received funding from the Research Council of Norway through Integreat – The Norwegian Centre for Knowledge-Driven Machine Learning (project number 332645). In addition, this work was partially funded by a joint program between the Norwegian Artificial Intelligence Research Consortium (NORA) and the HIDA Helmholtz Information & Data Science Academy (HIDA) to facilitate a research visit to the Karlsruhe Institute of Technology (KIT).

Ethical approval: REK ethical approval number 96604.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The data and code utilized for the simulation study are accessible in a personal GitHub repository at https://github.com/johandh2o. Due to the sensitive nature of the research topic, data from the application case cannot be obtained from the authors, but have to be requested from the relevant Norwegian authorities.

References

Balakrishnan, S., Kennedy, E., and Wasserman, L. Conservative inference for counterfactuals. *arXiv preprint arXiv:2310.12757*, 2023.

Balke, A. and Pearl, J. Counterfactual probabilities: Computational methods, bounds and applications. In de Mantaras, R. L. and Poole, D., editors, *Uncertainty in Artificial Intelligence*, pages 46–54. Morgan Kaufmann, San Francisco (CA), 1994a. ISBN 978-1-55860-332-5. doi: https://doi.org/10.1016/B978-1-55860-332-5.50011-0. URL https://www.sciencedirect.com/science/article/pii/B9781558603325500110.

Balke, A. and Pearl, J. Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1)*, Aaai '94, pages 230–237, Usa, 1994b. American Association for Artificial Intelligence. ISBN 0262611023.

Balke, A. and Pearl, J. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997. ISSN 01621459. URL http://www.jstor.org/stable/2965583.

Bareinboim, E., Correa, J., Ibeling, D., and Icard, T. *On Pearl's Hierarchy and the foundations of causal inference*, pages 507–556. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL https://doi.org/10.1145/3501714.3501743.

Beckers, S. Causal explanations and XAI. In *Conference on Causal Learning and Reasoning*, pages 90–109. Pmlr, 2022.

Bibaut, A. and van der Laan, M. Data-adaptive smoothing for optimal-rate estimation of possibly non-regular parameters. *arXiv e-prints*, pages arXiv–1706, 2017.

Chakraborty, B., Strecher, V., and Murphy, S. Inference for nonregular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*, 19(3):317–343, 2010.

Chernozhukov, V., Lee, S., and Rosen, A. Intersection bounds: estimation and inference. *Econometrica*, 81(2):667–737, 2013. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/23524295.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–c68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Cinelli, C. and Pearl, J. Generalizing experimental results by leveraging knowledge of mechanisms. *European Journal of Epidemiology*, 36(2):149–164, 2021. doi: 10.1007/s10654-020-00687-4. URL https://doi.org/10.1007/s10654-020-00687-4.

Cortese, S., Adamo, N., Del Giovane, C., Mohr-Jensen, C., Hayes, A., Carucci, S., Atkinson, L., Tessari, L., Banaschewski, T., Coghill, D., Hollis, C., Simonoff, E., Zuddas, A., Barbui, C., Purgato, M., Steinhausen, H.-C., Shokraneh, F., Xia, J., and Cipriani, A. Comparative efficacy and tolerability of medications for attention-deficit hyperactivity disorder in children, adolescents, and adults: a systematic review and network meta-analysis. *The lancet. Psychiatry*, 5(9):727–738, September 2018. ISSN 2215-0374, 2215-0366. doi: 10.1016/s2215-0366(18)30269-4. URL http://dx.doi.org/10.1016/S2215-0366(18)30269-4.

Cortese, S., Song, M., Farhat, L., Yon, D., Lee, S., Kim, M., Park, S., Oh, J., Lee, S., and Cheon, K.-A. Incidence, prevalence, and global burden of ADHD from 1990 to 2019 across 204 countries: data, with critical re-analysis, from the global burden of disease study. *Molecular Psychiatry*, pages 1–8, 2023.

Cui, Y. Individualized decision making under partial identification: Three perspectives, two optimality results, and one paradox. *Harvard Data Science Review*, 2021. doi: 10.1162/99608f92.d07b8d16. URL http://dx.doi.org/10.1162/99608f92.d07b8d16.

d'Adamo, R. Orthogonal policy learning under ambiguity. *arXiv preprint arXiv:2111.10904*, 2023.

Daniel, R. *Double Robustness*, pages 1–14. John Wiley and Sons, Ltd, 2018. ISBN 9781118445112. doi: https://doi.org/10.1002/9781118445112.stat08068. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08068.

Díaz, I., Luedtke, A., and van der Laan, M. *Sensitivity Analysis*, pages 511–522. Springer International Publishing, Cham, 2018. ISBN 978-3-319-65303-7. doi: 10.1007/978-3-319-65304-4_27. URL https://doi.org/10.1007/978-3-319-65304-4%5F27.

Ding, P. and VanderWeele, T. Sensitivity analysis without assumptions. *Epidemiology*, 27(3): 368–377, 2016.

Dümbgen, L. On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95:125–140, 1993.

Fan, Y. and Park, S. Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(3):931–951, 2010. URL https://EconPapers.repec.org/RePEc:cup:etheor:v:26:y:2010:i:03:p:931-951%5F99.

Fang, Z. and Santos, A. Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1 (306)):377–412, 2019.

Faraone, S., Banaschewski, T., Coghill, D., Zheng, Y., Biederman, J., Bellgrove, M., Newcorn, J., Gignac, M., Al Saud, N., Manor, I., et al. The world federation of ADHD international consensus statement: 208 evidence-based conclusions about the disorder. *Neuroscience & Biobehavioral Reviews*, 128:789–818, 2021.

Firpo, S. and Ridder, G. Bounds on functionals of the distribution treatment effects. Textos para discussão 201, FGV EESP - Escola de Economia de São Paulo, Fundação Getulio Vargas (Brazil), 6 2010. URL https://ideas.repec.org/p/fgv/eesptd/201.html.

Fréchet, M. Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université de Lyon. Série A: Sciences mathématiques et astronomie*, 14:53–77, 1951.

Graham, J. and Coghill, D. Adverse effects of pharmacotherapies for attention-deficit hyperactivity disorder: epidemiology, prevention and management. *CNS drugs*, 22:213–237, 2008.

Hannart, A., Pearl, J., Otto, F., Naveau, P., and Ghil, M. Causal counterfactual theory for the attribution of weather and climate-related events. *Bulletin of the American Meteorological Society*, 97(1):99–110, 2016. doi: 10.1175/bams-d-14-00034.1. URL https://journals.ametsoc.org/view/journals/bams/97/1/bams-d-14-00034.1.xml.

Hedden, B. Counterfactual Decision Theory. *Mind*, 132(527):730–761, 04 2023. ISSN 0026-4423. doi: 10.1093/mind/fzac060. URL https://doi.org/10.1093/mind/fzac060.

Hines, O., Dukes, O., Díaz-Ordaz, K., and Vansteelandt, S. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76(3):292–304, 2022.

Hirano, K. and Porter, J. Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701, 2009. URL http://www.jstor.org/stable/25621374.

Hirano, K. and Porter, J. Impossibility results for nondifferentiable functionals. *Econometrica*, 80(4): 1769–1790, 2012. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/23271416.

Ji, W., Lei, L., and Spector, A. Model-agnostic covariate-assisted inference on partially identified causal effects, 2023. URL https://arxiv.org/abs/2310.08115.

Kawakami, Y., Kuroki, M., and Tian, J. Probabilities of causation for continuous and vector variables. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024. URL https://openreview.net/forum?id=em3fzm95So.

Kitagawa, T., Montiel-Olea, J., Payne, J., and Velez, A. Posterior distribution of nondifferentiable functions. *Journal of Econometrics*, 217(1):161–175, 2020. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2019.10.009. URL https://www.sciencedirect.com/science/article/pii/S0304407620300014.

Lee, M. Mathematical analysis and performance evaluation of the GELU activation function in deep learning. *Journal of Mathematics*, 2023(1):4229924, 2023. doi: https://doi.org/10.1155/2023/4229924. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2023/4229924.

Li, A. and Pearl, J. Unit selection with causal diagram. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 5765–5772, 2022.

Li, A. and Pearl, J. Probabilities of causation with nonbinary treatment and effect. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20465–20472, 2024a.

Li, A. and Pearl, J. Unit selection with nonbinary treatment and effect. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20473–20480, 2024b.

Li, W., Lu, Z., Jia, J., Xie, M., and Geng, Z. Retrospective causal inference with multiple effect variables. *Biometrika*, 111(2):573–589, 09 2023. ISSN 1464-3510. doi: 10.1093/biomet/asad056. URL https://doi.org/10.1093/biomet/asad056.

Luedtke, A. and van der Laan, M. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713, 2016.

Luedtke, A. and van der Laan, M. Parametric-rate inference for one-sided differentiable parameters. *Journal of the American Statistical Association*, 113(522):780–788, 2018. doi: 10.1080/01621459.2017.1285777. URL https://doi.org/10.1080/01621459.2017.1285777. Pmid: 30078921.

Mueller, S. and Pearl, J. Personalized decision making – a conceptual introduction. *Journal of Causal Inference*, 11(1):20220050, 2023a. doi: doi:10.1515/jci-2022-0050. URL https://doi.org/10.1515/jci-2022-0050.

Mueller, S. and Pearl, J. Perspective on harm in personalized medicine – an alternative perspective. Technical report, Technical Report R-530, Department of Computer Science, University of California, Los Angeles, CA, 2023. Forthcoming, American Journal of Epidemiology, 2024, 2023b.

Mueller, S. and Pearl, J. Monotonicity: Detection, refutation, and ramification. Technical Report R-529, Department of Computer Science, University of California, Los Angeles, Ca, 2023c. URL http://ftp.cs.ucla.edu/pub/stat%5Fser/r529.pdf.

Mueller, S., Li, A., and Pearl, J. Causes of effects: Learning individual responses from population data. In Raedt, L. D., editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2712–2718. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/376. URL https://doi.org/10.24963/ijcai.2022/376. Main Track.

Oberst, M. and Sontag, D. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890. Pmlr, 2019.

Pearl, J. Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, 121(1):93–149, 1999. doi: 10.1023/a:1005233831499. URL https://doi.org/10.1023/A:1005233831499.

Pearl, J. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009. ISBN 978-0-521-89560-6. doi: 10.1017/cbo9780511803161.

Pelham, W., Altszuler, A., Merrill, B., Raiker, J., Macphee, F., Ramos, M., Gnagy, E., Greiner, A., Coles, E., Connor, C., et al. The effect of stimulant medication on the learning of academic curricula in children with ADHD: A randomized crossover study. *Journal of consulting and clinical psychology*, 90(5):367, 2022.

Plečko, D. and Bareinboim, E. Causal fairness analysis: A causal toolkit for fair machine learning. *Foundations and Trends® in Machine Learning*, 17(3):304–589, 2024. ISSN 1935-8237. doi: 10.1561/2200000106. URL http://dx.doi.org/10.1561/2200000106.

Ponomarev, K. *Essays in Econometrics*. PhD thesis, University of California in Los Angeles, 2022. URL https://escholarship.org/uc/item/1kz2n299.

Possebom, V. and Riva, F. Probability of causation with sample selection: A reanalysis of the impacts of Jóvenes en Acción on formality. *arXiv preprint arXiv:2210.01938*, 2022.

Robins, J. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, pages 113–159, 1989.

Robins, J. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, Lecture Notes in Statistics, pages 179–326. Springer, New York, 2004.

Robins, J. and Greenland, S. The probability of causation under a stochastic model for individual risk. *Biometrics*, 45(4):1125–1138, 1989. ISSN 0006341x, 15410420. URL http://www.jstor.org/stable/2531765.

Russell, T. Sharp bounds on functionals of the joint distribution in the analysis of treatment effects. *Journal of Business & Economic Statistics*, 39(2):532–546, 2021. doi: 10.1080/07350015.2019.1684300. URL https://doi.org/10.1080/07350015.2019.1684300.

Saha, S. and Garain, U. On noise abduction for answering counterfactual queries: A practical outlook. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=4FU8Jz1Oyj.

Sarvet, A. and Stensrud, M. Perspective on Harm in Personalized Medicine. *American Journal of Epidemiology*, page kwad162, 07 2023. ISSN 0002-9262. doi: 10.1093/aje/kwad162. URL https://doi.org/10.1093/aje/kwad162.

Sarvet, A. and Stensrud, M. Rejoinder to" perspectives on harm in personalized medicine–an alternative perspective". *arXiv preprint arXiv:2403.14869*, 2024.

Song, K. Local asymptotic minimax estimation of nonregular parameters with translation-scale equivariant maps. *Journal of Multivariate Analysis*, 125:136–158, 2014.

Storebø, O., Krogh, H., Ramstad, E., Moreira-Maia, C., Holmskov, M., Skoog, M., Nilausen, T., Magnusson, F., Zwi, M., Gillies, D., Rosendal, S., Groth, C., Rasmussen, K., Gauci, D., Kirubakaran, R., Forsbøl, B., Simonsen, E., and Gluud, C. Methylphenidate for attention-deficit/hyperactivity disorder in children and adolescents: Cochrane systematic review with meta-analyses and trial sequential analyses of randomised clinical trials. *British Medical Journal*, 351:h5203, November 2015. ISSN 0959-8138, 1756-1833. doi: 10.1136/bmj.h5203. URL http://www.bmj.com/content/351/bmj.h5203.

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

van der Laan, M. and Lendle, S. Online targeted learning. Technical report, University of California in Berkeley – Division of Biostatistics, 9 2014. URL https://biostats.bepress.com/ucbbiostat/paper330.

van der Laan, M. and Rose, S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Series in Statistics. Springer New York, 2011. ISBN 9781441997821. URL https://books.google.no/books?id=RGnSX5aCAgQC.

van der Laan, M. and Rose, S. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer Publishing Company, Incorporated, 1st edition, 2018. ISBN 3319653032.

van der Laan, M., Bibaut, A., and Luedtke, A. *CV-TMLE for Nonpathwise Differentiable Target Parameters*, pages 455–481. Springer International Publishing, Cham, 2018. ISBN 978-3-319-65304-4. doi: 10.1007/978-3-319-65304-4_25. URL https://doi.org/10.1007/978-3-319-65304-4%5F25.

VanderWeele, T. and Ding, P. Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274, 2017.

Vlontzos, A., Kainz, B., and Gilligan-Lee, C. Estimating categorical counterfactuals via deep twin networks. *Nature Machine Intelligence*, 5(2):159–168, 2023. doi: 10.1038/s42256-023-00611-x. URL https://doi.org/10.1038/s42256-023-00611-x.

World Health Organization. International Classification of Diseases 11th Revision, 2022. URL https://icd.who.int/en.

Appendix A. Proofs and derivations

A.1. Proof of proposition 4

Let $q_{a,b}(w,x) := \mathbb{P}(Y^0 \in I_a, Y^1 \in I_b \mid W = w, X = x)$ for $a, b \in [K]$. This is, $q_{a,b}$ is the (w,x)-specific joint probability of attaining outcome interval I_a under no treatment and, counterfactually, attaining I_b under treatment. Let $Q(w,x)$ be the matrix containing all these probabilities.

Under strong monotonicity (assumption 1) and a nonharmful treatment, one has that $q_{a,b} = 0$ for all $a > b$, and so $Q(w,x)$ can be represented by a lower triangular matrix:

$$Q(w,x) = \left(\begin{array}{ccccc:c}
q_{1,1} & 0 & 0 & \cdots & 0 & 0 \\
\hdashline
q_{1,2} & q_{2,2} & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
q_{1,K-2} & q_{2,K-2} & q_{3,K-2} & \cdots & 0 & 0 \\
q_{1,K-1} & q_{2,K-1} & q_{3,K-1} & \cdots & q_{K-1,K-1} & 0 \\
q_{1,K} & q_{2,K} & q_{3,K} & \cdots & q_{K-1,K} & \underbrace{q_{K,K}}_{\mathbb{P}(Y^0 \in I_K \mid W = w, X = x)}
\end{array}\right) \overbrace{\big\}}^{\mathbb{P}(Y^1 \in I_1 \mid W = w, X = x)} \tag{19}$$

All entries in $Q(w,x)$ add up to one and all off-diagonal entries add up to the (w,x)-specific probability of tiered benefit $\mathrm{PB}_c(w,x)$, thus:

$$1 = \mathrm{PB}_c(w,x) + \operatorname{tra} Q(w,x). \tag{20}$$

Strong monotonicity implies $q_{1,1}(w,x) = \mathbb{P}(Y^1 \in I_1 \mid W = w, X = x)$ and $q_{K,K}(w,x) = \mathbb{P}(Y^0 \in I_K \mid W = w, X = x)$ via fulfillment of the margin constraints. Moreover, under conditional ignorability (assumption 2) and propensity score positivity (assumption 3), these quantities are identified by $R_1(w,x,1)$ and $R_K(w,x,0)$ respectively. Therefore:

$$\begin{aligned}
\mathrm{PB}_c(w,x) &= 1 - q_{1,1}(w,x) - q_{K,K}(w,x) - \sum_{k=2}^{K-1} q_{k,k}(w,x) && (21)\\
&= [1 - R_1(w,x,1)] - R_K(w,x,0) - \sum_{k=2}^{K-1} q_{k,k}(w,x) && (22)\\
&= S_1(w,x,1) - R_K(w,x,0) - \sum_{k=2}^{K-1} q_{k,k}(w,x). && (23)
\end{aligned}$$

The quantities $q_{k,k}(w,x)$ are not identifiable for $k \in \{2, \ldots, K-1\}$; however, they can be bounded using the Fréchet inequalities (Fréchet, 1951).

$$\begin{aligned}
q_{k,k}(w,x) &\geq \max\{0;\ \mathbb{P}(Y^0 \in I_k \mid W = w, X = x) + \mathbb{P}(Y^1 \in I_k \mid W = w, X = x) - 1\}, && (24)\\
q_{k,k}(w,x) &\leq \min\{\mathbb{P}(Y^0 \in I_k \mid W = w, X = x);\ \mathbb{P}(Y^1 \in I_k \mid W = w, X = x)\}. && (25)
\end{aligned}$$

These bounds are sharp, meaning they can be attained and are the narrowest possible in the absence of additional information or constraints. The final bounds for $\mathrm{PB}_c(x)$ are given by marginalizing W out after replacing inputs by their identified quantities, resulting in:

$$\mathrm{PB}_c(x) \geq \mathbb{E}_{W|X=x}\left[S_1(W,x,1) - R_K(W,x,0) - \sum_{k=2}^{K-1} \min\{R_k(W,x,0);\ R_k(W,x,1)\}\right], \tag{26}$$

$$\mathrm{PB}_c(x) \leq \mathbb{E}_{W|X=x}\left[S_1(W,x,1) - R_K(W,x,0) - \sum_{k=2}^{K-1} \max\{0;\ R_k(W,x,0) + R_k(W,x,1) - 1\}\right] \tag{27}$$

The sharpness of these bounds is preserved through Jensen's inequality, which ensures that beginning with (w,x)-specific queries and subsequently applying a marginalization step yields narrower bounds than those obtained by starting with x-specific queries, as demonstrated by (Mueller et al., 2022).

A.2. Proof of proposition 5

Let $q_{a,b}(w,x)$ be defined as in appendix A.1. By the Fréchet inequalities, one gets:

$$q_{a,b}(w,x) \geq \max\{0;\ \mathbb{P}(Y^0 \in I_a \mid W = w, X = x) + \mathbb{P}(Y^1 \in I_b \mid W = w, X = x) - 1\}, \tag{28}$$

$$q_{a,b}(w,x) \leq \min\{\mathbb{P}(Y^0 \in I_a \mid W = w, X = x);\ \mathbb{P}(Y^1 \in I_b \mid W = w, X = x)\}. \tag{29}$$

Let $I_a = I_k$ and $I_b = (c_k, c_K)$. Under conditional ignorability (assumption 2) and propensity score positivity (assumption 3), these bounds are identified by:

$$q_{a,b}(w,x) \geq \max\{0;\ R_k(w,x,0) + S_k(w,x,1) - 1\}, \tag{30}$$

$$q_{a,b}(w,x) \leq \min\{R_k(w,x,0);\ S_k(w,x,1)\}. \tag{31}$$

By marginalizing W out and summing over all $k \in [K-1]$, one gets:

$$\mathrm{PB}_c(x) \geq \Lambda(x) := \sum_{k=1}^{K-1} \mathbb{E}_{W|X=x} \max\{0;\ R_k(W,x,0) + S_k(W,x,1) - 1\}, \tag{32}$$

$$\mathrm{PB}_c(x) \leq \Upsilon(x) := \sum_{k=1}^{K-1} \mathbb{E}_{W|X=x} \min\{R_k(W,x,0);\ S_k(W,x,1)\}. \tag{33}$$

A.3. Derivation of one-step corrected estimator

Consider a parametric submodel $P_\epsilon \in \mathfrak{P}$ indexed by a small fluctuation parameter $\epsilon \in \mathbb{R}$ and a point-mass contamination $O_i = (W_i, X_i, A_i, , Y_i)$, such that $P^\epsilon(O) = \epsilon\, \mathbb{I}(O = O_i) + (1-\epsilon)\, P^*(O)$. Under some technical conditions involving *(i)* fully nonparametric or saturated model $\mathfrak{P}$, *(ii)* smoothness of the paths within the model, *(ii)* positivity of the propensity score and of each stratum $x \in \mathcal{X}$, and *(iv)* boundedness of the outcome mean, the Gâteaux derivative, and their variances, we have that $\Lambda[\cdot](x)$ is pathwise differentiable at nonexceptional law $P^* \in \mathfrak{P}$ (Hines et al., 2022). Its *efficient influence* *function* (EIF) at P^* being evaluated at point $O_i \sim P^*$, with $X_i = x$, can be computed using the chain rule and gradient algebra for the Gâteaux derivative. Thus:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\epsilon}\Lambda[P^\epsilon](x)\bigg|_{\epsilon=0} &= \sum_{k=1}^{K-1}\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathbb{E}^{\epsilon}_{W|X=x}\max\left\{0;\ R^{\epsilon}_k(W,x,0)+S^{\epsilon}_k(W,x,1)-1\right\}\bigg|_{\epsilon=0} \qquad (34)\\
&= \sum_{k=1}^{K-1}\left(D^R_k(O_i)+D^S_k(O_i)\right)\cdot\lambda_k[P^*](W_i,x) \qquad (35)\\
&\quad+\sum_{k=1}^{K-1}\max\left\{0;\ R^*_k(W_i,x,0)+S^*_k(W_i,x,1)-1\right\}-\Lambda[P^*](x).
\end{aligned}$$

The expression in line (35) is justified by the application of the chain rule at nonexceptional P^*, and so the term $\max\{0;\ R_k(W,x,0)+S_k(W,x,1)-1\}$ is differentiable *almost everywhere*, with the functional derivative being the Heaviside step $\lambda_k(W,x) := \mathbb{I}[R_k(W,x,0)+S_k(W,x,1)-1>0]$.

Let $D^\Lambda(O_i) := \frac{\mathrm{d}}{\mathrm{d}\epsilon}\Lambda[P^\epsilon](x)\big|_{\epsilon=0}$ evaluated at P^* and observation $O_i \sim P^*$ with $X_i = x$. One key property of the EIF is that it satisfies the moment condition $\mathbb{E}_{O|X=x}D^\Lambda(O) = 0$, making its empirical counterpart suitable as an estimating equation for $\Lambda[P^*](x)$. Let $\widehat{P}$ be an initial estimator of P^* and J an evaluation dataset. Then, a one-step corrected estimator of $\Lambda[P^*](x)$, denoted $\widehat{\Lambda}_{1S}(x)$, can be constructed as a solution to the empirical moment condition of the EIF, namely:

$$\begin{aligned}
\frac{1}{|J(x)|}\sum_{i\in J(x)}\widehat{D}^\Lambda(O_i) &= \frac{1}{|J(x)|}\sum_{i\in J(x)}\sum_{k=1}^{K-1}\left(\widehat{D}^R_k(O_i)+\widehat{D}^S_k(O_i)\right)\cdot\lambda_k[\widehat{P}](O_i)\\
&\underbrace{+\frac{1}{|J(x)|}\sum_{i\in J(x)}\sum_{k=1}^{K-1}\max\left\{0;\ \widehat{R}_k(W_i,x,0)+\widehat{S}_k(W_i,x,1)-1\right\}}_{\widehat{\Lambda}_{\text{plug}}(x)}-\widehat{\Lambda}_{1S}(x)=0,
\end{aligned}\tag{36}$$

where $J(x)$ denotes the subset of indices within J for which $X = x$. By rearranging terms, one gets:

$$\widehat{\Lambda}_{1S}(x) = \widehat{\Lambda}_{\text{plug}}(x) + \frac{1}{|J(x)|}\sum_{i\in J(x)}\sum_{k=1}^{K-1}\left(\widehat{D}^R_k(O_i)+\widehat{D}^S_k(O_i)\right)\cdot\lambda_k[\widehat{P}](O_i). \tag{37}$$

A one-step corrected estimator formulation for the upper bound $\widehat{\Upsilon}_{1S}(x)$ follows analogously.

A.4. Derivation of stabilized one-step corrected estimator

Even when the rules λ_k, υ_k are well-defined by enforcing a strictly positive condition, ambiguities introduced by exceptional laws may cause their estimates to remain unstable, even as the sample size grows indefinitely. If these rules were instead *known* and not estimated from the data, $\Psi[\cdot]$ would be pathwise differentiable at any $P \in \mathfrak{P}$, given certain technical conditions presented in appendix A.3.

We handle the data-dependent nature of the estimated rules by building upon *stabilized one-step correction* (S1S) approach by Luedtke and van der Laan (2016, 2018). This setup enables us to treat a sequence of rule estimates as *known*, allowing a martingale version of the central limit theorem (CLT) to characterize the limiting distribution of the estimator.

To simplify the notation, assume that all functionals, distributions, and samples henceforth are conditioned on a given stratum $X = x$. Let $\Psi[P; \lambda, \upsilon] := (\Lambda[P; \lambda]; \Upsilon[P; \upsilon])^\top$ be the evaluation of $\Psi[P]$ when the rules are fixed at given $\lambda = \{\lambda_k\}_k^{K-1}$ and $\upsilon = \{\upsilon_k\}_k^{K-1}$, with:

$$\Lambda[P; \lambda] := \sum_{k=1}^{K-1} \mathbb{E}_{W|X=x} \left[R_k(w, x, 0) + S_k(w, x, 1) - 1\right] \cdot \lambda_k(w, x), \tag{38}$$

$$\Upsilon[P; \upsilon] := \sum_{k=1}^{K-1} \mathbb{E}_{W|X=x} \left\{R_k(w, x, 0) - \left[R_k(w, x, 0) - S_k(w, x, 1)\right] \cdot \upsilon_k(w, x)\right\}. \tag{39}$$

This reparametrization is consistent in the sense that $\Psi[P^*; \lambda^*, \upsilon^*] = \Psi[P^*]$, where the asterisk indicates true values.

Let P^n denote the empirical distribution of $O_{1:n} = \{O_i\}_{i=1}^n$, with each observation being an independent sample from P^*, and λ^n, υ^n the estimated rules from P^n. Let $\partial\Psi^n(O)$ be the bivariate one-step corrections given by equations (10) and (11) at P^n and observation O. Then, we can express the bias of the corrected estimate $\widehat{\psi}^n$ as follows:

$$\begin{aligned} \widehat{\psi}^n - \Psi[P^*] &= \Psi[P^n; \lambda^n, \upsilon^n] + \partial\Psi^n(O_{n+1}) - \Psi[P^*] \qquad (40) \\ &= \Psi[P^*; \lambda^n, \upsilon^n] + \partial\Psi^n(O_{n+1}) - \Psi[P^*] + \mathrm{Rem}(P^n) - \mathbb{E}^*\left[D^n(O) \mid O_{1:n}\right]. \qquad (41) \end{aligned}$$

Here, $\mathrm{Rem}(P^n) := \Psi[P^n; \lambda^n, \upsilon^n] - \Psi[P^*; \lambda^n, \upsilon^n] + \mathbb{E}^*\left[D^n(O) \mid O_{1:n}\right]$ represents the second-order remainder term from the von Mises expansion of the target at P^n with fixed rules. Under an exceptional law, the variance of the estimated bias, minus the second-order remainder, is unstable, as it fluctuates with each new observation due to underlying ambiguities. However, by fixing the rules based on accumulated observations $O_{1:n}$ and deriving uncertainty from the next observation O_{n+1}, a bivariate martingale structure is introduced. This setup enables the use of a generalized CLT that applies weights determined by the inverse *matrix standard deviation* —the inverse square-root covariance matrix— of each new realization, $T_n = \widehat{\mathrm{Cov}}(\Psi^n(O_n) + \partial\Psi^n(O_n))^{-1/2}$. Hence, for $0 < l < n$ and $M_n = \sum_{j=l}^{n-1} T_j$, one has that, as $(n-l) \to \infty$, $(n-l)^{1/2} M_n^{-1} \sum_{j=l}^{n-1} T_j(\widehat{\psi}^j - \Psi[P^*])$ converges to a bivariate Gaussian distribution with a unit covariance matrix under the following consistency and boundedness conditions (Luedtke and van der Laan, 2016, 2018):

(i) $(n-l)^{-\frac{1}{2}} \sum_{j=l}^{n-1} T_j \,\mathrm{Rem}(P^j) \xrightarrow{\mathbb{P}} (0, 0)$,

(ii) $(n-l)^{-1} \sum_{j=l}^{n-1} \left| T_j^2 \,\mathrm{Cov}\left(\{\Psi^j(O_i) + \partial\Psi^j(O_i)\}_{i=1}^j\right) - \mathrm{Id}_2 \right| \xrightarrow{\mathbb{P}} 0 \cdot \mathrm{Id}_2$,

(iii) $\exists\, \xi < \infty : (n-l)^{-1} \sum_{j=l}^{n-1} \mathbb{P}\left(\left\| T\left(\Psi^j(O_{j+1}) + \partial\Psi^j(O_{j+1})\right)^\top \right\|_2 < \xi \mid O_{1:j-1} \right) \xrightarrow{\mathbb{P}} 1$.

The mean of this limiting distribution is influenced by the asymptotic behavior of $\Psi[P^*; \lambda^n, \upsilon^n] - \Psi[P^*]$. For the lower bound, however, this mean is guaranteed to be less than or equal to zero, as it reflects the gap to the true maximum. To construct *potentially conservative* one-sided confidence intervals from below for both bounds, one can either target the negative of the upper bound or invert the sign of samples for the upper bound component from the asymptotic distribution.

In contrast to the 1S estimator discussed in section 4, the S1S approach eliminates the need for sample splitting. This is due to the martingale process structure, which inherently depends on *fixed* past data and out-of-sample evaluations (van der Laan and Lendle, 2014).

Appendix B. Data details for application case

We assess the benefits from pharmacological treatment with stimulant medication upon the numeracy test performance at grade 8 obtained by Norwegian children diagnosed with ADHD. By integrating information from national registries, we compile data on the medication history and national test scores of all children diagnosed with ADHD born between 2000 and 2007 in Norway, who would go to take the national test up to 2021. We exclude those with severe comorbid disorders and those with missing test scores at grade 5 and impute the missing values at grade 8 (totaling $9\,352$ individuals). Variables at the student, family, and school levels are linked from the Norwegian Prescription Database (NorPD), the Norwegian Patient Registry (NPR), the Database for Control and Payment of Health Reimbursement (KUHR), Statistics Norway (SSB), and the Medical Birth Registry of Norway (MBRN). We leverage data on students' and parents' diagnoses and their consultations with medical services during pre-exposure and post-exposure periods. Indicators of post-exposure medical status serve as proxies for adverse effects of treatment and its consequences. To operationalize relevant variables, we employ the following grouping:

- **pre-exposure covariates** W, X: sex at birth, birth year/month cohorts, birth parity number, raw scores at grade 5 national test for numeracy and reading, missingness indicator for scores at grade 5 English national test, mother's education level, mother's age at birth, student's and parents' diagnoses and medical consultations for related comorbid disorders, school identification (fixed effect), prior dispensations of ADHD stimulant medication for at least 90 days, and duration of prior treatment.
- **Exposure** A: having received dispensations of ADHD stimulant medication for at least 75% of the prescribed treatment period between the start of grade 6 and the national test in grade 8.
- **Outcomes** Y: raw scores at grade 8 national test for numeracy and corresponding official thresholds for tiers 1, 2 and 3.