Estimating Continuous Treatment Effects with Recourse Data

Alessandro Marchese* Vrije Universiteit Brussel
Jeroen Berrevoets King's College London
Niels Martin Universiteit Hasselt
Sam Verboven Vrije Universiteit Brussel

Abstract

Recourse explanations describe what would need to change for a different outcome to occur, yet they are not typically used for treatment effect estimation. We study how such recourse data can be used for causal inference with continuous treatments. This provides counterfactual supervision beyond observed outcomes at realized treatment values, and is particularly useful when routine treatment assignment leaves parts of the treatment domain unsupported, like in diverse applications such as drug dosing, pricing, operations, and recommender systems. We formalize recourse explanations as structural boundary samples and show how they identify continuous dose-response curves under a positivity condition fundamentally distinct from standard overlap. Since recourses are observed only for negative-outcome units, the boundary distribution is left-truncated relative to the population; we correct this selection using a Lynden–Bell inversion. We translate this result into an auxiliary recourse loss for neural dose-response estimators, converting recourse-derived boundaries into counterfactual supervision. Experiments on synthetic and semi-synthetic data show consistent improvements over standard baselines. More broadly, our work shows that recourse explanations can serve as structured causal evidence, expanding what can be learned from observational data.

1 Introduction

Estimating dose-response curves from observational data is essential when randomized experiments are infeasible, yet reliable estimation of the underlying response surface remains difficult under confounding and limited data support [69, 63, 26]. Compensating for this incompleteness has motivated a substantial methodological literature, ranging from structural assumptions on the response [11, 61, 75] to approaches that leverage all available observational signals — including unstructured ones [45, 23, 60] — to strengthen estimation and guide learning toward the stable causal mechanisms that govern the data-generating process [63, 73, 7].

Observational data is routinely accompanied by explanations. In many applied domains, observational data used for dose-response estimation are accompanied by explanations that distill expert or user knowledge about why an outcome occurred and, crucially, what could be changed to alter it. Expert explanations encode domain expertise accumulated through years of practice and direct observation of decision outcomes [56, 41]. In healthcare, clinicians write diagnostic rationales, treatment justifications, discharge summaries, and post-mortem analyses that reflect years of clinical experience and domain knowledge [78, 66, 83, 79]. Alternatively, user explanations originate from the decision subjects themselves, who can report first-hand on the factors that drove their response. In digital marketing and recommender systems, customer feedback, product reviews, and post-interaction surveys provide narrative accounts explaining why a recommendation was ignored or an advertisement failed to convert [76, 16, 88, 59].

*Corresponding author: alessandro.marchese@vub.be

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

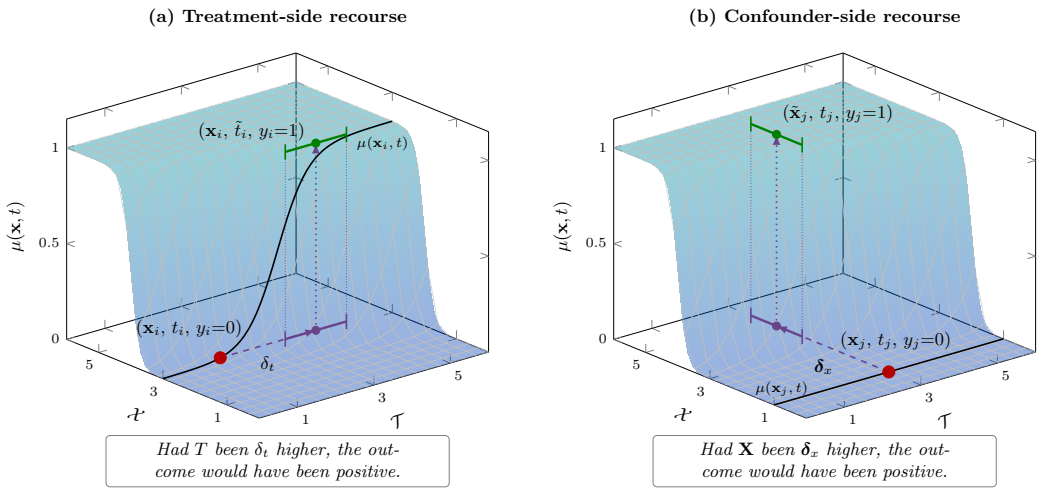

Figure 1: **Illustration of treatment-side vs. confounder-side counterfactual recourse in relation to the dose-response curve. (a)** For individual i, the explanation prescribes a treatment shift δ_t; the black curve shows $\mu(\mathbf{x}_i, t)$ rising along the treatment dimension $\mathcal{T}$, and the recourse moves i from $y_i = 0$ to $Y_i(\tilde{t}_i) = 1$. The shifted dose $\tilde{t}_i$ pins a point of the surface at the factual covariates. **(b)** For individual j, $\mu(\mathbf{x}_j, t) = 0$ throughout $\mathcal{T}$, so no treatment shift would yield $Y_j(\tilde{t}_j) = 1$ and recourse requires a shift $\boldsymbol{\delta}_x$ along the confounder dimensions $\mathcal{X}$ instead. The transition point then sits at $\tilde{\mathbf{x}}_j$, carrying no boundary information about j's own curve, only the certificate that none exists within $\mathcal{T}$.

Expert and user explanations often contain counterfactual recourse. Beyond describing the status quo, explanations frequently prescribe actionable modifications under which the outcome would have differed [56]. This type of explanation is known as counterfactual recourse [87, 43, 42]: a minimal-cost intervention on an individual's attributes such that, had the intervention been applied, the outcome would have changed. Importantly, unlike other types of explanations, counterfactual recourses allow for an *interventional* interpretation [43, 44, 13]. A clinician explaining a failed treatment may note that a higher dosage or an alternative drug combination would have been effective; a customer explaining why they ignored a recommendation may note that the product had undesirable features or the advertisement was irrelevant to their needs. In each case, the explanation implicitly specifies a conceptually nearby point on the response surface at which the outcome flips.

Explanatory supervision improves generalization performance. Explanatory guidance seeks to improve reasoning alignment by incorporating human-provided rationales during training, encouraging models to align their predictions with semantically meaningful factors rather than fitting purely observational correlations [70, 29, 91]. By enforcing consistency between model behavior and expert reasoning, these methods reduce reliance on spurious associations and promote invariance across environments, while also improving predictive performance and sample efficiency [24, 54, 2]. From a causal perspective, explanations act as soft constraints that guide learning toward stable mechanisms — the type of cause-effect relationships that persist under changes in decision policies or covariate distributions [63, 30, 73]. This makes explanatory guidance a natural complement to treatment effect estimation: rationales encode precisely the domain or first-hand knowledge needed to distinguish the stable response relationship from environment-specific artifacts.

This setting inverts aspects of subgroup discovery and recourse generation. A long-standing goal in treatment effect estimation is to identify *interpretable* subgroups of the population that exhibit homogeneous treatment responses [3, 28]. Partitioning methods address this by splitting the covariate space into axis-aligned regions where treatment effects are similar within groups and different across groups, yielding representations that domain experts can act on [82, 4, 5, 48, 95]. A parallel literature asks how to *produce* such prescriptions, generating recourse that is faithful to the underlying causal mechanism and feasible to enact, rather than merely flipping a predictor's decision [84, 43, 52, 64, 42]. Conversely, recourse explanations directly encode interpretable response-surface structure by specifying the modification under which an outcome would change. Rather than discovering that structure from data, we exploit it to improve estimation of the response surface itself; and rather than generating recourse from a known response surface, we treat observed recourse as evidence about an unknown one.

Contributions. The setting poses structural challenges that distinguish it from standard dose-response estimation, and our contributions address them in turn:

Recourse formalization. Explanations encode unit-level transition points rather than outcomes at observed doses, and vary in kind. We formalize them as structural boundaries of the response mechanism, and show that treatment-side recourses pin those boundaries at the unit's own covariates while confounder-side recourses certify non-responder status.

Alternative identification. Recourses are observed only on negative-outcome units, which truncates the boundary distribution and biases any estimator that ignores it. We invert this truncation and prove identification of the conditional average dose-response under an *alternative* positivity condition to pointwise overlap.

Auxiliary recourse loss. Turning boundaries into supervision a dose-response learner can consume requires reformulating the regression target. We do so as an auxiliary loss compatible with any neural continuous-treatment estimator, supplying structurally faithful supervision where standard adjustment fails.

Empirical validation. We validate the method on synthetic data and on semi-synthetic data with registry-grounded recourse annotations, showing consistent improvements over baselines.

2 Estimating the Dose-Response Curve: Problem Formulation

We study Conditional Average Dose Response (CADR) estimation from observational data in which each negative unit is accompanied by a structured explanation encoding a counterfactual scenario consistent with the underlying outcome mechanism.

Structural causal model. Let $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^{d_x}$ denote pre-treatment confounders, $T \in \mathcal{T} = [T_{\min}, T_{\max}] \subset \mathbb{R}$ a continuous treatment, and $Y \in \{0, 1\}$ a binary outcome. We adopt the structural causal model (SCM) framework [62] and assume the data-generating process is governed by

$$\mathbf{X} = f_X(U_X), \quad T = f_T(\mathbf{X}, U_T), \quad Y = f_Y(\mathbf{X}, T, U_Y), \tag{1}$$

where U_X, U_T, U_Y are mutually independent exogenous noise variables capturing all variation not explained by the endogenous parents. The implied causal graph $\mathcal{G}$ is $\mathbf{X} \to T \to Y,\ \mathbf{X} \to Y$.

Counterfactual recourse.[2] Each observation $(\mathbf{x}, t, y)$ with $y = 0$ is accompanied by a counterfactual recourse $\boldsymbol{\delta} = (\boldsymbol{\delta}_x,\ \delta_t) \in \mathbb{R}^{d_x} \times \mathbb{R}$ obtained as the solution to the minimal-cost recourse problem [43, 42, 87], where $\mathrm{cost}\colon \mathbb{R}^{d_x} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ quantifies the effort or feasibility of implementing $\boldsymbol{\delta}$:

$$\boldsymbol{\delta} \ = \ \underset{\boldsymbol{\delta}_x,\, \delta_t}{\arg\min}\ \mathrm{cost}(\boldsymbol{\delta}_x,\, \delta_t) \quad \text{s.t.} \quad f_Y(\tilde{\mathbf{x}},\, \tilde{t},\, u_Y) = 1, \quad \tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}_x \in \mathcal{X}, \quad \tilde{t} = t + \delta_t \in \mathcal{T}, \tag{2}$$

where u_Y is held at the factual noise realization of the individual, giving (2) its counterfactual semantics. The counterfactual state $(\tilde{\mathbf{x}}, \tilde{t})$ constitutes a *transition point* of f_Y: the minimal-cost point at which the dose-response surface crosses from $Y = 0$ to $Y = 1$ for that individual. We assume the feasible set $\{\boldsymbol{\delta} : f_Y(\tilde{\mathbf{x}}, \tilde{t}, u_Y) = 1\}$ is non-empty for all individuals with $y = 0$.

A recourse $\boldsymbol{\delta}$ prescribes a *joint* intervention $do(\mathbf{X} := \mathbf{x} + \boldsymbol{\delta}_x,\ T := t + \delta_t)$, which by standard do-calculus operates on the mutilated graph $\mathcal{G}_{\overline{\{X,T\}}}$ with all incoming edges to $\mathbf{X}$ and T removed, including $\mathbf{X} \to T$. Consequently, $\boldsymbol{\delta}_x$ cannot affect T and influences Y only via the direct path $\mathbf{X} \to Y$. A visual representation of such recourse operations is shown in Figure 1.

Treatment-preferential recourse. We further assume that interventions on the treatment are preferred over interventions on the confounders:

$$\mathrm{cost}(\boldsymbol{\delta}_x,\, 0) \ \gg \ \mathrm{cost}(\mathbf{0},\, \delta_t) \ > \ 0, \quad \text{for all } \boldsymbol{\delta}_x \neq \mathbf{0}\ ,\ \delta_t \neq 0 \tag{3}$$

so that the recourse in (2) exhausts treatment-side modifications before altering confounders[3]. When the recourse is generated by a human, the assumed cost structure is supported by a well-documented tendency to mutate controllable features when reasoning counterfactually [32, 13, 56]. Additionally, this reflects settings where the treatment is more readily actionable than the confounders or where the elicitation protocol itself prioritizes treatment explanations.

[2] We further use the terms recourse and explanation interchangeably to refer to the joint intervention and its representation.

[3] Additionally, one can assume $\mathrm{cost}(\boldsymbol{\delta}_x,\, \delta_t) > \mathrm{cost}(\boldsymbol{\delta}_x,\, 0)$ or prioritize sparsity for mixed recourses.

Data and explanations. We assume $\boldsymbol{\delta}$ to be directly observed in structured form for every unit with a negative outcome. We further assume access to an observational dataset $\mathcal{D}_{\text{obs}} = \{(\mathbf{x}_i, t_i, y_i, \boldsymbol{\delta}_i)\}_{i=1}^n$ drawn i.i.d., where $\boldsymbol{\delta}_i$ is observed iff $y_i = 0$.

Goal: Conditional Average Dose-Response (CADR). Our objective is to estimate the causal dose-response surface

$$\mu(\mathbf{x},\, t) \;=\; \mathbb{E}\big[Y(t) \,\big|\, \mathbf{X} = \mathbf{x}\big], \tag{4}$$

where $Y(t)$ denotes the potential outcome under $do(T := t)$. The quantity $\mu(\mathbf{x}, \cdot)$ is the *conditional dose-response curve* at covariate level $\mathbf{x}$, also called the conditional average dose response (CADR) or heterogeneous dose response [11, 61, 75, 12]. Contrasts $\tau(\mathbf{x}, t, t') = \mu(\mathbf{x}, t') - \mu(\mathbf{x}, t)$ recover the effect between two doses as a special case; for binary $\mathcal{T} = \{0, 1\}$, $\tau(\mathbf{x}, 0, 1)$ is the conditional average treatment effect (CATE) [71, 39].

Identification. Estimation of $\mu(\mathbf{x}, t)$ from observational data requires the following standard assumptions [72, 69, 35, 38]:

- **(A1) Positivity (overlap).** $p(T = t \mid \mathbf{X} = \mathbf{x}) > 0$ for all $t \in \mathcal{T}$ and $\mathbf{x} \in \mathcal{X}$ with $p(\mathbf{x}) > 0$. *Every treatment value has positive density at every covariate value, so counterfactuals are observationally supported.*
- **(A2) Unconfoundedness.** $Y(t) \perp\!\!\!\perp T \mid \mathbf{X}$ for all $t \in \mathcal{T}$. *Observed covariates* $\mathbf{X}$ *block all backdoor paths between treatment and outcome, leaving no unmeasured confounding.*
- **(A3) Consistency.** $Y = Y(T)$. *The factual outcome equals the potential outcome under the treatment actually received — no interference, no version-of-treatment ambiguity.*

Under **(A1)**–**(A3)**, the dose-response surface is identified via the adjustment formula $\mu(\mathbf{x}, t) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}, T = t]$; **(A2)** follows directly from (1), since $U_Y \perp\!\!\!\perp (\mathbf{X}, T)$. In the continuous-treatment setting **(A1)** is particularly stringent [74, 100, 68], requiring positive density, not just positive mass, at every evaluation dose. This motivates Section 3, where recourse-derived boundary information supplies what is missing at doses where **(A1)** fails.

3 Identification from Counterfactual Recourse

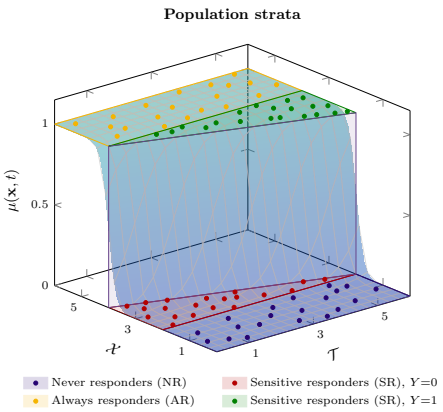

Figure 2: **The three response strata.** Units with Y=0 lie on the μ=0 plane, units with Y=1 on μ=1. The boundary (purple) is the μ=$\frac{1}{2}$ level set; where it exits $\mathcal{T}$ it separates NR and AR from the SR band. Strata are drawn at the median boundary, and are exact in the noiseless limit.

We show that recourse-derived boundary information yields an identification strategy that bypasses the adjustment formula $\mathbb{E}[Y \mid \mathbf{X}, T]$ and admits a fundamentally distinct alternative to the positivity assumption **(A1)**. Because $\mathcal{T}$ is bounded, the population splits into three response strata — *never-*, *always-* and *sensitive-responders* — of which only the last has a non-degenerate dose-response curve. Our main result identifies that curve by inverting the truncation that the negative-outcome stratum induces on the observed recourses; the other two strata contribute structural constants, and the three recombine into the full-population CADR. We develop the continuous single-threshold case below, defer other cases to Appendix B, and summarize all strata in Appendix A.

Three response strata. The success region $\mathcal{R}_1(\mathbf{x}, u_Y) = \{t \in \mathcal{T} : f_Y(\mathbf{x}, t, u_Y) = 1\}$ determines the response surface via $\mu(\mathbf{x}, t) = P(t \in \mathcal{R}_1(\mathbf{x}, U_Y))$. We develop the main result in the scalar threshold-crossing case,

$$f_Y(\mathbf{x}, t, u_Y) \;=\; \mathbf{1}\{t \geq g(\mathbf{x}, u_Y)\}, \tag{5}$$

with U_Y continuously distributed and $g(\mathbf{x}, \cdot)$ non-degenerate [58, 86], so that $\mathcal{R}_1(\mathbf{x}, u_Y) = [B, \infty) \cap \mathcal{T}$ for the *boundary* random variable $B := g(\mathbf{X}, U_Y)$. Throughout we assume the recourse is *structurally faithful* **(F)**[4]: $\tilde{t}_i \in \partial\mathcal{R}_1(\tilde{\mathbf{x}}_i, u_{Y,i})$, so that a treatment-side recourse reveals the unit's own

[4]Faithfulness is a property of the recourse-optimization problem (2): it assumes the explanation generating mechanism — typically a human — is faithful to the outcome generating mechanism f_Y.

boundary, $b_i = t_i + \delta_{t,i}$. The position of B relative to $\mathcal{T} = [T_{\min}, T_{\max}]$ then fully determines a unit's potential-outcome profile, defining the stratum indicator S (Figure 2):

$$S = \mathrm{NR} \ \text{ if } B > T_{\max}, \qquad S = \mathrm{SR} \ \text{ if } T_{\min} < B \le T_{\max}, \qquad S = \mathrm{AR} \ \text{ if } B \le T_{\min}. \tag{6}$$

Writing $\mu^s(\mathbf{x}, t) := \mathbb{E}[Y(t) \mid \mathbf{x},\ S = s]$, the never- and always-responder curves are the structural constants $\mu^{\mathrm{NR}} \equiv 0$ and $\mu^{\mathrm{AR}} \equiv 1$, while

$$\mu^{\mathrm{SR}}(\mathbf{x}, t) = P(B \le t \mid \mathbf{x}, \mathrm{SR}) = F_{B|\mathbf{x},\mathrm{SR}}(t). \tag{7}$$

Only the sensitive curve depends on the treatment, and it is the one the recourse data identify. That stratum moreover announces itself: a negative outcome with a treatment-only recourse certifies $S = \mathrm{SR}$, since feasibility in (2) gives $b_i \le T_{\max}$ while $y_i = 0$ gives $b_i > t_i \ge T_{\min}$, so the conditioning events below are observable.

Left-truncation structure of observed boundaries. Boundary samples are observed only for units in the $\{\mathrm{SR}, Y = 0\}$ stratum: $y_i = 0$ enforces $b_i > t_i$, and $S_i = \mathrm{SR}$ ensures the unit i is a sensitive responder. A boundary value is observable only when the factual treatment falls below it, so thresholds with more mass below them are overrepresented [50, 94]. Conditioning on $Y = 0$ opens the collider $U_Y \to Y \leftarrow T$ in (1): although $B \perp T \mid \mathbf{X}$ marginally (via $U_Y \perp T \mid \mathbf{X}$), the conditional boundary distribution becomes left-truncated at the factual treatment. The joint density at $\mathbf{X} = \mathbf{x}_0$ factorizes as

$$p(t, b, Y = 0 \mid \mathbf{x}_0, \mathrm{SR}) = p(t \mid \mathbf{x}_0)\, f_{B|\mathbf{x}_0,\mathrm{SR}}(b)\, \mathbf{1}\{b > t\}, \tag{8}$$

and marginalizing over t yields the observed boundary density

$$p(b \mid \mathbf{x}_0, Y = 0, \mathrm{SR}) = \frac{f_{B|\mathbf{x}_0,\mathrm{SR}}(b)\, F_{T|\mathbf{x}_0}(b^-)}{P(Y = 0 \mid \mathbf{x}_0, \mathrm{SR})} \tag{9}$$

where $F_{T|\mathbf{x}_0}(b^-) := P(T < b \mid \mathbf{X} = \mathbf{x}_0)$. The boundary distribution under $\{Y = 0\}$ is the untruncated $f_{B|\mathbf{x}_0,\mathrm{SR}}$ multiplied by the treatment lower CDF — a selection invertible whenever T has sufficient support below the evaluation point. $F_{T|\mathbf{x}_0}(b^-)$ is the Lynden–Bell selection kernel [50].

One-sided cumulative positivity. This motivates replacing **(A1)** with a CDF condition:

(A1′) **Recourse positivity.**[5] For the evaluation covariate $\mathbf{x}_0$, $F_{T|\mathbf{x}_0}(b^-) > 0$ for every $b \in \mathrm{supp}(B \mid \mathbf{x}_0, \mathrm{SR})$.

(A1′) is fundamentally different than **(A1)**: it requires treatment mass *somewhere* below each boundary in the sensitive support, rather than positive density *at* the evaluation dose t_0, which may lie outside $\mathrm{supp}(T \mid \mathbf{x}_0)$ entirely.

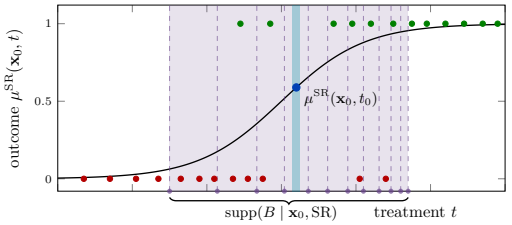

Figure 3: **Identification with an interior gap in treatment support.** Purple ticks show recourse-revealed thresholds b_i; t_0 falls where no treatment is observed, so **(A1)** fails. Under **(A1′)**, the cumulative mass $F_{T|\mathbf{x}_0}(b^-)$ is positive on $\mathrm{supp}(B \mid \mathbf{x}_0, \mathrm{SR})$ (bracket), and (10) identifies $\mu^{\mathrm{SR}}(\mathbf{x}_0, t_0)$ by inverting the left-truncation the $\{Y = 0\}$ stratum induces. A threshold is revealed with probability $F_{T|\mathbf{x}_0}(b^-)$, so high ones are over-represented.

Proposition 1 (Identification under Recourse positivity)**.** *Assume the SCM in* (1)*, treatment-preferential cost* (3)*, faithfulness* ***(F)****, threshold-model* (5) *and one-sided positivity* ***(A1′)****. Working with the treatment-side recourse data, the sensitive CADR at* $\mathbf{x}_0$ *is identified by*

$$\mu^{\mathrm{SR}}(\mathbf{x}_0, t_0) \;=\; P(Y = 0 \mid \mathbf{x}_0, \mathrm{SR}) \int_{T_{\min}}^{t_0} \frac{p(b \mid \mathbf{x}_0, Y = 0, \mathrm{SR})}{F_{T|\mathbf{x}_0}(b^-)}\, db, \tag{10}$$

for $t_0 \in \mathcal{T}$*. Every quantity on the right is identified from the observational distribution: the prefactor is fixed by total mass,* $P(Y{=}0 \mid \mathbf{x}_0, \mathrm{SR})^{-1} = \int_{T_{\min}}^{T_{\max}} p(b \mid \mathbf{x}_0, Y{=}0, \mathrm{SR})/F_{T|\mathbf{x}_0}(b^-)\, db$*, since* $\mathrm{supp}(B \mid \mathbf{x}_0, \mathrm{SR}) \subseteq \mathcal{T}$*. In particular,* (10) *holds even when* $p(t_0 \mid \mathbf{x}_0) = 0$*.*

[5] The Lynden–Bell estimator [50] requires a support-containment condition for identifiability of the underlying distribution, later formalized in [94]. Our **(A1′)** is its conditional analogue at the evaluation covariate.

Proof. (8) follows from $U_Y \perp T \mid \mathbf{X}$ in (1) and $B = g(\mathbf{X}, U_Y)$ under **(F)**. Marginalizing over t yields (9). By **(A1′)**, $F_{T|\mathbf{x}_0}(b^-) > 0$ on $\mathrm{supp}(B \mid \mathbf{x}_0, \mathrm{SR})$, so (9) inverts to

$$f_{B|\mathbf{x}_0,\mathrm{SR}}(b) \;=\; \frac{p(b \mid \mathbf{x}_0, Y=0, \mathrm{SR})\, P(Y=0 \mid \mathbf{x}_0, \mathrm{SR})}{F_{T|\mathbf{x}_0}(b^-)}. \tag{11}$$

Integrating over $(T_{\min}, t_0]$ and using (7) gives (10); integrating instead over all of $\mathcal{T}$, where $f_{B|\mathbf{x}_0,\mathrm{SR}}$ has total mass one, gives the prefactor. □

Full-population CADR. The remaining two strata contribute no curve to estimate, only a proportion $\pi_s(\mathbf{x}_0) := P(S = s \mid \mathbf{x}_0)$. Never-responders are certified by their recourse type: under treatment-preferential cost (3), a unit with $y_i = 0$ and $\boldsymbol{\delta}_{x,i} \neq \mathbf{0}$ must have $b_i > T_{\max}$, since otherwise the cheaper shift $(\mathbf{0}, b_i - t_i)$ would have been feasible. Sensitive responders are certified by a treatment-only recourse, and dividing that proportion by the truncation prefactor of (10) recovers the stratum as a whole:

$$\pi_{\mathrm{NR}}(\mathbf{x}_0) = P\big(Y=0,\, \boldsymbol{\delta}_x \neq \mathbf{0} \mid \mathbf{x}_0\big), \qquad \pi_{\mathrm{SR}}(\mathbf{x}_0) = \frac{P\big(Y=0,\, \boldsymbol{\delta}_x = \mathbf{0} \mid \mathbf{x}_0\big)}{P(Y=0 \mid \mathbf{x}_0, \mathrm{SR})}. \tag{12}$$

Always-responders produce neither a negative outcome nor a recourse, so no observable event isolates them as, in the data, they are pooled with the sensitive responders who happened to receive a dose above their own boundary. By the law of total probability over the partition (6), however, $\pi_{\mathrm{AR}} = 1 - \pi_{\mathrm{SR}} - \pi_{\mathrm{NR}}$, so their mass is identified as a residual and never has to be modelled directly. The law of total expectation over (6) then gives

$$\mu(\mathbf{x}_0, t_0) \;=\; \pi_{\mathrm{SR}}(\mathbf{x}_0)\, \mu^{\mathrm{SR}}(\mathbf{x}_0, t_0) \;+\; \pi_{\mathrm{AR}}(\mathbf{x}_0), \tag{13}$$

and substituting the residual together with (12)–(10) cancels the prefactor, leaving every term a proportion of an observable event:

$$\mu(\mathbf{x}_0, t_0) \;=\; 1 - P\big(Y=0,\, \boldsymbol{\delta}_x \neq \mathbf{0} \mid \mathbf{x}_0\big) - P\big(Y=0,\, \boldsymbol{\delta}_x = \mathbf{0} \mid \mathbf{x}_0\big) \int_{t_0}^{T_{\max}} \frac{p(b \mid \mathbf{x}_0, Y=0, \mathrm{SR})}{F_{T|\mathbf{x}_0}(b^-)}\, db. \tag{14}$$

Everyone succeeds at t_0 except the never-responders and the sensitive units whose boundary still lies above it; at $t_0 = T_{\max}$ the integral vanishes and $\mu(\mathbf{x}_0, T_{\max}) = 1 - \pi_{\mathrm{NR}}(\mathbf{x}_0)$, the maximal attainable response at $\mathbf{x}_0$. The stratum curves $\mu^{\mathrm{NR}} \equiv 0$ and $\mu^{\mathrm{AR}} \equiv 1$ are structural constants requiring no condition at all, and π_{NR} needs only $p(\mathbf{x}_0) > 0$; the treatment mechanism enters only through **(A1′)**, which the sensitive term and — via the residual — the always-responder mass both rest on. Consequently (14) identifies the CADR at evaluation doses that carry no treatment mass at all, where **(A1)** fails outright, and the gain is largest exactly where routine assignment rarely realizes the evaluation dose.

4 Learning Dose-Response Surfaces from Recourse Data

Equation (14) writes $\mu(\mathbf{x}_0, t_0)$ in terms of two proportions of events visible in $\mathcal{D}_{\mathrm{obs}}$ — units carrying a confounder-side and a treatment-side recourse — together with a Lynden–Bell reweighted integral over the revealed boundaries. We therefore introduce the two per-unit recourse-type indicators

$$\mathrm{TR}_i \;=\; \mathbf{1}\{y_i = 0 \;\wedge\; \boldsymbol{\delta}_{x,i} = \mathbf{0}\}, \qquad \mathrm{CR}_i \;=\; \mathbf{1}\{y_i = 0 \;\wedge\; \boldsymbol{\delta}_{x,i} \neq \mathbf{0}\}, \tag{15}$$

with $\mathrm{TR}_i = 1$ certifying $S_i = \mathrm{SR}$ and supplying the boundary $b_i = t_i + \delta_{t,i}$, and $\mathrm{CR}_i = 1$ certifying $S_i = \mathrm{NR}$. Units with $y_i = 1$ carry neither: they are a mixture of AR and successful SR units that no observable event separates. A single learner $\hat{\mu}_\theta(\mathbf{x}, t)$ is trained on all three groups together with the factual outcomes.

From identification to a regression target. Because the always-responder mass enters (14) only as a residual, it cannot be attached to any individual unit as a label; the identified curve is a population-level difference rather than a conditional average over the recourse subsample. We therefore target it with a *pseudo-outcome*. Attaching the Lynden–Bell weight

$$w(\mathbf{x}, b) \;=\; \frac{1}{F_{T|\mathbf{x}}(b^-)} \tag{16}$$

to each revealed boundary, define for every unit i and every $t \in \mathcal{T}$

$$\tilde{y}_i(t) \;=\; 1 \;-\; \mathrm{CR}_i \;-\; \mathrm{TR}_i \cdot w_i \cdot \mathbf{1}\{b_i > t\}. \tag{17}$$

Taking conditional expectations term by term and applying (12) and the inversion (11) recovers exactly (14):

$$\mathbb{E}\big[\tilde{y}(t) \,\big|\, \mathbf{X} = \mathbf{x}\big] \;=\; 1 - \pi_{\mathrm{NR}}(\mathbf{x}) - \pi_{\mathrm{SR}}(\mathbf{x})\big[1 - \mu^{\mathrm{SR}}(\mathbf{x}, t)\big] \;=\; \mu(\mathbf{x}, t), \tag{18}$$

so $\tilde{y}_i(t)$ is an unbiased signal for the *full-population* CADR, and a squared-error regression of $\hat{\mu}_\theta$ on it has μ as its population minimizer. The three strata enter through their certificates alone: confounder-side recourses subtract a constant, treatment-side recourses subtract the reweighted mass that would still fail at t, and always-responders are absorbed into the leading 1 without ever being identified individually.

Counterfactual-treatment formulation. A direct regression at a fixed t_0 would target one query at a time. Instead we exploit threshold-crossing more aggressively: by (5), $\tilde{y}_i(t)$ is defined at *every* t once b_i is known, so sampling counterfactual treatments $t_j^{\mathrm{cf}} \sim \mathrm{Unif}(T_{\min}, T_{\max})$ supervises $\mu(\mathbf{x}, \cdot)$ across the whole range in a single pass. For $\{\mathrm{CR}_i = 1\}$ units the pseudo-outcome is the constant 0 and for $\{y_i = 1\}$ units the constant 1, in both cases independent of t_j^{cf}; only the treatment-side units carry a t-dependent signal, and only they require the weight.

The auxiliary recourse loss. Every unit in the batch contributes a counterfactual pseudo-sample, with the pseudo-outcome determined by its recourse type:

$$\tilde{y}_{ij} = \begin{cases} 1 - \hat{w}_i\, \mathbf{1}\{t_j^{\mathrm{cf}} < b_i\}, & \mathrm{TR}_i = 1 \quad \text{(sensitive)}, \\ 0, & \mathrm{CR}_i = 1 \quad \text{(never-responder)}, \\ 1, & y_i = 1 \quad \text{(always- or successful sensitive)}. \end{cases} \tag{19}$$

The auxiliary loss averages the squared error of $\hat{\mu}_\theta$ against these pseudo-outcomes over k counterfactual draws per unit,

$$\mathcal{L}_{\mathrm{Rec}}(\theta) = \frac{1}{n\,k} \sum_{i=1}^{n} \sum_{j=1}^{k} \big(\hat{\mu}_\theta(\mathbf{x}_i, t_j^{\mathrm{cf}}) - \tilde{y}_{ij}\big)^2, \tag{20}$$

and the total objective combines it with the learner's factual loss,

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{Base}}(\theta) + \lambda \cdot \mathcal{L}_{\mathrm{Rec}}(\theta). \tag{21}$$

By (18) the pointwise minimizer of $\mathcal{L}_{\mathrm{Rec}}$ at $(\mathbf{x}_0, t_0)$ is $\mu(\mathbf{x}_0, t_0)$ as $\hat{w}$ converges to w, so $\mathcal{L}_{\mathrm{Rec}}$ supplies structurally faithful supervision throughout $\mathbf{x}$-space wherever **(A1′)** holds — including evaluation doses at which $\mathcal{L}_{\mathrm{Base}}$ has no factual outcomes to learn from.

Implementation notes. The weight $\hat{w}_i = 1/\hat{F}_{T|\mathbf{x}_i}(b_i^-)$ requires a single nuisance component, the conditional treatment CDF, fit on the observational sample; plug-in weights are clipped at the 95th percentile [17, 18], trading variance against a bias that vanishes as the clip level rises. For a decreasing response the strata mirror, $F_{T|\mathbf{x}}(b^-)$ becomes $1 - F_{T|\mathbf{x}}(b)$, and the indicator flips. Counterfactual draws are resampled each epoch; λ cannot be identified from data and must be set from domain knowledge. Algorithm 1 operationalizes the minimization of the objective $\mathcal{L}(\theta)$ as a training step.

5 Related Work

Continuous-treatment dose-response estimation. Classical work relies on the generalized propensity score and IPW under pointwise positivity [35, 38]. Modern neural estimators extend this lineage: DRNet uses bin-based heads with shared representation [75]; VCNet uses treatment-conditioned varying coefficients [61]; TransTEE attends over covariate-treatment tokens [99]; ADMIT balances representations via integral probability metrics [89]; DRVAE embeds the dose-response in a variational latent space [19]; A recent line formalizes identification and diagnostics when positivity fails or treatment support is restricted [74, 100, 68, 6, 12]. All of these methods rely on observational support at the evaluation dose; recourse data instead identifies the full-population CADR wherever recourse positivity **(A1′)** holds, and our auxiliary loss makes that route available to any of them.

Causal effect estimation with natural-language data. A recent line of work integrates language models into the estimation pipeline as *auxiliary machinery* for surrogate generation, imputation, augmentation, or representation [51, 23, 37, 36]; across the treatment effect literature, text takes several roles in the causal model: (a) text as *treatment* variable [90, 65, 93, 53]; (b) text as *mediator* [46, 85, 33, 98]; (c) text as *outcome* [31, 80, 47, 57]; and (d) text as *confounder* [92, 97, 15, 22]. In contrast to all of the above, our setting proposes using *explanations* the response mechanism, which are often available in textual format. Each explanation acts as a structural artifact: it either pins a boundary point of the unit-level success region or certifies that the region is empty, and under faithfulness the two together yield a new identification route for the dose-response surface.

6 Experiments

In the following section we evaluate our methodology empirically by combining the auxiliary loss with different CADR base learners. All code used for experiments is available at https://github.com/AlessandroMarchese/Recourse-CADR.

Experimental setup. Similarly to previous work in treatment effect estimation [1, 34, 40, 20], we construct a synthetic threshold-crossing data-generating process that gives us ground-truth control over $\mu(\mathbf{x}, t)$ and the recourse-positive region. Covariates $\mathbf{X} \in \mathbb{R}^d$ are sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and the threshold $g(\mathbf{x}, u_Y)$ is generated as a non-linear function of $\mathbf{x}$ with noise u_Y. The factual treatment T is drawn from a confounded distribution truncated to $[T_{\min}, T_{\max}]$, with the truncation chosen to leave $\mathrm{supp}(B \mid \mathbf{x})$ partially uncovered — the regime in which standard pointwise positivity fails but recourse positivity **(A1′)** holds. The factual outcome is $Y = \mathbf{1}\{T \geq g(\mathbf{x}, u_Y)\}$. For each $Y = 0$ unit, we generate a recourse explanation by solving (2) under treatment-preferential cost (3): when feasible, the recourse is treatment-only ($\mathrm{TR}_i = 1$, $b_i = t_i + \delta_{t,i}$). The resulting dataset contains observational data, subject to treatment truncation. Finally, an independent unbiased evaluation set is sampled with $T \sim \mathrm{Unif}(T_{\min}, T_{\max})$ on which we measure error against ground-truth $\mu(\mathbf{x}, t)$. Further details on the data generating process are available in Appendix E.1.

Algorithm 1: Recourse-supervised CADR estimation: single training step

Input: batch $\{(\mathbf{x}_i, t_i, y_i, \mathrm{TR}_i, \mathrm{CR}_i, b_i)\}_{i=1}^n$; LB weights $\{\hat{w}_i\}_{i:\mathrm{TR}_i=1}$; range $[T_{\min}, T_{\max}]$; weight λ; CF draws k
Init: learner $\hat{\mu}_\theta(\mathbf{x}, t)$

Stage ① : *Base learner loss*
$\hat{y}_i \leftarrow \hat{\mu}_\theta(\mathbf{x}_i, t_i)$ **for all** i
$\mathcal{L}_{\text{Base}} \leftarrow \mathcal{L}_{\text{Base}}(\{y_i\}, \{\hat{y}_i\})$

Stage ② : *Counterfactual pseudo-outcome loss*
$\mathcal{L}_{\text{Rec}} \leftarrow 0$
for *each* $i = 1$ **to** n **do**
  for $j = 1$ **to** k **do**
    ▷ `Sample counterfactual treatment`
    $t_j^{\text{cf}} \sim \mathrm{Unif}(T_{\min}, T_{\max})$
    if $\mathrm{TR}_i = 1$ **then**
      ▷ `Treatment-side: SR`
      ▷ `Lynden-Bell weighting`
      $\tilde{y}_{ij} \leftarrow 1 - \hat{w}_i \cdot \mathbf{1}\{b_i > t_j^{\text{cf}}\}$
    else if $\mathrm{CR}_i = 1$ **then**
      ▷ `Confounder-side: NR`
      $\tilde{y}_{ij} \leftarrow 0$
    else
      ▷ `Positive outcome: AR or SR`
      $\tilde{y}_{ij} \leftarrow 1$
    $\mathcal{L}_{\text{Rec}} \mathrel{+}= \left(\hat{\mu}_\theta(\mathbf{x}_i, t_j^{\text{cf}}) - \tilde{y}_{ij}\right)^2$
$\mathcal{L}_{\text{Rec}} \leftarrow \mathcal{L}_{\text{Rec}}/(n \cdot k)$

Stage ③ : *Joint update*
$\mathcal{L} \leftarrow \mathcal{L}_{\text{Base}} + \lambda \cdot \mathcal{L}_{\text{Rec}}$
$\theta \leftarrow \mathtt{Update}(\nabla_\theta \mathcal{L})$
Output: updated parameters θ

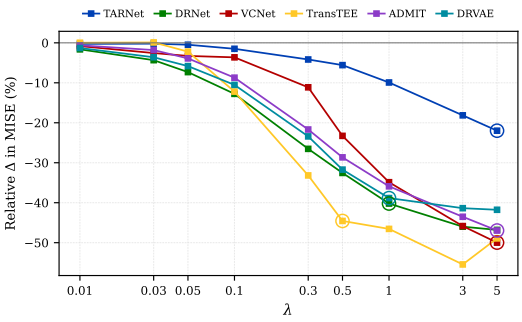

Figure 4: **Bias corrected via Lynden-Bell reweighting.** The difference in MISE is shown for each estimator when Lynden–Bell weights are used in the auxiliary loss $\mathcal{L}_{\text{Rec}}$ compared to no reweighting of the recourse-derived boundaries. The difference is shown across different values for the recourse loss weight λ. Circled points are where λ is tuned for an optimal MISE.

UNOS-PTR organ offer data. The United Network for Organ Sharing (UNOS) Potential Transplant Recipient (PTR) records document organ offers made to candidate patients and their medical teams, the binary acceptance decision, and—if refused—the refusal reason [14]. The UNOS dataset has been used extensively in prior treatment-effect work [10, 8, 9, 96, 55, 54]. We estimate the CADR of liver-offer acceptance with respect to *cold ischemic time* (CIT): the duration between donor cross-clamp and recipient transplantation during which the organ is preserved cold. CIT significantly affects organ

Table 1: **Performance metrics on conditional average dose-response estimation.** Performances of baseline dose-response models are compared with our Recourse and Lynden-Bell adaptations. Performance is measured on MISE, AMSE and DPE metrics on synthetic and semi-synthetic data; entries are means with standard deviations over 10 seeds.

Model	Synthetic data			UNOS-PTR data		
	MISE	AMSE	DPE	MISE	AMSE	DPE
Recourse-only baseline	.0900 .0130	.0451 .0072	.0025 .0078	.5662 .0962	.4534 .0794	**.0000** .0000
+ Lynden-Bell	**.0466** .0146	**.0052** .0071	**.0020** .0058	**.5172** .0951	**.4282** .0865	**.0000** .0000
TARNET [77]	.4305 .0504	.1746 .0590	.2035 .0166	1.1628 .0954	1.0196 .0835	.3871 .0980
+ Recourse	.1041 .0089	.0665 .0053	**.0856** .0015	1.1329 .0943	.9900 .0818	.3724 .1153
+ Lynden-Bell	**.0866** .0049	**.0488** .0026	**.0856** .0015	**1.1043** .0860	**.9808** .0784	**.3607** .1177
DRNET [75]	.4076 .0421	.1625 .0446	.0946 .0283	1.2510 .1128	1.1191 .1114	.4806 .0020
+ Recourse	.0651 .0076	.0328 .0053	**.0200** .0005	.7436 .0780	.6145 .0826	.0639 .1312
+ Lynden-Bell	**.0428** .0055	**.0139** .0018	**.0200** .0005	**.4542** .0788	**.3555** .0779	**.0000** .0000
VCNET [61]	.0589 .0097	.0101 .0041	**.0000** .0000	.8172 .0514	.6580 .0487	.0033 .0095
+ Recourse	.0610 .0067	.0331 .0033	**.0000** .0000	.8090 .0503	.6520 .0485	.0031 .0090
+ Lynden-Bell	**.0312** .0043	**.0059** .0021	**.0000** .0000	**.7820** .0448	**.6443** .0453	**.0001** .0002
TRANSTEE [99]	.0933 .0374	.0261 .0229	.0120 .0114	1.0486 .1651	.9251 .1641	.0973 .1418
+ Recourse	.0594 .0070	.0189 .0043	.0002 .0004	.4853 .1214	.3463 .1168	**.0000** .0000
+ Lynden-Bell	**.0379** .0036	**.0024** .0011	**.0000** .0000	**.3030** .0216	**.1648** .0167	**.0000** .0000
ADMIT [89]	.4015 .0362	.1597 .0385	.0952 .0284	1.2441 .1045	1.1175 .1035	.4801 .0019
+ Recourse	.0735 .0083	.0433 .0050	**.0200** .0005	.8290 .0855	.7038 .0847	.2349 .2079
+ Lynden-Bell	**.0430** .0061	**.0134** .0017	**.0200** .0005	**.5609** .0939	**.4465** .0939	**.0000** .0000
DRVAE [19]	.4230 .0313	.1684 .0269	.1096 .0282	1.1829 .1389	1.0446 .1421	.4789 .0009
+ Recourse	.0679 .0087	.0324 .0049	**.0200** .0005	.6000 .0530	.4703 .0517	**.0000** .0001
+ Lynden-Bell	**.0466** .0049	**.0148** .0013	**.0200** .0005	**.4545** .0880	**.3416** .0894	**.0000** .0001

quality, acceptance rates, and post-transplant survival [21, 81]. Importantly, CIT is an *operational* variable which can be intervened on [49, 27]. We treat CIT as the treatment T and all remaining patient and organ features as confounders $\mathbf{X}$; refusals citing CIT serve as the recourse set.

Benchmarks and metrics. The methodology is applied on multiple state of the art, neural, dose-response estimators: TARNET [77], DRNET [75], VCNET [61], TRANSTEE [99], ADMIT [89], DRVAE [19]. For each estimator, a baseline version is trained on observational data and compared to a version with our recourse loss (with and without Lynden–Bell reweighting). Additionally, and analogously, we report two recourse-only baselines that do not require a base learner. Performance is measured with three standard CADR metrics, all computed against an interventional test set with treatments drawn uniformly over the evaluation range: mean integrated squared error (MISE), average mean squared error (AMSE) and decision policy error (DPE).

Results. Table 1 displays our main results. Recourse data are integrated within pre-existing base CADR learners. Generally, the models improve when having access to these data in both the synthetic and semi-synthetic settings. The effect of the debiasing from the Lynden–Bell inversion is shown in Figure 4. The Lynden–Bell inversion appears to consistently reduce the bias, induced by negative-only recourses, across different values for the loss trade-off λ. Importantly, when λ is tuned for optimality, the performance of the models is significantly increased.

7 Conclusion

In this work, we show that counterfactual recourse explanations provide structured causal information for continuous treatment effect estimation. By treating treatment-side recourses as response boundaries, we relax standard positivity requirements and turn explanations into auxiliary supervision for dose-response estimators. The experiments show improvements over standard baselines across synthetic and semi-synthetic settings. More broadly, our results suggest a promising direction for causal inference in domains where recourse-like signals are routinely collected alongside outcomes.

8 Discussion

8.1 Limitations

Faithfulness is hard to verify. Identification rests on the explanation naming the exact point at which the outcome would have flipped, and every quantity we estimate inherits whatever error that step carries. Human explanations are unlikely to be exact in many ways. A physician asked what dose would have produced a response will tend to name one comfortably inside the effective range rather than the smallest that would have worked, and explanations cluster on protocol values and round numbers, so what is recorded may reflect a shared decision rule rather than an individual threshold. Observational data rarely contain a second look at the same unit, so such distortions usually go unseen. They are not beyond reach in principle: where a unit is later observed in a comparable situation, its behaviour can be compared against what the explanation prescribed, and repeated exposures of this kind offer a route to validating, or calibrating, the assumption.

The strata depend on explanatory preferences. Treatment-preferential recourse is what turns the *type* of an explanation into a stratum label: it is the reason an explanation pointing at the confounders marks a never-responder, and the reason one pointing at the treatment marks a sensitive unit. It is therefore not a technical convenience but the hinge on which two of the three strata hang, and it can fail in both directions. Where changes to the confounders are cheap or routine, a unit that treatment could in fact have helped may be labelled a never-responder and removed from the response altogether; where the elicitation forces a treatment explanation, genuine never-responders enter the inversion carrying thresholds that do not exist. Because the label is read directly off the recorded type, neither error is averaged away by more data.

8.2 Potential societal implications

Explanations carry the conditions that produced them. Treating an explanation as causal evidence means inheriting whatever wrote it. Refusal reasons are recorded under time pressure, within institutional taxonomies, and with an eye to what will be defensible later, so a refusal that was really about the recipient may be entered as one about the treatment. The never-responder label is the most consequential case. If the groups historically refused for non-treatment reasons are not evenly distributed across the population, that label records who was turned away rather than who could not have benefited, and a policy trained on such an estimate withholds treatment from exactly those groups and generates further refusals of the same kind. Where this mass concentrates in groups defined by protected or proxy attributes we recommend withholding the never-responder component and using only the responsive one, since a wrong structural zero is a stronger and more harmful claim than a noisy response curve.

Explanations may be written strategically. Once it is known that explanations feed an estimator, they stop being passive records. For example, a team seeking more resources for relieving a particular constraint has an incentive to name that constraint more often, and one that would rather leave its own practice unexamined has an incentive to name anything else. This differs from the imprecision discussed above in being deliberate, directional, and correlated with the quantity being estimated, so it biases the result rather than blurring it, and the incentive grows precisely as the estimates come to matter more.

8.3 Future work

From structured recourse to free text. We assume the recourse is observed in structured form, which might not always be the case. The immediate extension is an extraction step that maps an explanation to an intervention with calibrated uncertainty and carries that uncertainty through to the estimator, rather than collapsing it to a point value.

Beyond explanations with an interventional reading. Not every explanation prescribes an intervention. Those that do map onto a recourse and can be used as developed here; those that do not may still express a structural constraint on the response mechanism, and discarding them loses information that is in principle usable. Encoding whatever constraint an explanation expresses, rather than only those that resolve to a boundary, points at estimators that can process more explanation types.

References

[1] Ahmed M Alaa and Mihaela Van Der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30, 2017.

[2] Panagiotis Alimisis and Christos Diou. Learning robust intervention representations with delta embeddings. *arXiv preprint arXiv:2508.04492*, 2025.

[3] Susan F Assmann, Stuart J Pocock, Laura E Enos, and Linda E Kasten. Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet*, 355(9209):1064–1069, 2000.

[4] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[5] Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *Observational studies*, 5(2):37–51, 2019.

[6] Han Bao and Michael Schomaker. Addressing positivity violations in continuous interventions through data-adaptive strategies. *arXiv preprint arXiv:2502.14566*, 2025.

[7] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

[8] Jeroen Berrevoets, Ahmed Alaa, Zhaozhi Qian, James Jordon, Alexander ES Gimson, and Mihaela Van Der Schaar. Learning queueing policies for organ transplantation allocation using interpretable counterfactual survival analysis. In *International Conference on Machine Learning*, pages 792–802. PMLR, 2021.

[9] Jeroen Berrevoets, Daniel Jarrett, Alex Chan, and Mihaela van der Schaar. Allsim: Simulating and benchmarking resource allocation policies in multi-user systems. *Advances in Neural Information Processing Systems*, 36:851–866, 2023.

[10] Jeroen Berrevoets, James Jordon, Ioana Bica, Mihaela van der Schaar, et al. Organite: Optimal transplant donor organ offering using an individual treatment effect. *Advances in neural information processing systems*, 33:20037–20050, 2020.

[11] Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33:16434–16445, 2020.

[12] Christopher Bockel-Rickermann, Toon Vanderschueren, Tim Verdonck, and Wouter Verbeke. Sources of gain: Decomposing performance in conditional average dose response estimation. *arXiv preprint arXiv:2406.08206*, 2024.

[13] Ruth MJ Byrne. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *IJCAI*, pages 6276–6282. California, CA, 2019.

[14] J Michael Cecka. The optn/unos renal transplant registry. *Clinical transplants*, pages 1–16, 2005.

[15] Jacob Chen, Rohit Bhattacharya, and Katherine Keith. Proximal causal inference with text data. *Advances in Neural Information Processing Systems*, 37:135983–136017, 2024.

[16] Yubo Chen and Jinhong Xie. Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management science*, 54(3):477–491, 2008.

[17] Stephen R Cole and Miguel A Hernán. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664, 2008.

[18] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

[19] Ruijing Cui, Jianbin Sun, Bingyu He, Kewei Yang, and Bingfeng Ge. Disentangled representation via variational autoencoder for continuous treatment effect estimation. *arXiv preprint arXiv:2406.02310*, 2024.

[20] Alicia Curth and Mihaela Van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1810–1818. PMLR, 2021.

[21] Agnes Debout, Yohann Foucher, Katy Trébern-Launay, Christophe Legendre, Henri Kreis, Georges Mourad, Valérie Garrigue, Emmanuel Morelon, Fanny Buron, Lionel Rostaing, et al. Each additional hour of cold ischemia time significantly increases the risk of graft failure and mortality following renal transplantation. *Kidney international*, 87(2):343–349, 2015.

[22] Shachi Deshpande, Kaiwen Wang, Dhruv Sreenivas, Zheng Li, and Volodymyr Kuleshov. Deep multi-modal structural equations for causal effect estimation with unstructured proxies. *Advances in Neural Information Processing Systems*, 35:10931–10944, 2022.

[23] Nikita Dhawan, Leonardo Cotta, Karen Ullrich, Rahul G Krishnan, and Chris J Maddison. End-to-end causal effect estimation from unstructured natural language data. *Advances in Neural Information Processing Systems*, 37:77165–77199, 2024.

[24] Liang Dong, Leiyang Chen, Chengliang Zheng, Zhongwang Fu, Umer Zukaib, Xiaohui Cui, and Zhidong Shen. Ocie: augmenting model interpretability via deconfounded explanation-guided learning. *Knowledge-Based Systems*, 302:112390, 2024.

[25] Brent D Ershoff, Christine K Lee, Christopher L Wray, Vatche G Agopian, Gregor Urban, Pierre Baldi, and Maxime Cannesson. Training and validation of deep neural networks for the prediction of 90-day post-liver transplant mortality using unos registry data. In *Transplantation proceedings*, volume 52, pages 246–258. Elsevier, 2020.

[26] Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

[27] Wojciech Figiel, Piotr Smoter, Maciej Krasnodębski, Paweł Rykowski, Marcin Morawski, Michał Grąt, Waldemar Patkowski, and Krzysztof Zieniewicz. Predictors of long-term outcomes after liver transplantation depending on the length of cold ischemia time. In *Transplantation Proceedings*, volume 54, pages 1025–1028. Elsevier, 2022.

[28] Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

[29] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. Going beyond xai: A systematic survey for explanation-guided learning. *ACM Computing Surveys*, 56(7):1–39, 2024.

[30] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[31] Michael Gill and Andrew Hall. How judicial identity changes the text of legal rulings. *Available at SSRN 2620781*, 2015.

[32] Vittorio Girotto, Paolo Legrenzi, and Antonio Rizzo. Event controllability in counterfactual thinking. *Acta Psychologica*, 78(1-3):111–133, 1991.

[33] Limor Gultchin, David Watson, Matt Kusner, and Ricardo Silva. Operationalizing complex causes: A pragmatic view of mediation. In *International Conference on Machine Learning*, pages 3875–3885. PMLR, 2021.

[34] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[35] K Hirano and GW Imbens. The propensity score with continuous treatments. applied bayesian modeling and causal inference from incomplete-data perspectives, 2004.

[36] Nicolas Huynh, Julianna Piskorz, Jeroen Berrevoets, Max Ruiz Luyten, and Mihaela van der Schaar. Improving treatment effect estimation with llm-based data augmentation. In *1st ICML Workshop on Foundation Models for Structured Data*, 2025.

[37] Kosuke Imai and Kentaro Nakamura. Causal representation learning with generative artificial intelligence: Application to texts as treatments. *arXiv preprint arXiv:2410.00903*, 2024.

[38] Kosuke Imai and David A Van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866, 2004.

[39] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

[40] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

[41] Daniel Kahneman and Gary Klein. Conditions for intuitive expertise: a failure to disagree. *American psychologist*, 64(6):515, 2009.

[42] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022.

[43] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 353–362, 2021.

[44] Amir-Hossein Karimi, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in neural information processing systems*, 33:265–277, 2020.

[45] Katherine Keith, David Jensen, and Brendan O'Connor. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, 2020.

[46] Katherine A. Keith, Douglas Rice, and Brendan O'Connor. Text as causal mediators: research design for causal estimates of differential treatment of social groups via language aspects. In *Proceedings of the Workshop on Causal Inference and NLP*, 2021.

[47] Anna Koroleva, Sanjay Kamath, and Patrick Paroubek. Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations. *Journal of Biomedical Informatics*, 100:100058, 2019.

[48] Seungyeon Lee, Ruoqi Liu, Wenyu Song, Lang Li, and Ping Zhang. Subgroupte: Advancing treatment effect estimation with subgroup identification. *ACM transactions on intelligent systems and technology*, 16(3):1–23, 2025.

[49] Vladimir J Lozanovski, Bernd Döhler, Karl Heinz Weiss, Arianeb Mehrabi, and Caner Süsal. The differential influence of cold ischemia time on outcome after liver transplantation for different indications—who is at risk? a collaborative transplant study report. *Frontiers in immunology*, 11:892, 2020.

[50] Donald Lynden-Bell. A method of allowing for known observational selection in small samples applied to 3cr quasars. *Monthly Notices of the Royal Astronomical Society*, 155(1):95–118, 1971.

[51] Yuchen Ma, Dennis Frauen, Jonas Schweisthal, and Stefan Feuerriegel. Llm-driven treatment effect estimation under inference time text confounding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2024.

[52] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

[53] Arun S Maiya. Causalnlp: A practical toolkit for causal inference with text. *arXiv preprint*, 2021.

[54] Alessandro Marchese, Jeroen Berrevoets, and Sam Verboven. Causal explanation-guided learning for organ allocation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[55] Alessandro Marchese, Hans de Ferrante, Jeroen Berrevoets, and Sam Verboven. Dynamite: Optimal time-sensitive organ offers using ite. *Machine Learning for Health (ML4H)*, pages 696–713, 2025.

[56] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[57] Iman Modarressi, Jann Spiess, and Amar Venugopal. Causal inference on outcomes learned from text. *arXiv preprint arXiv:2503.00725*, 2025.

[58] Byron JT Morgan. *Analysis of quantal response data*. Chapman and Hall/CRC, 2018.

[59] Susan M Mudambi and David Schuff. What makes a helpful online review? a study of customer reviews on amazon. com. *MIS quarterly*, 34(1):185–200, 2010.

[60] Scott Mueller and Judea Pearl. Personalized decision making–a conceptual introduction. *Journal of Causal Inference*, 11(1):20220050, 2023.

[61] Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. *arXiv preprint arXiv:2103.07861*, 2021.

[62] Judea Pearl. *Causality*. Cambridge university press, 2009.

[63] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

[64] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

[65] Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. Causal effects of linguistic properties. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4095–4109, 2021.

[66] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

[67] Rang Rang, James M Ritter, Rod J Flower, and Graeme Henderson. *Rang & dale farmacologia*. Elsevier Brasil, 2015.

[68] Katharina Ring and Michael Schomaker. A diagnostic to find and help combat stochastic positivity issues–with a focus on continuous treatments. *Journal of Causal Inference*, 14(1):20250007, 2026.

[69] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[70] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

[71] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[72] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

[73] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[74] Michael Schomaker, Helen McIlleron, Paolo Denti, and Iván Díaz. Causal inference for continuous multiple time point interventions. *Statistics in Medicine*, 43(28):5380–5400, 2024.

[75] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5612–5619, 2020.

[76] Shahana Sen and Dawn Lerman. Why are you telling me this? an examination into negative consumer reviews on the web. *Journal of interactive marketing*, 21(4):76–94, 2007.

[77] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.

[78] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

[79] Kaveh G Shojania, Elizabeth C Burton, Kathryn M McDonald, and Lee Goldman. Changes in rates of autopsy-detected diagnostic errors over time: a systematic review. *Jama*, 289(21):2849–2856, 2003.

[80] Dhanya Sridhar and Lise Getoor. Estimating causal effects of tone in online debates. *arXiv preprint arXiv:1906.04177*, 2019.

[81] James E Stahl, Jennifer E Kreke, Fawaz Ali Abdul Malek, Andrew J Schaefer, and Joseph Vacanti. Consequences of cold-ischemia time on primary nonfunction and patient and graft survival in liver transplantation: a meta-analysis. *PloS one*, 3(6):e2468, 2008.

[82] Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, Bogong Li, et al. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(2), 2009.

[83] Ariel Soares Teles, Ivan Rodrigues de Moura, Francisco Silva, Angus Roberts, and Daniel Stahl. Ehr-based prediction modelling meets multimodal deep learning: A systematic review of structured and textual data fusion methods. *Information Fusion*, 118:102981, 2025.

[84] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.

[85] Victor Veitch, Dhanya Sridhar, and David Blei. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

[86] Edward Vytlacil. Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, 70(1):331–341, 2002.

[87] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[88] Lei Wang, Ee-Peng Lim, Zhiwei Liu, and Tianxiang Zhao. Explanation guided contrastive learning for sequential recommendation. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 2017–2027, 2022.

[89] Xin Wang, Shengfei Lyu, Xingyu Wu, Tianhao Wu, and Huanhuan Chen. Generalization bounds for estimating causal effects of continuous treatments. *Advances in Neural Information Processing Systems*, 35:8605–8617, 2022.

[90] Zhao Wang and Aron Culotta. When do words matter? understanding the impact of lexical choice on audience perception using individual treatment effect estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[91] Leander Weber, Sebastian Lapuschkin, Alexander Binder, and Wojciech Samek. Beyond explaining: Opportunities and challenges of xai-based model improvement. *Information Fusion*, 92:154–176, 2023.

[92] Galen Weld, Peter West, Maria Glenski, David Arbour, Ryan A Rossi, and Tim Althoff. Adjusting for confounders with text: Challenges and an empirical evaluation framework for causal inference. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2022.

[93] Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. Challenges of using text classifiers for causal inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, page 4586, 2018.

[94] Michael Woodroofe. Estimating a distribution function with truncated data. *The Annals of Statistics*, 13(1):163–177, 1985.

[95] Han Wu, Sarah Tan, Weiwei Li, Mia Garrard, Adam Obeng, Drew Dimmery, Shaun Singh, Hanson Wang, Daniel Jiang, and Eytan Bakshy. Distilling heterogeneity: From explanations of heterogeneous treatment effect models to interpretable policies. *arXiv preprint arXiv:2111.03267*, 2021.

[96] Jinsung Yoon, Ahmed Alaa, Martin Cadeiras, and Mihaela Van Der Schaar. Personalized donor-recipient matching for organ transplantation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[97] Jiaming Zeng, Michael F Gensheimer, Daniel L Rubin, Susan Athey, and Ross D Shachter. Uncovering interpretable potential confounders in electronic medical records. *Nature communications*, 13(1):1014, 2022.

[98] Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. Quantifying the causal effects of conversational tendencies. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–24, 2020.

[99] Yi-Fan Zhang, Hanlin Zhang, Zachary C Lipton, Li Erran Li, and Eric P Xing. Exploring transformer backbones for heterogeneous treatment effect estimation. *arXiv preprint arXiv:2202.01336*, 2022.

[100] Yikun Zhang, Yen-Chi Chen, and Alexander Giessing. Nonparametric inference on dose-response curves without the positivity condition. *arXiv preprint arXiv:2405.09003*, 2024.

A Table of strata and events

Table 2: Strata used throughout the identification and methodology sections: the three latent response strata of (6) (top) and the observable events that certify them (bottom).

Stratum	Description	How identified from data	Role in identification & methodology
$\{S = \mathrm{SR}\}$	Sensitive responders: $T_{\min} < B \le T_{\max}$, so the outcome flips at some admissible dose.	Proportion π_{SR} from (12): the treatment-side recourse proportion divided by the truncation prefactor.	Carries the only non-degenerate curve $\mu^{\mathrm{SR}} = F_{B\mid\mathbf{x},\mathrm{SR}}$, recovered by the Lynden–Bell inversion (Proposition 1).
$\{S = \mathrm{NR}\}$	Never-responders: $B > T_{\max}$, so no admissible dose would yield $Y = 1$.	$\pi_{\mathrm{NR}} = P(Y = 0, \boldsymbol{\delta}_x \neq \mathbf{0} \mid \mathbf{x})$, certified by confounder-side recourse under (3).	Structural zero $\mu^{\mathrm{NR}} \equiv 0$; subtracts a constant in (14), giving $\mu(\mathbf{x}, T_{\max}) = 1 - \pi_{\mathrm{NR}}$.
$\{S = \mathrm{AR}\}$	Always-responders: $B \le T_{\min}$, so every admissible dose yields $Y = 1$.	Isolated by no observable event; recovered as the residual $\pi_{\mathrm{AR}} = 1 - \pi_{\mathrm{SR}} - \pi_{\mathrm{NR}}$.	Structural one $\mu^{\mathrm{AR}} \equiv 1$; absorbed into the leading 1 of (14) and of the pseudo-outcome (17).
$\{\mathrm{TR} = 1\}$	Treatment-side recourse: $y_i = 0$ with $\boldsymbol{\delta}_{x,i} = \mathbf{0}$; a strict subset of $\{S = \mathrm{SR}\}$.	Directly observable in $\mathcal{D}_{\mathrm{obs}}$ from the recourse type (15).	Certifies $S_i = \mathrm{SR}$ and reveals $b_i = t_i + \delta_{t,i}$; the only stratum carrying boundary information, LB-weighted in (19).
$\{\mathrm{CR} = 1\}$	Confounder-side or mixed recourse: $y_i = 0$ with $\boldsymbol{\delta}_{x,i} \neq \mathbf{0}$.	Directly observable in $\mathcal{D}_{\mathrm{obs}}$ from the recourse type (15).	Certifies $S_i = \mathrm{NR}$; contributes the constant pseudo-outcome 0 in (19).
$\{Y = 1\}$	Positive outcome: a mixture of AR units and SR units dosed above their own boundary, which no observable event separates.	$y_i = 1$ in the factual data; no recourse, hence no boundary.	Carries the always-responder mass collectively; contributes the constant pseudo-outcome 1 in (19), which the residual in (12) makes legitimate.

B Identification in other settings

B.1 CATE identification

We specialize the CADR result to the binary-treatment setting. With $\mathcal{T} = \{0, 1\}$ the threshold model (5) gives $Y(t) = \mathbf{1}\{t \ge B\}$, which is non-decreasing in t by construction, so monotonicity is implied rather than assumed [86] and the type $(Y(0), Y(1)) = (1, 0)$ cannot occur. The three strata of (6) then coincide exactly with the classical compliance types: $\mathrm{NR} = (0, 0)$, $\mathrm{SR} = (0, 1)$ and $\mathrm{AR} = (1, 1)$. Since the treatment effect is $+1$ on SR and 0 on both constant strata, the CATE is simply the sensitive proportion, and the no-always-responder condition is unnecessary for the same reason as in Section 3: π_{AR} never has to be estimated.

Proposition 2 (Binary specialization)**.** *Assume the SCM* (1) *with* $\mathcal{T} = \{0, 1\}$*, the threshold parametrization* (5)*, structural faithfulness* ***(F)****, treatment-preferential cost* (3)*, and* $P(T = 0 \mid \mathbf{x}_0) > 0$*. Then*

$$\tau(\mathbf{x}_0) \;:=\; \mathbb{E}\big[Y(1) - Y(0) \,\big|\, \mathbf{x}_0\big] \;=\; \pi_{\mathrm{SR}}(\mathbf{x}_0) \;=\; \frac{P\big(Y = 0,\, \boldsymbol{\delta}_x = \mathbf{0} \mid \mathbf{x}_0\big)}{P(T = 0 \mid \mathbf{x}_0)}. \tag{22}$$

Proof. Under (5), $\tau_i = Y_i(1) - Y_i(0) = \mathbf{1}\{B_i \le 1\} - \mathbf{1}\{B_i \le 0\}$, which is 1 exactly on $\{0 < B_i \le 1\} = \{S_i = \mathrm{SR}\}$ and 0 on NR and AR. The law of total expectation over (6) gives $\tau(\mathbf{x}_0) = \pi_{\mathrm{SR}}(\mathbf{x}_0)$. For the second equality, a sensitive unit has $Y = 0$ iff $T = 0$, so $\{Y = 0, \boldsymbol{\delta}_x = \mathbf{0}\} = \{S = \mathrm{SR}, T = 0\}$ by (12), and $T \perp B \mid \mathbf{X}$ gives $P(S = \mathrm{SR}, T = 0 \mid \mathbf{x}_0) = \pi_{\mathrm{SR}}(\mathbf{x}_0)\, P(T = 0 \mid \mathbf{x}_0)$. □

The denominator is the binary specialization of the Lynden–Bell kernel: for $b \in (0, 1]$, $F_{T\mid\mathbf{x}_0}(b^-) = P(T = 0 \mid \mathbf{x}_0)$ is constant over the sensitive support, so the inversion of (10) reduces to a single division. Correspondingly, **(A1′)** reduces to $P(T = 0 \mid \mathbf{x}_0) > 0$ alone: unlike **(A1)** it places no requirement on $P(T = 1 \mid \mathbf{x}_0)$, so $\tau(\mathbf{x}_0)$ remains identified at covariate values where no unit was ever treated.

Dropping the threshold parametrization (5) re-admits defiers, and $\tau(\mathbf{x}_0)$ is then no longer a stratum proportion. Faithfulness nonetheless pins both potential outcomes of any $\{\mathrm{TR}_i = 1\}$ unit — $Y_i(t_i) =$

0 by consistency and $Y_i(1-t_i)=1$ by the recourse certificate, so $\tau_i = 1-2t_i$ — and averaging over $T \mid \mathbf{x}_0, \mathrm{TR}=1$ identifies $\tau^{\mathrm{TR}=1}(\mathbf{x}_0) = P(T=0 \mid \mathbf{x}_0, \mathrm{TR}=1) - P(T=1 \mid \mathbf{x}_0, \mathrm{TR}=1)$ on the subpopulation whose counterfactuals the recourse supplies.

B.2 Multi-threshold response

Beyond the scalar threshold-crossing parametrization (5), the success region $\mathcal{R}_1(\mathbf{x}, u_Y)$ may be a finite union of intervals — a regime common in dose-response settings with sweet-spot behavior, where both insufficient and excessive treatment fail. We treat the two-boundary case first and sketch the generalization.

Two-boundary (success window) case. Let $\mathcal{R}_1(\mathbf{x}, u_Y) = [B^L, B^U] \cap \mathcal{T}$ with lower boundary $B^L = g^L(\mathbf{X}, U_Y)$ and upper boundary $B^U = g^U(\mathbf{X}, U_Y)$, both continuously distributed, e.g., therapeutic windows in drug dosing [67]. The position of the window relative to $\mathcal{T}$ replaces (6) as the stratum definition:

$$S = \begin{cases} \mathrm{NR} & \text{if } [B^L, B^U] \cap \mathcal{T} = \emptyset, \text{ i.e. } B^L > T_{\max} \text{ or } B^U < T_{\min}, \\ \mathrm{AR} & \text{if } \mathcal{T} \subseteq [B^L, B^U], \text{ i.e. } B^L \le T_{\min} \text{ and } B^U \ge T_{\max}, \\ \mathrm{SR} & \text{otherwise: the window meets } \mathcal{T} \text{ without covering it.} \end{cases} \tag{23}$$

The structural curves and the recourse certificates are unchanged: $\mu^{\mathrm{NR}} \equiv 0$, $\mu^{\mathrm{AR}} \equiv 1$, a negative outcome with a treatment-only recourse certifies $S = \mathrm{SR}$, and one with a confounder-side recourse certifies $S = \mathrm{NR}$. A sensitive unit fails at t_0 in exactly one of two disjoint ways — below the window or above it — so, counting failures as in (14),

$$\mu^{\mathrm{SR}}(\mathbf{x}_0, t_0) \;=\; 1 - P\big(B^L > t_0 \,\big|\, \mathbf{x}_0, \mathrm{SR}\big) - P\big(B^U < t_0 \,\big|\, \mathbf{x}_0, \mathrm{SR}\big). \tag{24}$$

Two boundary streams. For a $\{\mathrm{TR}_i = 1\}$ unit the sign of $\delta_{t,i}$ identifies which boundary the recourse hits: $\delta_{t,i} > 0$ pins a lower-boundary observation $b_i^L = t_i + \delta_{t,i} > t_i$, left-truncated with kernel $F_{T|\mathbf{x}_0}(b^-)$, while $\delta_{t,i} < 0$ pins an upper-boundary observation $b_i^U = t_i + \delta_{t,i} < t_i$, right-truncated with kernel $1 - F_{T|\mathbf{x}_0}(b)$ [94]. The two streams admit symmetric Lynden–Bell inversions in the appropriate direction.

Two-directional cumulative positivity. We replace **(A1′)** with:

(B1′) Two-sided recourse positivity. For the evaluation covariate $\mathbf{x}_0$: (i) $F_{T|\mathbf{x}_0}(b^-) > 0$ for every $b \in \mathrm{supp}(B^L \mid \mathbf{x}_0, \mathrm{SR}) \cap (T_{\min}, T_{\max}]$, and (ii) $1 - F_{T|\mathbf{x}_0}(b) > 0$ for every $b \in \mathrm{supp}(B^U \mid \mathbf{x}_0, \mathrm{SR}) \cap [T_{\min}, T_{\max})$.

Treatment mass is required both below every attainable lower boundary and above every attainable upper boundary — the natural condition when the evaluation dose sits in a gap with data on either side — but still nowhere at t_0 itself.

Proposition 3 (Two-boundary identification)**.** *Assume the SCM* (1) *with success region* $[B^L, B^U] \cap \mathcal{T}$*, structural faithfulness* ***(F)****, treatment-preferential cost* (3) *and* ***(B1′)****. Then for every* $t_0 \in \mathcal{T}$ *the full-population CADR is identified by*

$$\begin{aligned} \mu(\mathbf{x}_0, t_0) \;=\; & 1 - P\big(Y=0,\, \boldsymbol{\delta}_x \neq \mathbf{0} \mid \mathbf{x}_0\big) \\ & - P\big(\mathrm{TR}=1,\, \delta_t > 0 \mid \mathbf{x}_0\big) \int_{t_0}^{T_{\max}} \frac{p(b \mid \mathbf{x}_0, \mathrm{TR}=1, \delta_t > 0)}{F_{T|\mathbf{x}_0}(b^-)}\, db \\ & - P\big(\mathrm{TR}=1,\, \delta_t < 0 \mid \mathbf{x}_0\big) \int_{T_{\min}}^{t_0} \frac{p(b \mid \mathbf{x}_0, \mathrm{TR}=1, \delta_t < 0)}{1 - F_{T|\mathbf{x}_0}(b)}\, db. \end{aligned} \tag{25}$$

Proof. Recourses with $\delta_t > 0$ are drawn from $\{S = \mathrm{SR}, Y = 0, T < B^L\}$, so by the argument of (8)–(9) applied to the B^L stream, $p(b \mid \mathbf{x}_0, \mathrm{TR}=1, \delta_t > 0) = f_{B^L|\mathbf{x}_0,\mathrm{SR}}(b)\, F_{T|\mathbf{x}_0}(b^-) / P(Y = 0, \delta_t > 0 \mid \mathbf{x}_0, \mathrm{SR})$. Inverting under **(B1′)**(i), integrating over $(t_0, T_{\max}]$ and multiplying by $\pi_{\mathrm{SR}}(\mathbf{x}_0)$ gives $\pi_{\mathrm{SR}} P(B^L > t_0 \mid \mathbf{x}_0, \mathrm{SR})$ as the second subtracted term, the truncation prefactor cancelling against π_{SR} exactly as in (14). The symmetric argument on the B^U stream, with right-truncation kernel $1 - F_{T|\mathbf{x}_0}$ and **(B1′)**(ii), gives the third term. Substituting both into (24), together with $\pi_{\mathrm{NR}} = P(Y = 0, \boldsymbol{\delta}_x \neq \mathbf{0} \mid \mathbf{x}_0)$ and $\pi_{\mathrm{AR}} = 1 - \pi_{\mathrm{SR}} - \pi_{\mathrm{NR}}$, yields (25). □

Writing μ^{SR} as one minus the two failure probabilities, rather than as the difference $F_{B^L} - F_{B^U}$ of two CDFs, is what keeps both integrals inside $\mathcal{T}$: the CDF form would require the left tail of B^L below $T_{\min}$, which no recourse reveals. This is the window analogue of the counting argument behind (14).

Generalization to M boundaries. For $\mathcal{R}_1(\mathbf{x}, u_Y) = \bigcup_{m=1}^{M/2}[B_m^L, B_m^U] \cap \mathcal{T}$, each boundary function is still identified by a Lynden–Bell inversion of its own stream in the appropriate truncation direction, under the corresponding two-directional positivity condition. The counting argument (24), however, no longer reduces to two terms: a unit whose recourse points upward from t_i to b_m^L may still succeed at some $t_0 < b_m^L$ lying in an earlier window, so failure at t_0 is not determined by the revealed boundary alone. Identification therefore requires three further ingredients: (i) M is known; (ii) each recourse $\boldsymbol{\delta}_i$ is assigned to a specific stream $(m, L/U)$ and this assignment is known; and (iii) every boundary lies within $\mathcal{T}$, since a window falling entirely outside it is revealed by no recourse. Under (i)–(iii) the per-stream inversions recover all M boundary distributions and $\mu^{\mathrm{SR}}(\mathbf{x}_0, t_0) = P\big(t_0 \in \mathcal{R}_1 \mid \mathbf{x}_0, \mathrm{SR}\big)$ follows.

C Population risk of the auxiliary loss

We show that the population minimizer of (20) is the full-population CADR μ of (4), not a stratum-level or subsample-level quantity.

Population risk. Write $\tilde{y}(t)$ for the pseudo-outcome (17) with the population weight $w(\mathbf{x}, b) = 1/F_{T|\mathbf{x}}(b^-)$ of (16). With counterfactual doses drawn as $t \sim \mathrm{Unif}(\mathcal{T})$ independently of the unit, the population risk corresponding to (20) is

$$\mathcal{R}(\hat{\mu}) \;=\; \mathbb{E}_{(\mathbf{X},\tilde{y})}\,\mathbb{E}_{t\sim\mathrm{Unif}(\mathcal{T})}\big[\big(\hat{\mu}(\mathbf{X}, t) - \tilde{y}(t)\big)^2\big] \;=\; \frac{1}{|\mathcal{T}|}\int_{\mathcal{T}} \mathbb{E}\big[\big(\hat{\mu}(\mathbf{X}, t) - \tilde{y}(t)\big)^2\big]\, dt. \tag{26}$$

Since t is drawn independently of the unit and $\hat{\mu}$ is unrestricted, $\mathcal{R}$ decomposes pointwise over $(\mathbf{x}_0, t_0) \in \mathcal{X} \times \mathcal{T}$. Fixing $(\mathbf{x}_0, t_0)$ and writing $m = \hat{\mu}(\mathbf{x}_0, t_0)$, the map $m \mapsto \mathbb{E}\big[(m - \tilde{y}(t_0))^2 \mid \mathbf{x}_0\big]$ is minimized at the conditional mean,

$$m^\star(\mathbf{x}_0, t_0) \;=\; \mathbb{E}\big[\tilde{y}(t_0) \,\big|\, \mathbf{x}_0\big]. \tag{27}$$

Evaluating the conditional mean. By (17), $\tilde{y}(t_0) = 1 - \mathrm{CR} - \mathrm{TR} \cdot w(\mathbf{x}_0, B)\, \mathbf{1}\{B > t_0\}$, so

$$\mathbb{E}\big[\tilde{y}(t_0) \mid \mathbf{x}_0\big] \;=\; 1 \;-\; \mathbb{E}\big[\mathrm{CR} \mid \mathbf{x}_0\big] \;-\; \mathbb{E}\Big[\mathrm{TR} \cdot \tfrac{\mathbf{1}\{B>t_0\}}{F_{T|\mathbf{x}_0}(B^-)} \,\Big|\, \mathbf{x}_0\Big]. \tag{28}$$

The second term is a proportion of an observable event, and by (12),

$$\mathbb{E}\big[\mathrm{CR} \mid \mathbf{x}_0\big] \;=\; P\big(Y = 0,\, \boldsymbol{\delta}_x \neq \mathbf{0} \mid \mathbf{x}_0\big) \;=\; \pi_{\mathrm{NR}}(\mathbf{x}_0). \tag{29}$$

For the third term, the integrand vanishes unless $\mathrm{TR} = 1$, and $\mathrm{supp}(B \mid \mathbf{x}_0, \mathrm{SR}) \subseteq (T_{\min}, T_{\max}]$ by (6), so

$$\mathbb{E}\Big[\mathrm{TR} \cdot \tfrac{\mathbf{1}\{B>t_0\}}{F_{T|\mathbf{x}_0}(B^-)} \,\Big|\, \mathbf{x}_0\Big] \;=\; P(\mathrm{TR} = 1 \mid \mathbf{x}_0) \int_{t_0}^{T_{\max}} \frac{p(b \mid \mathbf{x}_0,\, Y = 0,\, \mathrm{SR})}{F_{T|\mathbf{x}_0}(b^-)}\, db, \tag{30}$$

which is finite under **(A1′)**. Applying the Lynden–Bell inversion (11) to the integrand and using (7),

$$\int_{t_0}^{T_{\max}} \frac{p(b \mid \mathbf{x}_0, Y = 0, \mathrm{SR})}{F_{T|\mathbf{x}_0}(b^-)}\, db \;=\; \frac{P(B > t_0 \mid \mathbf{x}_0, \mathrm{SR})}{P(Y = 0 \mid \mathbf{x}_0, \mathrm{SR})} \;=\; \frac{1 - \mu^{\mathrm{SR}}(\mathbf{x}_0, t_0)}{P(Y = 0 \mid \mathbf{x}_0, \mathrm{SR})}. \tag{31}$$

Since $\{\mathrm{TR} = 1\} = \{S = \mathrm{SR},\, Y = 0\}$, we have $P(\mathrm{TR} = 1 \mid \mathbf{x}_0) = \pi_{\mathrm{SR}}(\mathbf{x}_0)\, P(Y = 0 \mid \mathbf{x}_0, \mathrm{SR})$, which is the second identity of (12). The truncation prefactor therefore cancels in (30), leaving

$$\mathbb{E}\Big[\mathrm{TR} \cdot \tfrac{\mathbf{1}\{B>t_0\}}{F_{T|\mathbf{x}_0}(B^-)} \,\Big|\, \mathbf{x}_0\Big] \;=\; \pi_{\mathrm{SR}}(\mathbf{x}_0)\big[1 - \mu^{\mathrm{SR}}(\mathbf{x}_0, t_0)\big]. \tag{32}$$

Conclusion. Substituting (29) and (32) into (28) and using $\pi_{\mathrm{AR}} = 1 - \pi_{\mathrm{SR}} - \pi_{\mathrm{NR}}$ together with the decomposition (13),

$$\begin{aligned} m^\star(\mathbf{x}_0, t_0) \;&=\; 1 - \pi_{\mathrm{NR}}(\mathbf{x}_0) - \pi_{\mathrm{SR}}(\mathbf{x}_0)\big[1 - \mu^{\mathrm{SR}}(\mathbf{x}_0, t_0)\big] \\ &=\; \pi_{\mathrm{SR}}(\mathbf{x}_0)\, \mu^{\mathrm{SR}}(\mathbf{x}_0, t_0) + \pi_{\mathrm{AR}}(\mathbf{x}_0) \;=\; \mu(\mathbf{x}_0, t_0). \end{aligned} \tag{33}$$

The population minimizer of $\mathcal{L}_{\mathrm{Rec}}$ is therefore the full-population CADR at every $(\mathbf{x}_0, t_0)$ with $p(\mathbf{x}_0) > 0$ at which **(A1′)** holds.

D Metrics

We evaluate each estimator $\hat{\mu}$ against ground truth μ on a held-out evaluation set $\{(\mathbf{x}_i, t)\}_{i\in\mathcal{I}_{\text{eval}}, t\in\mathcal{T}_{\text{grid}}}$, where $\mathcal{T}_{\text{grid}} \subset [T_{\min}, T_{\max}]$ is a uniform grid of $|\mathcal{T}_{\text{grid}}|$ treatment values and $\mathcal{I}_{\text{eval}}$ indexes the evaluation units. We report three complementary metrics, each capturing a different aspect of CADR estimation quality [75, 61, 11].

Mean Integrated Squared Error (MISE). Per-unit L^2 error of the dose-response curve, averaged across units:

$$\text{MISE} \;=\; \frac{1}{|\mathcal{I}_{\text{eval}}|} \sum_{i\in\mathcal{I}_{\text{eval}}} \int_{T_{\min}}^{T_{\max}} \big(\hat{\mu}(\mathbf{x}_i, t) - \mu(\mathbf{x}_i, t)\big)^2 \, dt, \tag{34}$$

approximated by numerical integration on $\mathcal{T}_{\text{grid}}$. MISE measures pointwise estimation accuracy at the unit level: it penalizes any deviation between $\hat{\mu}$ and μ at any $(\mathbf{x}, t)$, regardless of whether the deviation cancels in aggregate.

Average Mean Squared Error (AMSE). Population-level L^2 error of the average dose-response curve:

$$\text{AMSE} \;=\; \int_{T_{\min}}^{T_{\max}} \left(\frac{1}{|\mathcal{I}_{\text{eval}}|} \sum_{i\in\mathcal{I}_{\text{eval}}} \hat{\mu}(\mathbf{x}_i, t) - \frac{1}{|\mathcal{I}_{\text{eval}}|} \sum_{i\in\mathcal{I}_{\text{eval}}} \mu(\mathbf{x}_i, t) \right)^2 dt. \tag{35}$$

AMSE is never larger than MISE: it tolerates per-unit deviations that average out across the population. The gap $\text{MISE} - \text{AMSE}$ measures how much of the estimator's error is heterogeneous (unit-specific) versus shared (population-level bias). A small AMSE with large MISE indicates correct population mean but poor heterogeneity recovery.

Decision Policy Error (DPE). Squared regret of the estimator-induced dose policy relative to the oracle policy:

$$\text{DPE} \;=\; \frac{1}{|\mathcal{I}_{\text{eval}}|} \sum_{i\in\mathcal{I}_{\text{eval}}} \big(\mu(\mathbf{x}_i,\, \hat{t}_i^*) - \mu(\mathbf{x}_i,\, t_i^*)\big)^2, \tag{36}$$

where $\hat{t}_i^* = \operatorname{argmax}_{t\in\mathcal{T}_{\text{grid}}} \hat{\mu}(\mathbf{x}_i, t)$ is the dose recommended by the estimator for unit i and $t_i^* = \operatorname{argmax}_{t\in\mathcal{T}_{\text{grid}}} \mu(\mathbf{x}_i, t)$ is the oracle-optimal dose. DPE measures the cost of acting on $\hat{\mu}$ rather than μ: a small DPE with large MISE indicates that the estimator misranks doses globally but correctly identifies the optimal dose per unit; a large DPE with small MISE indicates the converse.

E Data and models

Experiments ran on a 13th Gen Intel(R) Core(TM) i9-13900HX processor with 32GB RAM. End-to-end wall-clock time for running the synthetic experiments $\approx$ 2min.

E.1 Synthetic setup

Synthetic data-generating process. We instantiate the SCM (1) with the threshold-crossing parametrization (5):

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0},\, \mathbf{I}_{d_x}), \qquad T \,\big|\, \mathbf{x} \;\sim\; \mathcal{N}\big(\mathbf{x}^\top \boldsymbol{\beta}_T,\, \sigma_T^2\big) \,\big|\, [T_{\min},\, T_{\max}^{\text{obs}}], \qquad Y \;=\; \mathbf{1}\big\{T \;\geq\; g(\mathbf{x},\, U_Y)\big\}, \tag{37}$$

with threshold function

$$g(\mathbf{x},\, u_Y) \;=\; \boldsymbol{\beta}_g^\top \mathbf{x} \;+\; \mathbf{x}^\top \boldsymbol{\Gamma}_g\, \mathbf{x} \;+\; c_g \;+\; u_Y, \qquad U_Y \;\sim\; \text{Gumbel}(0,\, \sigma_{U_Y}), \tag{38}$$

where $\boldsymbol{\Gamma}_g$ is symmetric, c_g is a global shift, and the conditional treatment distribution is a Gaussian truncated to the observable interval $[T_{\min}, T_{\max}^{\text{obs}}] \subsetneq [T_{\min}, T_{\max}]$ — the upper truncation $T_{\max}^{\text{obs}}$ is chosen strictly below $T_{\max}$ so that $\text{supp}(B \mid \mathbf{x})$ is partially uncovered by $\text{supp}(T \mid \mathbf{x})$, the regime in which classical positivity **(A1)** fails but recourse positivity **(A1′)** holds.

The boundary random variable $B = g(\mathbf{X}, U_Y)$ has realization b_i for each unit. For each $\{Y_i = 0\}$ unit with $b_i \leq T_{\max}$, we treat b_i as a boundary-faithful treatment-side recourse ($\text{TR}_i = 1$, $\delta_{t,i} = b_i - t_i$); this is the idealization in which structural faithfulness **(F)** holds exactly, isolating the LB inversion from optimizer-fidelity effects. The remaining $\{Y_i = 0\}$ units have $b_i > T_{\max}$ and receive a confounder-side recourse ($\text{CR}_i = 1$).

The CADR is closed-form,

$$\mu(\mathbf{x},\, t) \;=\; F_{U_Y}\big(t - \boldsymbol{\beta}_g^\top \mathbf{x} - \mathbf{x}^\top \boldsymbol{\Gamma}_g\, \mathbf{x} - c_g\big), \tag{39}$$

where F_{U_Y} is the noise CDF, providing pointwise ground truth against which estimators are evaluated on an independent unbiased evaluation set with $T \sim \mathrm{Unif}(T_{\min}, T_{\max})$.

Table 3: Synthetic DGP parameters.

Parameter	Value
Covariate dimension d_x	5
Treatment range $[T_{\min}, T_{\max}]$	$[-3, 3]$
Observable upper bound $T_{\max}^{\mathrm{obs}}$	0
Treatment coefficients $\boldsymbol{\beta}_T$	$\sim \mathcal{N}(\mathbf{0}, 0.3^2\mathbf{I})$
Treatment noise σ_T	0.8
Threshold linear coefficients $\boldsymbol{\beta}_g$	$\sim \mathcal{N}(\mathbf{0}, 0.8^2\mathbf{I})$
Threshold quadratic matrix $\boldsymbol{\Gamma}_g$	$\frac{1}{2}(\mathbf{A} + \mathbf{A}^\top)$, $\mathbf{A} \sim \mathcal{N}(\mathbf{0}, 0.15^2\mathbf{I})$
Threshold global shift c_g	-2
Threshold noise scale σ_{U_Y}	2

E.1.1 Semi-synthetic UNOS-PTR data

Cohort and features. We use the UNOS Patient and Transplant Registry, restricted to deceased-donor liver transplants. We use a similar preprocessing as in prior work [25, 10, 8, 9, 55, 54]. Each candidate offer consists of patient-side features (demographics, primary diagnosis, lab values, anthropometrics) and organ-side features (donor demographics, mechanism of death, smoking and alcohol history). The treatment is cold-ischemia time (CIT, in hours), and the outcome Y is the binary acceptance/refusal of the offer. The average acceptance rate is approximately 3%, reflecting the highly skewed nature of the offer-decision distribution.

Recourse structure. Each refusal in the registry is annotated with a structured reason. Approximately 7% of refusals cite CIT as reason; these are interpreted as treatment-side recourses ($\mathrm{TR}_i = 1$) where reducing CIT would have flipped the decision. The remaining refusals cite non-CIT reasons (donor age, organ quality, recipient medical condition, etc.); under treatment-preferential cost (3) these are confounder-side recourses ($\mathrm{CR}_i = 1$) certifying $S_i = \mathrm{NR}$ — units for whom no change in CIT alone would have changed the outcome. Accepted offers ($Y_i = 1$) carry neither recourse type.

Boundary construction. While the refusal reasons themselves are real registry data, the registry does not directly record *how much* CIT reduction would have flipped the decision. We synthesize the boundary b_i for each $\{\mathrm{TR}_i = 1\}$ unit as $b_i = t_i - 2\,\widehat{\sigma}_T$, where $\widehat{\sigma}_T$ is the empirical standard deviation of CIT in the cohort. This fixes the boundary at two CIT-standard-deviations below the factual treatment, a stylized but consistent boundary-faithful recourse construction as the $\{\mathrm{TR}_i = 1\}$ stratum membership remains *real*. Because boundaries are constructed below the factual T_i, the Lynden–Bell inversion runs in its right-truncation form (recourses point downward; denominator $1 - F_{T|\mathbf{x}}(b)$).

Evaluation. An oracle DRNet [75] model with the Lynden–Bell loss and $\lambda_{\mathrm{oracle}} = 1$ is trained on the full obs sample for N_{oracle} epochs with patience-based early stopping; this serves as the reference dose-response surface for evaluating all baselines on a held out set via MISE/AMSE/DPE. Each variant is trained for at most half the oracle's converged epoch budget ($N_{\mathrm{variant}} = \lfloor N_{\mathrm{oracle}}/2 \rfloor$) and at a λ strictly lower than $\lambda_{\mathrm{oracle}}$. The nuisance component for the Lynden–Bell weights are estimated empirically as further explained in Section E.1.2

E.1.2 Models

Nuisance estimation for Lynden–Bell weights. The weight $\hat{w}_i = 1/\hat{F}_{T|\mathbf{x}_i}(b_i^-)$ requires a single nuisance component, the conditional treatment CDF. We report results under two nuisance regimes:

- Estimated nuisance. $\hat{F}_{T|\mathbf{x}}(\cdot)$ is fit as a Gaussian with linear conditional mean (least-squares regression of T on $\mathbf{x}$) and homoscedastic residual variance, then truncated to the observable interval $[T_{\min}, T_{\max}^{\mathrm{obs}}]$. This regime is mildly misspecified under the Gumbel threshold noise of (38) and reflects a practitioner-friendly setting.

- Analytical nuisance (oracle). $\hat{F}_{T|\mathbf{x}}(\cdot)$ uses the true truncated-Gaussian parameters $(\boldsymbol{\beta}_T, \sigma_T)$ of (37). This isolates Lynden–Bell -inversion error from nuisance-estimation error.

Plug-in weights are clipped at the 95th percentile [17, 18].

The synthetic experiments use the analytical-nuisance regime to isolate Lynden–Bell -inversion error from nuisance-estimation error. The semi-synthetic UNOS-PTR experiments use the estimated-nuisance regime, since the true SCM is unknown for real data; the nuisance component is fit on the observational sample.

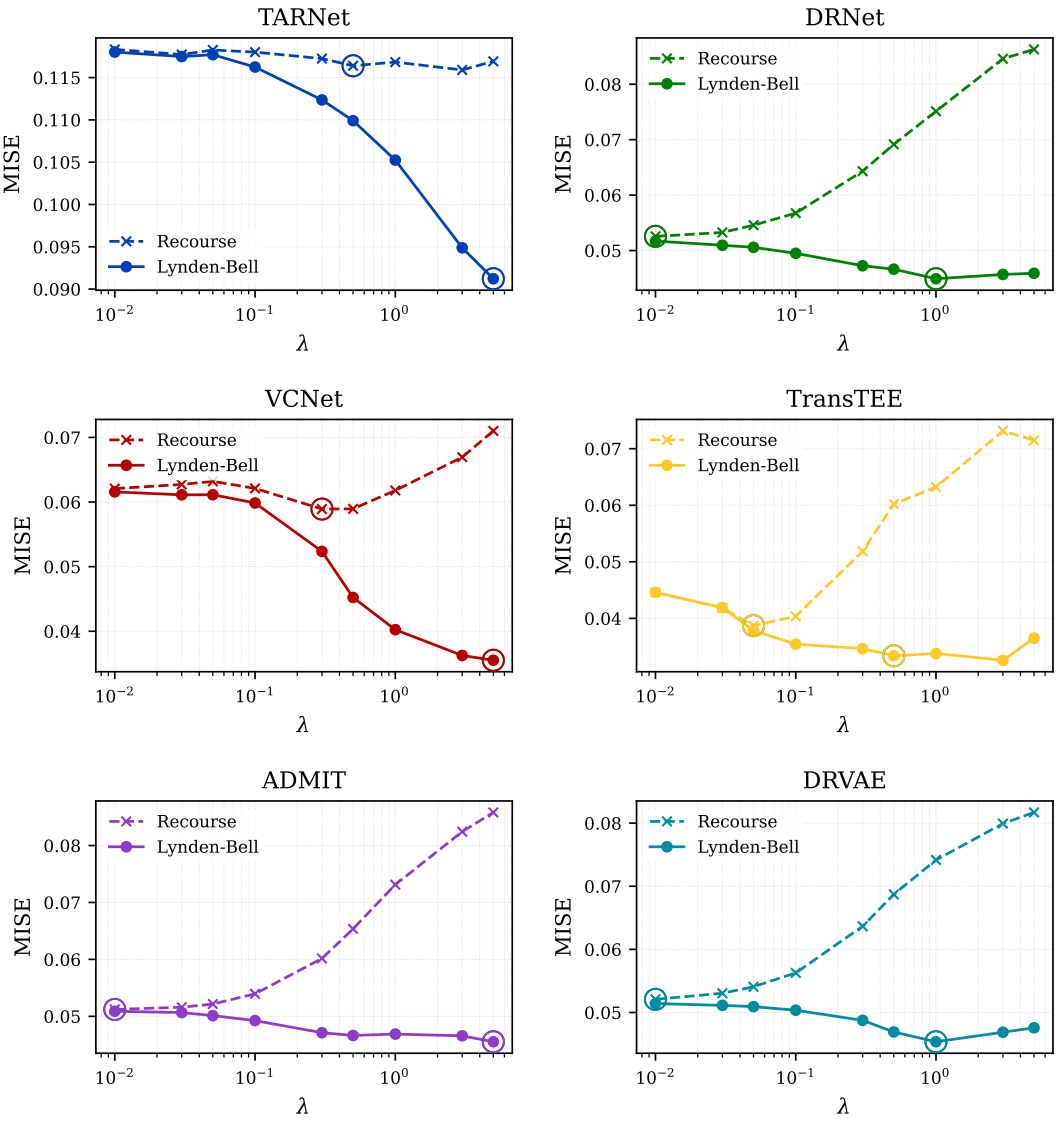

Figure 5: Test MISE on the synthetic DGP as a function of the recourse-loss multiplier λ, for the recourse and Lynden–Bell variants of each model. Circled points mark the optimally tuned λ.

Model hyperparameters. All neural baselines share a common training recipe and backbone-width budget; architecture-specific hyperparameters follow each method's reference implementation. Training uses Adam with learning rate 10^{-3}, batch size equal to the training set ($N_{\text{obs}} = 3000$), maximum 1000 epochs with patience 100 on the validation set ($N_{\text{val}} = 1500$), evaluated every 10 epochs. The auxiliary recourse loss uses $k = 4$ counterfactual draws per unit, resampled at each epoch. Every baseline consists of at least 2 dense layers. If needed, more layers are provided to preserve each architectural blueprint. The recourse-loss multiplier λ is set per (model, variant) pair

by a grid search over $\{0.01, 0.03, 0.05, 0.1, 0.3, 0.5, 1, 3, 5\}$ on validation MISE; the sweep curves and selected values are reported in Figure 5. Architecture defaults:

- TARNET [77]: representation dim 32, head hidden 32.
- DRNET [75]: 4 treatment bins, representation dim 32, head hidden 32.
- VCNET [61]: 4 knots, polynomial degree 2, hidden 32.
- TRANSTEE [99]: model dim 32, 4 heads, 1 encoder layer.
- ADMIT [89]: 4 bins, representation dim 32, head hidden 32, IPM weight 0.5.
- DRVAE [19]: 4 bins, representation dim 32, head hidden 32, KL weight 0.01.
- Standalone estimators (Recourse-only): 2-hidden-layer MLP with hidden width 5, taking $[\mathbf{x}, t] \in \mathbb{R}^{d_x+1}$ as input and producing a scalar logit; trained with on the auxiliary recourse loss alone.

F Acknowledgments

The organ-patient pair data reported in this work has been supplied by the United Network for Organ Sharing as the contractor for the Organ Procurement and Transplantation Network. The interpretation and reporting of the data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the OPTN or the U.S. Government.

This work was supported by the Research Foundation — Flanders (FWO) under grant number G0AAC25N, through the project *Causal Optimization of Data-Driven Healthcare Processes*.

Finally, we would like to thank the NeurIPS reviewers and area chair for their valuable feedback.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Sections 2, 3, 4 and 6, deliver what is promised in the abstract.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed as assumptions are introduced.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Section 3 and Appendix B provide the necessary theory and proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All code is made available for independent assessment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A link to the code for the experiments and the implementation of our methodology. All relevant experimental details are included in the Appendix. UNOS data needs to be requested to OPTN, authors cannot share this data.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All the settings are explained in Appendix E. Moreover, the code is made directly accessible.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appropriate variances are shown in the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: They are discussed in Appendix E

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethic.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All use models and datasets are correctly cited and credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer:[N/A]

    Justification: We introduce an auxiliary loss for other models. The code is made available.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [No]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: While an LLM may be used to extract recourse information from text data, it is not core to our methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.