N-OF-1 SURROGACY: EVALUATING AND LEARNING SURROGATE MAPS FOR INDIVIDUAL-LEVEL TREATMENT DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

The N-of-1 study design, i.e., experimentation on a single individual, is the gold standard for fully personalized treatment decisions. In practice, however, the primary outcome of interest can be too delayed or costly to measure, and is often observed only once, making an individual's counterfactual outcomes under different treatments unavailable. Treatment decisions must therefore rely on intermediate measurements, known as surrogate outcomes. Existing surrogate frameworks establish validity with respect to population-average treatment effects. However, population-level surrogacy need not imply correct *individual-level* decisions. We formalize individual-level surrogacy through the regret of the treatment policy induced by a surrogate map, a function of intermediate measurements. Because individual-level regret depends on unobservable subject-level counterfactual outcomes, we propose an evaluation framework based on outcome imputation and establish identifiability of the individual-level regret under suitable assumptions. We propose an estimator that targets the individual-level regret of surrogate maps using observed data, and further discuss when and how surrogate maps can be learned from data. In controlled experiments, we validate the finite-sample behavior of our evaluation procedure and its behavior under assumption violations. Finally, we apply the framework to longitudinal data from a large-scale wearable health study, and evaluate fixed biomarkers and learned surrogate maps as individual-level surrogates for the effect of changes in physical activity on the onset of cardiometabolic disease.

1 INTRODUCTION

Learning optimal treatment decisions at the level of a single individual is the central goal of personalized decision-making. The N-of-1 trial (Guyatt et al., 1990; Lillie et al., 2011) is the canonical experimental design for this goal. Rather than estimating an average effect over a population, an N-of-1 trial randomly exposes a single subject to different treatments to identify the strategy best tailored to that individual. This design spans various domains, from experimental physiology to individualized recommendations in digital advertising. A central challenge is that the true outcome of interest can typically be observed only once per subject, and only after a long delay, e.g., life expectancy, onset of a chronic disease, or lifetime spending. Because an individual decision has to be made before the true outcome is observed, one has to replace the true outcome by a surrogate derived from intermediate measurements.

Formal criteria for surrogacy have been developed almost exclusively at the population level. The earliest are statistical surrogacy criteria (Prentice, 1989; Freedman et al., 1992), which have been shown to be prone to the *surrogate paradox*. Meta-analytic approaches (Buyse et al., 2000) evaluate a surrogate across multiple randomized trials by correlating trial-level average treatment effects on the surrogate and the primary outcome, and are the recommended standard for surrogate validation (Bujkiewicz et al., 2026). Causal frameworks aim to define surrogacy in terms of criteria on underlying potential outcomes or causal graphs; examples include principal surrogacy (Frangakis & Rubin, 2002; Gilbert & Hudgens, 2008), strong surrogacy (Lauritzen et al., 2004), and transportability-based definitions (Pearl & Bareinboim, 2011). Notably, all of these criteria define surrogate validity with respect to population-level treatment effects. Such criteria need not transfer to the N-of-1 de-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sign due to individual-level hidden confounding which cannot be ignored on the subject level. In aggregated analyses, subject-level confounding can be marginalized over the population, and a surrogate can remain valid on a population level if such confounding is present. In an N-of-1 design, individual-level hidden confounding cannot be removed by aggregation. Consequently, a surrogate that is valid on average can fail at the individual level, and vice versa (as illustrated in Appendix A.3). This gap calls for a notion of surrogacy with respect to individual-level treatment decisions, which we term *individual-level surrogacy*. We formulate the following desiderata for individual-level surrogacy: (i) an individual-level surrogate should correctly *rank* a set of treatments $\mathcal{A}^j$ performed by a single subject j, but need not accurately estimate the magnitude of corresponding effects, (ii) an individual-level surrogate should generalize to new N-of-1 subjects, (iii) it should generalize to new *treatments* for which no population-level data is available.

In many applications, no single scalar surrogate endpoint can be pre-specified. Instead, one has access to a high-dimensional vector of short-term measurements, from which a scalar surrogate must be derived. Such a vector could be a collection of biomarkers in health applications, or short-term engagement metrics in advertising. A growing literature aims to construct such scalar surrogates from data via learning the best predictor of the outcome given the intermediate measurements, termed the *surrogate index* (Athey et al., 2026; Chen & Ritzwoller, 2023). These works impose Prentice-style surrogacy conditions and, similarly to the meta-analytic approach, evaluate how well the learned surrogate approximates average treatment effect across studies. Instead, we study the evaluation and learning of *surrogate maps* for the purpose of individual-level treatment ranking, and do not impose Prentice surrogacy conditions. We defer a more detailed comparison to prior surrogacy frameworks and N-of-1 literature to Appendix A.1.

Evaluation of a surrogate map on an individual level is a hard problem for multiple reasons: for an N-of-1 subject exposed to multiple treatments, the true potential long-term outcomes are often unobservable, and need not be well-defined. A surrogate can therefore never be validated against a subject's own outcomes, and must instead be assessed by combining the subject's intermediate responses with data from other individuals. In this work, we develop concrete conditions under which validity of a surrogate on an individual level can be tested from such data.

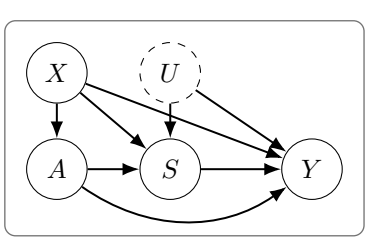
We make the following contributions: In Section 2, we formalize the problem of individual-level surrogacy and propose the notion of *individual-level regret* of a surrogate map. In Section 3, we establish identifiability of individual-level regret from N-of-1 and population data under explicit assumptions, and propose a concrete evaluation pipeline via imputation of outcome labels. In Section 3.2, we establish conditions under which our evaluation guarantees that a surrogate map correctly ranks *new* treatments performed on unseen subjects. In Section 4, we validate our framework on synthetic data and study the finite-sample behavior of our estimator. Finally, in Section 5, we apply our framework to longitudinal health data from the Apple Heart and Movement Study, and produce an evaluation of wearable markers as well as learned surrogate maps as individual-level surrogates for a time-to-event cardiometabolic outcome. Additionally, in Appendix A.7, we discuss how surrogate maps can be learned, show suboptimality of the surrogate index as an individual-level surrogate, and establish an optimal surrogate map under additional assumptions.

2 Setting

2.1 Data-generating process

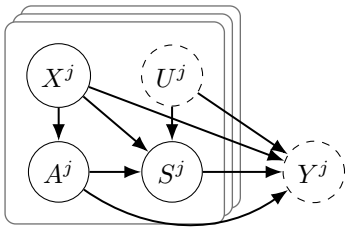
We consider access to two study structures: a population-level observational study, and a set of subject-level N-of-1 experimentation studies. We leverage the population study to characterize relationships between covariates X, treatments A, surrogates S, and the long-term outcome Y. In turn, we aim to leverage the subject-level N-of-1 study to provide individualized treatment recommendations. Formally, we assume both study structures are governed by a structural causal model (SCM) describing the relationships between observed covariates $X \in \mathcal{X} \subseteq \mathbb{R}^m$, treatments $A \in \mathcal{A}$, surrogate variables $S \in \mathcal{S} \subseteq \mathbb{R}^d$, and the long-term outcome $Y \in \mathbb{R}$: the population study by the SCM $\mathcal{C}^{\text{obs}}$ (Figure 1a), and each subject $j \in \mathcal{J}$ by a *subject-specific* SCM $\mathcal{C}^j$ (Figure 1b).

We assume the causal mechanisms g_0, f_0, as well as the distribution of the independent noise variables $\epsilon_A, \epsilon_Y \sim \mathbb{P}_\epsilon$, are shared across subjects, while the distribution of observed covariates X^j, and, importantly, the distribution of the unobserved confounding $U^j \in \mathbb{R}^l$ may be subject-specific.



$$\mathcal{C}^{\text{obs}}: \begin{cases} X \sim \mathbb{P}^{\text{obs}}_X,\ U \sim \mathbb{P}^{\text{obs}}_U; \\ A \coloneqq \pi_0(X, \epsilon_A),\ S \coloneqq g_0(A, X, U); \\ Y \coloneqq f_0(A, S, X, U, \epsilon_Y). \end{cases}$$

(a) Population study $\mathcal{C}^{\text{obs}}$



$$\mathcal{C}^{j}: \begin{cases} X^j \sim \mathbb{P}^j_X,\ U^j \sim \mathbb{P}^j_U; \\ A^j \coloneqq \pi_0(X^j, \epsilon_A),\ S^j \coloneqq g_0(A^j, X^j, U^j) \\ Y^j \coloneqq f_0(A^j, S^j, X^j, U^j, \epsilon_Y). \end{cases}$$

(b) N-of-1 experiment on subject j

Figure 1: Data-generating process of a population observational study (a) and within-individual experimentation (b). Dashed nodes are unobserved. Rounded boxes illustrate the number of samples per subject: in the population study (a), each subject is treated once. In (b), (X^j, A^j, S^j) is drawn repeatedly within subject, but the long-term outcome Y^j is not observed.

We denote by $\mathbb{P}^{\text{obs}}$ the joint distribution of the observed variables (X, A, S, Y) in the population study. For each N-of-1 subject $j \in \mathcal{J}$, we denote by $\mathbb{P}^j$ the joint distribution of the variables (X^j, A^j, S^j, Y^j) induced by the subject-specific causal model $\mathcal{C}^j$, and by $\mathbb{E}^j$ and Var^j the corresponding expectation and variance. Further, for any $B^j \in \{A^j, X^j, S^j, Y^j\}$, we denote by $\text{supp}(B^j)$ the support of B^j in $\mathbb{P}^j$. Let $\mathcal{A}^j \coloneqq \text{supp}(A^j)$ be the set of treatments to which the N-of-1 subject j was exposed. For any treatment $a \in \mathcal{A}^j$, we define the subject-level potential outcomes $S^j(a) \coloneqq g_0(a, X^j, U^j)$ and $Y^j(a) \coloneqq f_0(a, S^j(a), X^j, U^j, \epsilon_Y)$ of the surrogate vector and the primary outcome, respectively. Note that under our data-generating process, subject-level potential outcomes are generally random variables. For the rest of the manuscript, we assume integrability of all potential outcomes both under $\mathbb{P}^{\text{obs}}$ and $\mathbb{P}^j$.

The following example provides concrete intuition for our data-generating process:

Example 1 (N-of-1 setting in online experimentation)**.** *We are given a collection $\mathcal{J}$ of users of an e-commerce platform. Each subject j is exposed to a collection $\mathcal{A}^j$ of advertising campaigns for short time periods one by one. This collection can be fully randomized or influenced by X^j, comprised of the user's demographics (age, sex, socioeconomic level etc.) and some varying context (season, holidays, inflation). After each short time period of exposure, an intermediate response $S^j(a)$ (such as number of clicks, items in the basket and more) to each campaign $a \in \mathcal{A}^j$ is collected. This intermediate response is heterogeneous across users and is influenced by the user's latent type U^j (e.g., holiday shopper, price conscious, trend follower etc.). This type is unobserved and cannot be determined from X^j, but can be crucial in determining the long-term effectiveness of a campaign on a particular subject. Our objective is to find a campaign a which maximizes* this particular *subject's lifetime spending $Y^j(a)$. However, for the N-of-1 subject, on which experimentation is performed, no data on lifetime spending is available by definition. Our goal is thus to rely on the collection $\{S^j(a)\}_{a \in \mathcal{A}^j}$ of the subject's short-term responses, as well as an independent longitudinal customer study $\mathbb{P}^{\text{obs}}$, to determine the individual-level best campaign.*

2.2 Identifying individual-level best treatments via surrogate maps

We now formalize the individual-level quantity our framework targets. For a subject j and a treatment $a \in \mathcal{A}^j$, write $\mu^j(a) \coloneqq \mathbb{E}^j[Y^j(a)]$ for subject j's expected primary potential outcome under a. These induce a total ground-truth *ranking* of the treatments subject j attempted,

$$a \succeq^j a' \quad :\Longleftrightarrow \quad \mu^j(a) \geq \mu^j(a'), \qquad a, a' \in \mathcal{A}^j, \tag{1}$$

where the maximal elements $\mathcal{A}^*_j \coloneqq \text{argmax}_{a \in \mathcal{A}^j} \mu^j(a)$ are the *individual-level best treatments*. Our goal is to recover the treatment ranking (1): for every pair of treatments the subject has tried, we want to correctly decide which yields the better long-term outcome *for that individual*. Recovering the ranking also identifies the set $\mathcal{A}^*_j$ of best treatments.

As previously stated, many primary outcomes of interest can be observed at most once per subject, and cannot be observed within the short time scale of a within-subject experiment $\mathcal{C}^j$. Recovering (1) directly is therefore infeasible, and our approach is instead to rely on the subject's covariates, as well as intermediate measurements $\{S^j(a)\}_{a\in\mathcal{A}^j}$ collected after each exposure $a \in \mathcal{A}^j$. Since these measurements are vector-valued, we require an appropriate scalarization to define an individual-level treatment ranking. We call such a scalarization a *surrogate map*:

Definition 1 (Surrogate map)**.** *A valid surrogate map is a measurable function $f\colon \mathcal{S}\times\mathcal{X}\to\mathbb{R}$ such that $\mathbb{E}^j\left[\,|f(S^j(a),X^j)|\,\right] < \infty$ for every $j \in \mathcal{J}$ and every $a \in \mathcal{A}^j$.*

A surrogate map assigns to each subject j and each attempted treatment $a \in \mathcal{A}^j$ the *surrogate value* $\mathbb{E}^j\left[f(S^j(a),X^j)\right]$, and thereby induces its own ranking of the individual's treatments $\mathcal{A}^j$:

$$a \succeq_f^j a' \quad :\Longleftrightarrow \quad \mathbb{E}^j\left[f(S^j(a),X^j)\right] \geq \mathbb{E}^j\left[f(S^j(a'),X^j)\right].$$

Equivalently, f induces a treatment-assigning *policy* on every subset of treatments $\mathcal{B}\subseteq\mathcal{A}^j$ the subject may be asked to choose from via $\pi_f(j;\mathcal{B}) \coloneqq \operatorname{argmax}_{a\in\mathcal{B}}\mathbb{E}^j\left[f(S^j(a),X^j)\right]$. We propose measuring the value of a surrogate map f as an *individual-level surrogate* through the long-term cost of the decisions that it makes for subject j. More concretely, we introduce the notion of individual-level regret, which measures the average cost of picking the best treatment according to the surrogate-induced policy as compared to the true best treatment, out of every *pair* of the subject's attempted treatments:

Definition 2 (Individual-level regret)**.** *For a surrogate map f, a subject $j \in \mathcal{J}$ and any subset of attempted treatments $\mathcal{B}\subseteq\mathcal{A}^j$, the individual-level regret of the surrogate policy π_f on $\mathcal{B}$ is*

$$r^j(f;\mathcal{B}) \coloneqq \max_{a\in\mathcal{B}}\mu^j(a) \;-\; \min_{a'\in\pi_f(j;\mathcal{B})}\mu^j(a') \;\geq\; 0. \tag{2}$$

The total individual-level regret *of f for subject j is the average of (2) over all pairs of treatments attempted by j,*

$$r^j(f) \coloneqq \binom{|\mathcal{A}^j|}{2}^{-1} \sum_{\{a,a'\}\subseteq\mathcal{A}^j,\; a\neq a'} r^j\big(f;\{a,a'\}\big). \tag{3}$$

Each summand of (3) has the interpretation of a regret: a single term $r^j(f;\{a,a'\})$ is the long-term cost to subject j of following the surrogate-based recommendation rather than the truly best of the two treatments. This regret could be, for instance, the missed years of life expectancy or the missed lifetime spending incurred on that comparison. We discuss the pairwise choice for (3) in Appendix A.2. If $S^j(a)$ is constant, $r^j(f)$ is invariant under strictly increasing transformations of f such as a constant offset. In particular, our notion of regret does not require that a surrogate map is calibrated to the scale of Y or accurately estimates the average treatment effect. Next, for a group of N-of-1 subjects $j \in \mathcal{J}$, we can define the *average individual-level regret* of a surrogate map as

$$R_{\mathcal{J}}(f) \coloneqq \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} r^j(f). \tag{4}$$

We call f a *perfect individual-level surrogate map* on $\mathcal{J}$ if $R_{\mathcal{J}}(f) = 0$. In Appendix A.2, we show that a perfect individual-level surrogate map recovers the true treatment ranking for every subject $j \in \mathcal{J}$, and correctly identifies the individual-level best treatment(s).

3 Evaluation of individual-level surrogate maps

In the previous section, we introduced the individual-level regret $r^j(f)$ as a measure of individual-level surrogacy, and argued why a perfect individual-level surrogate map should correctly rank treatments on an individual level. However, the regret $r^j(f)$ is defined through the subject-level potential outcomes $\mu^j(a)$, which are not observable in an N-of-1 experiment: the long-term outcome is often realized at most once per subject, cannot be observed within the experiment's short time scale, and need not be well defined for a subject exposed to several treatments in sequence. We therefore replace the labels $\mu^j(a)$ by *pseudo-outcomes* imputed from the population observational study, and give conditions under which the resulting *pseudo-regret* coincides with the true individual-level regret. For simplicity, we take covariates X^j to be constant within a subject. In Appendix A.6, we relax this to covariates with a fixed part X^j_{fix} and a within-subject varying part X^j_{rand}. Thus, our results cover within-subject non-randomized studies.

3.1 Pseudo-outcomes and individual-level pseudo-regret

Given N-of-1 subject j with covariates X^j and surrogate responses $\{S^j(a)\}_{a\in\mathcal{A}^j}$ to treatments $\mathcal{A}^j$, we analyze under which conditions the subject's individual-level regret can be identified from their own intermediate data and the population observational study $\mathbb{P}^{\text{obs}}(A, X, S, Y)$ (cf. Section 2) containing different subjects for whom we observe long-term outcomes, but no repeated experiments.

To enable estimation of the N-of-1 subject's long-term outcomes from the observational study, we require that similar intermediate responses have been observed in the population study:

Assumption 1 (Support inclusion)**.** *It holds that* $\big(a, X^j, S^j(a)\big) \in \operatorname{supp} \mathbb{P}^{\text{obs}}_{A,X,S}$ *for every* $a \in \mathcal{A}^j$.

Denoting the set of treatments covered by the observational study as $\mathcal{A}^{\text{obs}} \coloneqq \operatorname{supp} \mathbb{P}^{\text{obs}}_{A}$, we note that Assumption 1 requires $\mathcal{A}^j \subseteq \mathcal{A}^{\text{obs}}$. This allows us to define the study-based *pseudo-outcomes* of subject j as

$$\tilde{\mu}^j(a) \coloneqq \mathbb{E}^{\text{obs}}\left[\, Y \mid A = a,\ X = X^j,\ S = S^j(a) \,\right], \qquad a \in \mathcal{A}^j. \tag{5}$$

Substituting them for μ^j in Definition 2 yields the regret $\tilde{r}^j(f; \mathcal{B}) \coloneqq \max_{a\in\mathcal{B}} \tilde{\mu}^j(a) - \min_{a'\in\pi_f(j;\mathcal{B})} \tilde{\mu}^j(a')$ of f on any subset of treatments $\mathcal{B} \subseteq \mathcal{A}^j$, and the resulting total *individual-level pseudo-regret* of subject j

$$\tilde{r}^j(f) \coloneqq \binom{|\mathcal{A}^j|}{2}^{-1} \sum_{\{a,a'\}\subseteq\mathcal{A}^j, a\neq a'} \tilde{r}^j\big(f; \{a, a'\}\big). \tag{6}$$

Analogously to Section 2.2 we can define the average pseudo-regret over an N-of-1 population. We require one more assumption to enable identifiability of the true individual-level regret:

Assumption 2 (Invertibility and stability of the latent type)**.** *For every* $j \in \mathcal{J}$*, the latent type* U^j *is constant* $\mathbb{P}^j$*-a.s. Further, for every* $(a, x, s) \in \operatorname{supp} \mathbb{P}^{\text{obs}}_{A,X,S}$*, the map* $g_0(a, x, \cdot)$ *is injective and admits a measurable left inverse* $u = g_0^{-1}(a, x, s)$.

Intuitively, observing a subject's covariates, treatment, and surrogate response recovers their latent type U^j. Assumption 2 does not imply unconfoundedness, nor that U^j is recoverable from (X, S, Y) alone: for example, knowing which intervention produced an observed change in fitness is more informative than the change alone. We are now ready to state our identifiability result:

Theorem 1 (Identifiability of individual-level regret)**.** *Let the N-of-1 subject* j *and the observational population study* $\mathbb{P}^{\text{obs}}(A, X, S, Y)$ *satisfy Assumptions 1 and 2. Then the potential outcome means of subject* j *are identified by the pseudo-outcomes* (5)*,*

$$\mu^j(a) \;=\; \tilde{\mu}^j(a) \quad \textit{for all } a \in \mathcal{A}^j, \qquad \textit{and hence} \qquad r^j(f) = \tilde{r}^j(f). \tag{7}$$

We prove Theorem 1 in Appendix A.10. In Appendix A.4.1, we provide a way to estimate individual-level pseudo-regret from finite-sample data via Algorithm 1. In Appendix A.4.2, we show that our estimator is consistent and study its finite-sample properties. The main focus of our work is evaluation of surrogate maps. However, a natural question is when and how good individual-level surrogate maps can be learned. In Appendix A.7, we discuss how surrogate maps can be learned in different data settings. In Appendix A.7.3, we show that under the additional assumption of additive confounding, the *causal surrogate map* $f_{\text{causal}}(s, x) = \mathbb{E}^{\text{obs}}[Y \mid X = x, \text{do}(S = s)]$ is optimal.

3.2 Generalization to new treatments

Section 3.1 demonstrates how any surrogate map can be evaluated on a set of already observed treatments $\mathcal{A}^{\text{obs}}$. However, a surrogate map is only useful if, after evaluation, it can be deployed on new subjects and, importantly, new treatments (such as a novel exercise routine or new advertising campaign) for which no population data has been collected. In practice, individual-level treatments can almost always be assumed novel, since, for instance, the chance that a new subject replicates exactly one of observed exercise routines is negligible. Even within a well-defined family of medications, the treatments of practical interest are precisely new doses with no outcome data. Thus, it is reasonable to consider a new set of treatments $\mathcal{A}^{\text{tst}}$ with $\mathcal{A}^{\text{tst}} \cap \mathcal{A}^{\text{obs}} = \varnothing$.

First, observe that no evaluation procedure can generalize to an *arbitrary* new treatment: a treatment that bypasses the surrogate vector and only acts on Y directly can incur unbounded regret (cf. Appendix A.5). We therefore restrict our attention to treatments with no direct effect:

Assumption 3 (Full mediation). *Let* $m(a, s, x, u) \coloneqq \mathbb{E}_{\epsilon_Y}[f_0(a, s, x, u, \epsilon_Y)]$. *There is a measurable function* $\bar{m}\colon \mathcal{S} \times \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ *such that* $m(a, \cdot) = \bar{m}(\cdot)$ *for every* $a \in \mathcal{A}^{\text{obs}} \cup \mathcal{A}^{\text{tst}}$. *Equivalently, this corresponds to removing the direct edge* $A \to Y$ *in Figure 1.*

Assumption 3 is weaker than the classical no-direct-effect assumption for scalar surrogate endpoints (Prentice, 1989): it requires mediation through the full surrogate space, not through any candidate map $f(S, X)$ (cf. Appendix A.5). As more intermediate measurements are collected, Assumption 3 becomes more plausible. Note that this also makes it more likely that S and Y are confounded, making the Prentice-style assumption of full mediation and ignorability (e.g. in Athey et al. (2026)) less plausible than just full mediation as in our work.

We work with two populations of N-of-1 subjects. On the *evaluation-time population* $\mathcal{J}^{\text{eval}}$ we can estimate a map's individual-level regret, whereas the *test-time population* $\mathcal{J}^{\text{tst}}$ is exposed to new treatments $\mathcal{A}^{\text{tst}}$ not covered by $\mathbb{P}^{\text{obs}}$, so regret cannot be computed there. This is the population on which we ultimately wish to deploy the map. We ask when low pseudo-regret on $\mathcal{J}^{\text{eval}}$ implies low true regret on $\mathcal{J}^{\text{tst}}$. We observe that under Assumption 3, a treatment enters the regret only through its surrogate response, and, for any treatment pair (a, a'), can be expressed as

$$r^j(f; \{a, a'\}) = \rho_f\big(X^j,\, U^j,\, S^j(a),\, S^j(a')\big),$$

where ρ_f is a fixed function of (X, U, S, S'), and *does not* depend on the treatments. A map's regret is therefore governed entirely by the distribution of surrogate contrasts $(X^j, U^j, S^j(a), S^j(a'))$ within the N-of-1 population regardless of which treatments were applied. We formally show in Appendix A.5 that, under Assumption 3, perfect evaluation-time pseudo-regret implies perfect test-time regret as soon as the pairs of surrogate responses observed during evaluation are sufficiently diverse. In particular, the causal mechanism $g_0(a, \cdot)$ of the new treatment $a \in \mathcal{A}^{\text{tst}}$ can be arbitrary.

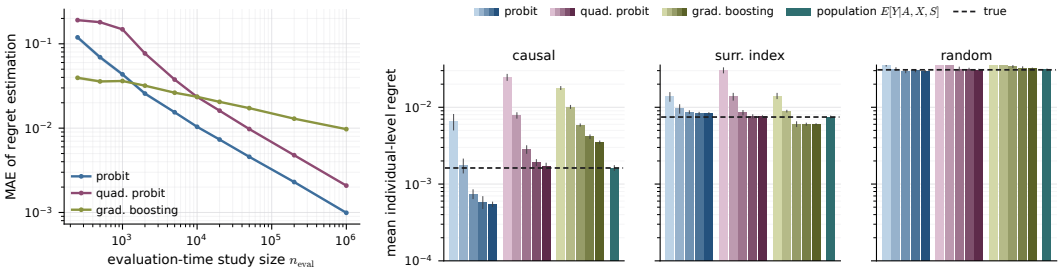
4 Synthetic Experiments

We validate our evaluation framework on a synthetic data-generating process (DGP) modeling the online-experimentation setting of Example 1. Because the DGP is known in closed form, we can compute true individual-level regret, allowing us to investigate the validity of our pseudo-regret as an estimand, its finite-sample behavior, and its behavior under misspecification.

Data-generating process. We consider a nonlinear structural causal model of the treatments A, covariates X, surrogates $S \in \mathbb{R}^6$, and binary outcomes $Y \in \{0, 1\}$ outlined in Appendix A.8. The shared structural parameters of the SCM are drawn once independently and shared by every subject and every split. We independently sample two disjoint treatment sets: the training-time treatments $\mathcal{A}^{\text{tr}}$ and evaluation-time treatments $\mathcal{A}^{\text{eval}}$.

Evaluation. We sample the following datasets: $D^{\text{tr}} = \{(A_i, X_i, S_i, Y_i)\}_{i=1}^{n_{\text{tr}}}$ containing observational data with treatments in $\mathcal{A}^{\text{tr}}$, used to learn surrogate maps (cf. Appendix A.7); $D^{\text{obs}} = \{(A_i, X_i, S_i, Y_i)\}_{i=1}^{n_{\text{eval}}}$, containing observational data with treatments in $\mathcal{A}^{\text{eval}}$, used to learn the outcome regression function at evaluation time; and a collection $\{D^j\}_{j \in \mathcal{J}^{\text{eval}}}$ of N_{eval} N-of-1 evaluation subjects exposed to evaluation treatments $\mathcal{A}^{\text{eval}}$ on which we compute the individual-level pseudo-regret. We evaluate the following surrogate maps: the *surrogate index* $f_{\text{SI}}(s, x) \coloneqq \mathbb{P}^{\text{tr}}(Y = 1 \mid S = s, X = x)$ (Athey et al., 2026), the causal surrogate map $f_{\text{causal}}(s, x) \coloneqq \mathbb{P}^{\text{tr}}(Y = 1 \mid \text{do}(S = s), X = x)$ (cf. Appendix A.7.3), and a fixed random linear functional of (S, X) as a reference. For the outcome model, we consider three different estimators: pooled probit regression with treatment-specific coefficients for (S, X) and pooled coefficients for $S \otimes X$; a more expressive quadratic probit regression, to which we add treatment-specific coefficients for $S \otimes X$ and $S \otimes S$; and *gradient boosting* as a nonparametric comparison.

In Figure 2a, we consider the instantiation of the synthetic DGP (26) with additive confounding (cf. Appendix A.7.3), in which case the probit regression estimator is correctly specified. We report the mean absolute error of regret estimation $\text{MAE} = \frac{1}{|\mathcal{J}_{\text{eval}}|} \sum_j |r^j(f) - \hat{r}^j(f)|$ induced by the three outcome models, where f is taken to be the surrogate index, and we vary n_{eval}. The estimation error vanishes as n_{eval} grows, at a rate inherited from the outcome model, confirming that the pseudo-regret is a consistent estimator of the true individual-level regret (Corollary 1).



(a) Regret estimation error vs. n

(b) True vs. estimated regret for three surrogate maps

Figure 2: (a) Regret estimation error (evaluated at the surrogate index f_{SI}) as a function of the observational study size n_{eval} (log-log). (b) Estimated and true individual-level regret of the surrogate index $f_{\text{SI}}(s, x)$, the causal map $f_{\text{causal}}(s, x)$ and the random map at n_{eval} growing from 5k (lightest shade) to 1M samples (darkest shade), for three different outcome models. Population pseudo-regret is shown in teal. Dashed lines show the true individual-level regret of each map.

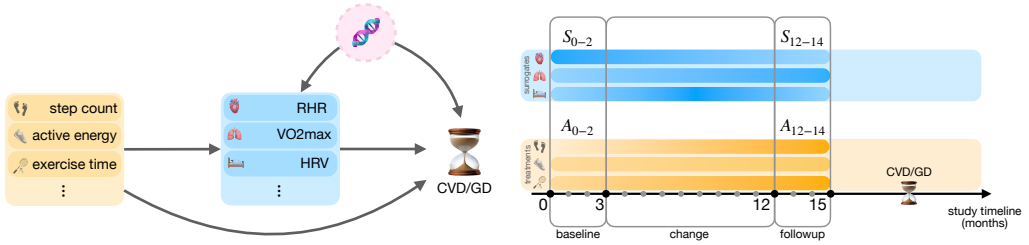
In Figure 2b, we consider a more general version of our synthetic DGP (26) with non-additive confounding, in which case the causal map incurs non-zero regret (cf. Appendix A.7.4), and show the estimated pseudo-regret as well as the true regret of three surrogate maps (f_{SI}, f_{causal} and the random map) at different sample sizes n_{eval} and for different outcome models. The regret values are *recovered correctly* by the correctly specified quadratic outcome model as the sample size grows. The maps are *ranked correctly* by the probit models (including misspecified pooled probit) at all considered sample sizes. In Appendix A.8, we outline additional ablation studies: we investigate individual-level regret estimation under violation of Assumption 2 (invertibility), effects of non-additive confounding, goodness of ranking of surrogate maps at a growing sample size, and more.

5 Case Study: The Apple Heart & Movement Study

We illustrate our framework using real-world wearable and health data from the Apple Heart and Movement Study, a large-scale longitudinal cohort of U.S. adults with continuous wearable measurements and survey data recorded over up to 5 years (MacRae, 2021; Truslow et al., 2024). We ask: can we evaluate which combinations of wearable biomarkers are good individual-level surrogates for the effect of lifestyle interventions on cardiometabolic health?

Since AHMS is a single observational study, we first map it onto our causal framework in Section 2: we segment each participant's time series data into *change windows* (cf. Appendix A.9.2 for details). Each change window contains a baseline, change and followup period (illustrated in Figure 3b). As treatment labels for each window, we consider discretized within-window changes in one of the six daily activity metrics (step count, walking/running distance, exercise time, stand time, active energy burned, flights climbed). As the surrogate vector, we consider standardized within-window change in six wearable biomarkers (resting heart rate, heart rate variability, VO_2max, walking heart rate, six-minute walk distance, walking speed) and, additionally, 256-dimensional embeddings of a photoplethysmogram (PPG) foundation model (Abbaspourazad et al., 2024). Our outcome of interest is the time until first onset of a cardiometabolic event: a composite of coronary and peripheral artery disease, myocardial infarction, coronary revascularization, heart failure, stroke, and type-2 diabetes. We combine these outcomes since their incidence is consistently associated with physical activity and lifestyle (Lee et al., 2011; Kyu et al., 2016). We describe our cohort filtering in Appendix A.9.1 and provide more details on the treatments, surrogates and outcomes in Appendix A.9.2.

A central difference from a true N-of-1 trial is that behavior changes in AHMS are not randomized: participants have activity patterns based on motivation, health status and more. We partially mitigate this confounding by our change-from-baseline design and use of change windows. Additionally, we adjust for subject-level covariates such as baseline age, socioeconomic status etc. (Appendix A.9.2), and within-subject varying covariates such as current age, current season and time since enrollment (Appendix A.6). We assume the residual confounding to be captured in some latent type U^j as in Section 2 (e.g., physiology or genetics). The resulting causal structure of the AHMS study is depicted in Figure 3a.



(a) Causal structure of the study

(b) Temporal structure of the study

Figure 3: (a) Mapping of the AHMS study onto the causal setting in Section 2. (b) Illustration of a 15-month within-subject change window. The treatment label is the change in an activity metric, the surrogate vector is the change in wearable markers, the outcome is time-to-onset of cardiometabolic disease (cardiovascular disease or/and diabetes mellitus type 2).

Candidate surrogate maps. We fit multiple surrogate maps on a separate observational cohort $D^{\text{tr}} = \{(A_i, S_i, X_i, Y_i)\}_{i=1}^{n_{\text{tr}}}$. We fit the maps on the surrogate space comprised of within-window wearable biomarker changes, as described above. First, we note that wearable biomarkers (metrics) can be evaluated on their own as a univariate surrogate map $f(S, X) = S_k$. This allows us to evaluate, e.g., the change in VO_2max (mL/kg/min) as an individual-level surrogate map.

Additionally to the fixed metrics, we learn and evaluate the following surrogate maps: 1) The unadjusted Cox model, which uses only S; 2) the adjusted Cox model, which takes in S and adjustment covariates (age, sex, socioeconomic status, baseline health status and more); 3) a two-stage control function version of the Cox model, which attempts to correct endogeneity of the surrogate space by using A as instruments (Newey et al., 1999). We fit versions of 1), 2) and 3) on the full surrogate space, only the subset of wearable biomarkers (metrics), and only on the subset of PPG embeddings. Additional details on the definition and fitting of the surrogate maps are given in Appendix A.9.4.

5.1 Evaluation procedure and results

On a disjoint part of the study $D^{\text{obs}} = \{(A_i, S_i, X_i, Y_i)\}_{i=1}^{n_{\text{eval}}}$ we fit an L2-penalized Cox model for the composite cardiometabolic outcome using as input A, X, S as in Algorithm 1. We take as the pseudo-label $\hat{\mu}^j(a)$ the negative linear predictor term of the Cox model, since its treatment ranking is identical to survival probability at any time point (Appendix A.9.3). Finally, on a disjoint collection of subjects $\mathcal{J}^{\text{eval}}$ with data $D^j = \{(X_i^j, A_i^j, S_i^j)\}_{i=1}^{n_j}$ from multiple change windows, we compute the individual-level pseudo-regret via imputing $\hat{\mu}^j(a)$ as in Algorithm 1 and average over the N-of-1 subjects. For the final evaluation, we perform cross-fitting with $K = 20$ folds, where in each fold we use 45% of data as D^{tr}, 45% of data as D^{obs} and the remaining 10% of data as $\{D^j\}_{j \in \mathcal{J}^{\text{eval}}}$. Full details are given in Appendix A.9.3.

Ranking of fixed biomarkers as individual-level surrogates. Figure 9 in Appendix A.9.5 shows six wearable biomarkers ranked by their estimated individual-level regret. We find that VO_2max and resting heart rate generally achieve the lowest individual-level regret across all change windows.

Evaluation of biomarker- and PPG-based surrogate maps. In Figure 4, we report the average individual-level regret of the learned surrogate maps relative to the regret of the reference surrogate VO_2max, as we vary the duration of the change period. First, we observe that surrogate maps fit on wearable metrics only have the lowest regret across all change-period durations. Adjustment for baseline covariates (adjusted Cox model) and endogeneity (control-function method) yield consistent improvements over the unadjusted map, which does not account for confounding; moreover, in almost all cases the control-function method attains the lowest regret, in line with our theoretical result (cf. Appendix A.7.3). Second, adding PPG embeddings does not improve performance: full-space and PPG-only maps consistently yield higher regret than metric-only maps. This suggests that the most relevant signal for individual decision making might be already contained in the low-dimensional wearable marker space. Third, the relative improvements over the fixed surrogate (VO_2max) grow larger as the change period increases. This suggests that there is more signal to determine the best treatment as the change-period duration increases.

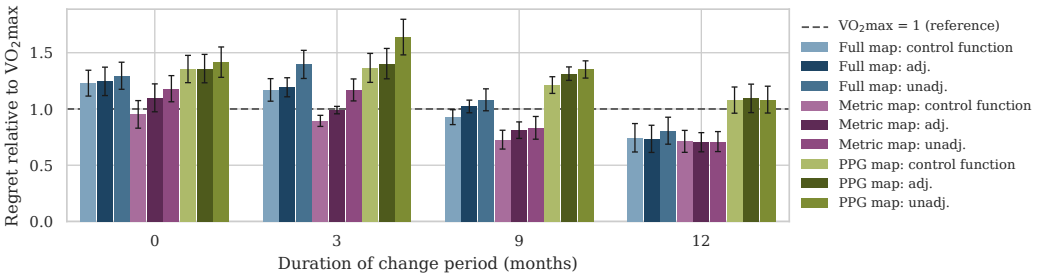


Figure 4: Regret of learned surrogate maps as a function of the change period duration. Each bar represents the mean over 20 cross-fitting folds of the individual-level regret of a learned surrogate map relative to that of VO_2max (values below 1 indicate lower regret than VO_2max). Metric-space maps attain the lowest regrets across all change windows. In most cases, the control function maps yield lower regrets than adjusted and non-adjusted maps.

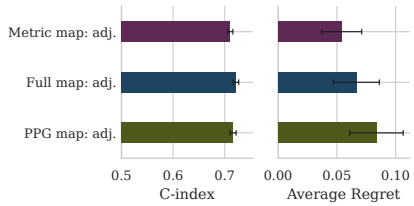


Figure 5: Held-out C-index versus individual-level regret for three surrogate maps with different surrogate input space at the 9-month change period. The result suggests that predictive performance is not indicative of individual-level treatment ranking ability.

Relationship of surrogate map predictive performance and individual-level regret. A natural question is whether a surrogate map's predictive accuracy for the long-term outcome is a reliable proxy for its individual-level treatment-ranking ability. If so, one could simply select maps by standard predictive metrics and bypass our regret-based evaluation entirely. Figure 5 shows that this is not the case. We compare three adjusted-Cox surrogate maps differing only in their surrogate input space (metrics only, full space, PPG only) along two axes: their held-out discriminative performance for the outcome as measured by the concordance index (C-index) (Harrell et al., 1982), and their individual-level regret. In Figure 5, the three maps attain similar c-indices, with the metrics-only map in fact the lowest. At the same time, the metrics-only map achieves the lowest regret, whereas the PPG-only map achieves nearly twice its regret. This reflects the central premise of our work: population-level predictiveness is a statistical property of the surrogate-outcome joint distribution, whereas individual-level surrogacy measures correctness of within-participant ordering of counterfactual outcomes, and the two goals can oppose each other. Our findings thus underscore why a dedicated evaluation framework for individual-level surrogacy is necessary.

6 Discussion

In this work, we formalize individual-level surrogacy in the N-of-1 experimentation setting through the individual-level regret of a surrogate map-induced treatment policy, and establish conditions under which it is identifiable and can be computed from observational data. We thus address a gap in the existing notions of surrogacy, which target estimation of population-level treatment effects. Although our work extends beyond the standard Prentice unconfoundedness assumption, our identifiability result relies on the latent invertibility assumption, which is untestable and might be violated in practice. Even though we provide ablations, an interesting direction for future work remains whether individual-level regret can be robustly evaluated in absence of invertibility. On the example of longitudinal wearable health data, we show how our framework allows us to rank surrogate maps and fixed biomarkers as individual-level surrogates for effects of lifestyle interventions. We note that our proposed evaluation extends beyond the biomedical setting, and can be applied to any context in which a decision-maker must select among actions for a single unit using intermediate measurements while the primary outcome remains unobserved.

AI USE STATEMENT

In this work, we used generative AI tools for refining and proofreading writing, developing and simplifying notation, assisting with minor technical steps such as the generation of numerical toy examples, and refining and checking theoretical results and their proofs. Additionally, we used generative AI tools for code generation for the experiments. We did not use generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. We have reviewed and verified all AI-assisted work. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work includes a secondary analysis of previously collected data from the Apple Heart and Movement Study (AHMS), an observational cohort study, whose participants are U.S. residents aged 18 or older who provided informed consent electronically. Participants control which data types they share, and may withdraw at any time. We did not interact with participants, assign any intervention, or collect any new data. All analyses were performed on de-identified data in accordance with the study protocol and informed consent form. We report aggregate results only.

AHMS participants own a wearable device paired with a smartphone and chose to enroll in a research app. The cohort is therefore not fully representative of the general population. More broadly, analyses conducted on consumer wearable datasets may not generalize to people who do not use such devices.

Our contribution is mostly theoretical and partially methodological. The case study in Section 5 aims to illustrate our framework, not to issue any clinical guidance. The surrogate maps evaluated here should not be used to inform individual health decisions without prospective validation.

REPRODUCIBILITY STATEMENT

All assumptions underlying our results, together with their motivation and illustrative examples, are stated in the main text (Section 3) and the appendix. Complete proofs of all theoretical results are given in the appendix. The data-generating process of Section 4 is fully specified in Appendix A.8, including all numerical parameter values, variable dimensions, and sample sizes needed to reproduce the experiments. Additionally, we are planning to release the full code for reproducing the synthetic experiments.

The data underlying our case study Section 5 cannot be made publicly available due to privacy assurances given in the informed consent signed by study participants and to contractual agreements between the entities conducting the study. Requests to access the data should be directed to the corresponding author. Requests will be reviewed by the study teams under the guidance of the study PI and evaluated under the terms of the informed consent form and the relevant contractual agreements. Code for the case-study analyses may be made available on the same basis, in a manner consistent with policies protecting participant confidentiality.

So that the case study can nonetheless be reproduced on a comparable cohort, we provide details for every processing step in Section 5, Appendix A.9, including the cohort filtering, construction of change windows, computation of behavior labels, the surrogate spaces, definition of the outcome, data splitting, learning of surrogate maps, and more.

REFERENCES

Salar Abbaspourazad, Oussama Elachqar, Andrew C. Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In *ICLR*, 2024. URL https://arxiv.org/abs/2312.05409.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

Artur Araujo, Steven Julious, and Stephen Senn. Understanding variation in sets of n-of-1 trials. *PloS one*, 11(12):e0167167, 2016.

Susan Athey and Guido W. Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006.

Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *The Review of Economic Studies*, 93(4):2284–2312, 07 2026. ISSN 0034-6527. doi: 10.1093/restud/rdaf087. URL https://doi.org/10.1093/restud/rdaf087.

Aurélien Bibaut, Nathan Kallus, Simon Ejdemyr, and Michael Zhao. Long-term causal inference with imperfect surrogates using many weak experiments, proxies, and cross-fold moments. *arXiv preprint arXiv:2311.04657*, 2023.

Aurélien Bibaut, Nathan Kallus, and Apoorva Lal. Nonparametric jackknife instrumental variable estimation and confounding robust surrogate indices. *arXiv preprint arXiv:2406.14140*, 2024.

Sylwia Bujkiewicz, Oriana Ciani, Bart Heeg, Dawn Lee, Jeanette M. Kusel, Kristian Thorlund, Petros Pechlivanoglou, Stephen Stefani, Wanrudee Isaranuwatchai, Marc Buyse, and Mario Ouwens. Methods for evaluation of surrogate endpoints for health technology assessment decision making: A good practices report of an ISPOR task force. *Value in Health*, 29(5):711–724, 2026. doi: 10.1016/j.jval.2026.01.020.

Marc Buyse, Geert Molenberghs, Tomasz Burzykowski, Didier Renard, and Helena Geys. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 1 (1):49–67, 2000.

Hua Chen, Zhi Geng, and Jinzhu Jia. Criteria for surrogate end points. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(5):919–932, 2007.

Jiafeng Chen and David M. Ritzwoller. Semiparametric estimation of long-term treatment effects. *Journal of Econometrics*, 237(2, Part A):105545, 2023. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2023.105545.

Taane G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman. Survival analysis part I: basic concepts and first analyses. *British journal of cancer*, 89(2):232–238, 2003.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008. doi: 10.1214/009052607000000910.

Eric J. Daza. Causal analysis of self-tracked time series data using a counterfactual framework for N-of-1 trials. *Methods of Information in Medicine*, 57(S01):e10–e21, 2018. doi: 10.3414/ME16-02-0044.

Naihua Duan, Richard L Kravitz, and Christopher H Schmid. Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *Journal of clinical epidemiology*, 66(8):S21–S28, 2013.

Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

Laurence S Freedman, Barry I Graubard, and Arthur Schatzkin. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in medicine*, 11(2):167–178, 1992.

Chenyin Gao, Peter B. Gilbert, and Larry Han. On the role of surrogates in conformal inference of individual causal effects. *arXiv preprint arXiv:2412.12365*, 2024.

Peter B Gilbert and Michael G Hudgens. Evaluating candidate principal surrogate endpoints. *Biometrics*, 64(4):1146–1154, 2008.

Gordon H Guyatt, Jana L Keller, Roman Jaeschke, David Rosenbloom, Jonathan D Adachi, and Michael T Newhouse. The n-of-1 randomized controlled trial: clinical usefulness: our three-year experience. *Annals of internal medicine*, 112(4):293–299, 1990.

Frank E. Harrell, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1414–1423, 2017.

Guido Imbens, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. Long-term causal inference under persistent confounding via data combination. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):362–388, 2025. doi: 10.1093/jrsssb/qkae095.

Guido W. Imbens and Whitney K. Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009.

Marshall M Joffe and Tom Greene. Related causal frameworks for surrogate outcomes. *Biometrics*, 65(2):530–538, 2009.

Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):480–509, 2025.

Predrag Klasnja, Eric B Hekler, Saul Shiffman, Audrey Boruvka, Daniel Almirall, Ambuj Tewari, and Susan A Murphy. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health psychology*, 34(S):1220, 2015.

Michael R. Kosorok and Eric B. Laber. Precision medicine. *Annual Review of Statistics and Its Application*, 6:263–286, 2019.

Hmwe H. Kyu, Victoria F. Bachman, Lily T. Alexander, John Everett Mumford, Ashkan Afshin, Kara Estep, J. Lennert Veerman, Kristen Delwiche, Marissa L. Iannarone, Madeline L. Moyer, Kelly Cercy, Theo Vos, Christopher J. L. Murray, and Mohammad H. Forouzanfar. Physical activity and risk of breast cancer, colon cancer, diabetes, ischemic heart disease, and ischemic stroke events: systematic review and dose-response meta-analysis for the Global Burden of Disease Study 2013. *BMJ*, 354:i3857, 2016. doi: 10.1136/bmj.i3857.

Steffen L. Lauritzen, Odd O. Aalen, Donald B. Rubin, and Elja Arjas. Discussion on causality [with reply]. *Scandinavian Journal of Statistics*, 31(2):189–201, 2004. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4616823.

Duck-chul Lee, Xuemei Sui, Enrique G. Artero, I-Min Lee, Timothy S. Church, Paul A. McAuley, Fatima C. Stanford, Harold W. Kohl, and Steven N. Blair. Long-term effects of changes in cardiorespiratory fitness and body mass index on all-cause and cardiovascular disease mortality in men. *Circulation*, 124(23):2483–2490, 2011. doi: 10.1161/CIRCULATIONAHA.111.038422. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.111.038422.

Elizabeth O Lillie, Bradley Patay, Joel Diamant, Brian Issell, Eric J Topol, and Nicholas J Schork. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Personalized medicine*, 8(2):161–173, 2011.

Calum A MacRae. Apple Heart & Movement Study. ClinicalTrials.gov identifier NCT04198194, 2021.

Alessandro-Umberto Margueritte, Ahmet Zahid Balcıoğlu, Jesse H. Krijthe, Dave Zachariah, and Fredrik D. Johansson. Learning plug-in surrogate endpoints for randomized experiments. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 4272–4298, 2026.

Hiral Master, Jeffrey Annis, Shi Huang, Joshua A. Beckman, Francis Ratsimbazafy, Kayla Marginean, Robert Carroll, Karthik Natarajan, Frank E. Harrell, Dan M. Roden, Paul Harris, and Evan L. Brittain. Association of step counts over time with the risk of chronic disease in the All of Us research program. *Nature Medicine*, 28(11):2301–2308, 2022. doi: 10.1038/s41591-022-02012-w.

Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Whitney K Newey, James L Powell, and Francis Vella. Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67(3):565–603, 1999.

Amanda E. Paluch, Shivangi Bajpai, Marcel Ballin, David R. Bassett, Thomas W. Buford, Mercedes R. Carnethon, Ariel Chernofsky, Erin E. Dooley, Ulf Ekelund, Kelly R. Evenson, et al. Prospective association of daily steps with cardiovascular disease: A harmonized meta-analysis. *Circulation*, 147(2):122–131, 2023.

Ambarish Pandey, Sushil Garg, Monica Khunger, Douglas Darden, Colby Ayers, Dharam J. Kumbhani, Helen G. Mayo, James A. de Lemos, and Jarett D. Berry. Dose-response relationship between physical activity and risk of heart failure: A meta-analysis. *Circulation*, 132(19):1786–1794, 2015.

J. Pearl and E. Bareinboim. Transportability across studies: A formal approach. Technical Report R-372, University of California, Los Angeles, 2011.

Judea Pearl. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pp. 373–392. 2022.

Ross L. Prentice. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8(4):431–440, 1989. doi: https://doi.org/10.1002/sim.4780080407.

Brenda L. Price, Peter B. Gilbert, and Mark J. van der Laan. Estimation of the optimal surrogate based on a randomized trial. *Biometrics*, 74(4):1271–1281, 2018.

James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.

Nicholas J. Schork. Personalized medicine: Time for one-person trials. *Nature*, 520(7549):609–611, 2015.

Florian Stijven and Peter B. Gilbert. Evaluation of surrogate endpoints based on meta-analysis with surrogate indices. *arXiv preprint arXiv:2509.01737*, 2025.

Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390, 2024.

James Truslow, Angela Spillane, Huiming Lin, Katherine Cyr, Adeeti Ullal, Edith Arnold, Ron Huang, Laura Rhodes, Jennifer Block, Jamie Stark, et al. Understanding activity and physiology at scale: the Apple Heart & Movement Study. *npj Digital Medicine*, 7(1):242, 2024.

Lars van der Laan, Nathan Kallus, and Aurélien Bibaut. Nonparametric instrumental variable inference with many weak instruments. *arXiv preprint arXiv:2505.07729*, 2025.

Tyler J VanderWeele. Surrogate measures and consistent surrogates. *Biometrics*, 69(3):561–565, 2013.

Zeyu Xu, Xiaojie Mao, Hao Mei, and Yue Liu. Evaluating surrogates in individual treatment regimes. *arXiv preprint arXiv:2512.00405*, 2025.

Deborah R Zucker, Christopher H Schmid, Martin W McIntosh, Ralph B D'Agostino, Harry P Selker, and Joseph Lau. Combining single patient (n-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. *Journal of clinical epidemiology*, 50(4):401–410, 1997.

A Appendix

Contents

A.1 Related work . . . 15
A.2 Choice of the definition of individual-level regret . . . 17
A.3 Population vs. individual-level surrogacy . . . 19
A.4 Empirical estimation of individual-level regret . . . 20
A.4.1 Empirical evaluation pipeline . . . 20
A.4.2 Consistency of the estimator . . . 20
A.5 Generalization to new treatments . . . 22
A.6 Extension to within-subject varying covariates . . . 23
A.6.1 Generalized setting . . . 23
A.6.2 Regret identifiability under additional assumptions . . . 24
A.6.3 Simplified regret identifiability in the linear case . . . 25
A.6.4 Application to the case study in Section 5 and the Cox model . . . 26
A.7 Learning surrogate maps . . . 28
A.7.1 Learning with N-of-1 data: empirical regret minimization . . . 28
A.7.2 Learning with only observational data . . . 28
A.7.3 Optimality of the causal map under additive confounding . . . 29
A.7.4 Beyond additive confounding . . . 30
A.8 Additional details on the synthetic experiments . . . 32
A.8.1 Data-generating process . . . 32
A.8.2 Evaluation setup . . . 32
A.8.3 Additional finite-sample experiments . . . 34
A.8.4 Learning surrogate maps . . . 34
A.8.5 Ablation: violation of invertibility . . . 34
A.9 Additional details on the AHMS case study in Section 5 . . . 36
A.9.1 Study and cohort . . . 36
A.9.2 Change windows, treatments, surrogates and outcome . . . 36
A.9.3 Evaluation protocol . . . 38
A.9.4 Surrogate maps . . . 39
A.9.5 Additional results . . . 40
A.10 Deferred proofs . . . 41
A.10.1 Proof of Theorem 1 . . . 41
A.10.2 Proof of Theorem 2 . . . 41

A.1 Related work

Our work draws on and contributes to several lines of research: the statistical and causal evaluation of surrogate endpoints, the more recent line of learning of surrogate maps from multi-dimensional intermediate measurements, the N-of-1 experimental design, and surrogates for individualized decisions.

Population-level surrogacy criteria. The first formal criterion is due to Prentice (1989), and requires the primary outcome to be conditionally independent of treatment given the surrogate in a population. In particular, the Prentice criterion induces a statistical test of surrogacy. Freedman et al. (1992) quantify the degree of surrogacy through the proportion of the treatment effect explained by the surrogate. Both criteria refer to the average effect in a single trial and are known to be prone to the so-called *surrogate paradox* (Chen et al., 2007; VanderWeele, 2013). The meta-analytic framework (Buyse et al., 2000) instead aims to measure correlation between trial-level average treatment effects on the surrogate endpoint and trial-level effects on the outcome across multiple randomized trials. This correlation is summarized by the coefficient of determination R^2_{trial}, and is often regarded as the gold standard for surrogate endpoint evaluation (Stijven & Gilbert, 2025; Bujkiewicz et al., 2026). Note that the meta-analytic framework also defines an individual-level, or within-trial correlation coefficient R^2_{indiv}, which measures the within-trial association between surrogate and outcome across subjects. Despite the shared name, this measure is unrelated to our notion of individual-level surrogacy, which aims to rank treatments within a single individual on which repeated experimentation is performed.

There exist multiple causal surrogacy frameworks, often formulated as conditions on the underlying causal graph of (A, X, S, Y) or/and the corresponding potential outcomes: principal surrogacy (Frangakis & Rubin, 2002; Gilbert & Hudgens, 2008), strong surrogacy (Lauritzen et al., 2004), and consistent surrogacy (Chen et al., 2007; VanderWeele, 2013). Although principal surrogacy, similarly to our work, operates on tuples $(S(a), S(a'))$ of surrogate potential outcomes, it mostly relies on their imputation to evaluate (population-level) surrogacy, and requires ignorability. We refer the reader to Joffe & Greene (2009) for a comparison between these frameworks and their connection to mediation analysis (Robins & Greenland, 1992; Pearl, 2022). Pearl & Bareinboim (2011) provide a definition of surrogacy based on transportability between trials. All of the listed criteria are defined with respect to treatment effects in a population. As we illustrate in Appendix A.3, surrogacy at the population level does not imply surrogacy at the individual level: a surrogate map can rank treatments correctly on a population level and misrank them on an individual level within a large sub-population.

Surrogates for individualized decisions. A few recent works attempt to characterize surrogates from an individual-level perspective. Xu et al. (2025) propose to evaluate surrogates by the regret of the individualized treatment rules (ITRs) learned from them, i.e., the loss in value relative to outcome-optimal ITRs, also under budget constraints, and propose doubly robust estimators. Despite a similar terminology, both the problem and the notion of regret in Xu et al. (2025) differ from ours. In Xu et al. (2025), an ITR is understood as an assignment of a binary treatment from pre-treatment covariates. The surrogate is then used to estimate the conditional average treatment effect (CATE) $\mathbb{E}[Y(1) - Y(0) \mid X = x]$ via the surrogate CATE $\mathbb{E}[S(1) - S(0) \mid X = x]$ and to assign the treatment whenever the estimated surrogate CATE is positive. The regret is the loss in population value relative to the ITR based on the outcome CATE. In particular, this line of work assumes that treatment effect heterogeneity can be explained by the observed covariates. In our N-of-1 setting, by contrast, the decision is made after an individual's own surrogate responses to several, possibly previously unseen, treatments have been observed. The regret of a surrogate map is measured against the individual's counterfactual outcome means $\mu^j(a)$ within its own structural causal model, which depend on a latent type beyond the covariates. Gao et al. (2024) use surrogates to construct more efficient conformal prediction intervals for individual treatment effects. In total, these works, similarly to the broad literature on ITRs (Kosorok & Laber, 2019), consider rules that assign known treatments based on pre-treatment covariates and are evaluated on a population level. As an example, such an ITR could aim to assign blood pressure medication $A = 1$ or no treatment $A = 0$ based on the subject's age, sex and other conditions to maximize total benefit in a population. In contrast, our goal would be to recommend to an individual who is sequentially trying multiple,

potentially novel, medications $A = 1, ..., m$ the best of the treatments based on their short-term biomarker values.

Learning surrogate maps from multiple intermediate measurements. The surrogate index (Athey et al., 2026) predicts the long-term outcome from a vector of intermediate measurements and identifies average treatment effects under Prentice surrogacy conditions and transportability between experimental and observational samples. Kallus & Mao (2025) characterize the efficiency gains from surrogates when the outcome is partially observed. Chen & Ritzwoller (2023) develop semiparametric estimators of long-term treatment effects. Related approaches estimate an optimal surrogate from a randomized trial (Price et al., 2018), learn plug-in surrogate maps by minimizing the error of the induced CATE (Margueritte et al., 2026), or combine the surrogate index with the meta-analytic framework (Stijven & Gilbert, 2025). When surrogates and outcome are confounded, statistical surrogacy fails even if the surrogates fully mediate the treatment effect. Imbens et al. (2025) address such confounding by exploiting the sequential structure of short-term outcomes. Bibaut et al. (2023; 2024); van der Laan et al. (2025) use historical experiments as instruments to learn a bridge function h with $\mathbb{E}[Y - h(S) \mid A] = 0$ as a confounding-robust surrogate index for the average effect of new treatments. In contrast to this line of work, we target within-individual treatment rankings, and not estimation of average or conditional treatment effects. Our proposed evaluation criterion (3) is invariant to monotone transformations of the surrogate map, does not require accurate estimation of effects, and thus is a complementary metric of surrogacy.

N-of-1 trial design. The N-of-1 trial design (Guyatt et al., 1990; Lillie et al., 2011; Schork, 2015) alternates treatments within a single individual and is the gold standard for individual-level treatment selection. Statistical methods for analyzing N-of-1 trials include mixed-effect models (Zucker et al., 1997; Duan et al., 2013; Araujo et al., 2016). Recent work adapts the N-of-1 design to digital health interventions (Klasnja et al., 2015). Daza (2018) formalizes N-of-1 observational studies of self-tracked time series. Most N-of-1 analyses assume that the outcome of interest is measured repeatedly within the individual. In contrast, our framework is most interesting in the setting in which the outcome can be observed at most once (e.g., life expectancy) or is very costly to measure (e.g., full body MRI).

A.2 Choice of the definition of individual-level regret

Pairwise definition of regret. Instead of computing individual-level regret (3) over all pairs of treatments in $\mathcal{A}^j$, we could have just computed it over full treatment set $\mathcal{A}^j$, which recovers the classical notion of regret as the cost of not selecting an individual-level best treatment. Here, we show that the pairwise definition is stricter, and thus also controls the classical individual-level regret. However, the converse fails as soon as $|\mathcal{A}^j| \geq 3$.

Lemma 1. *Let f be a valid surrogate map and $j \in \mathcal{J}$. The following are equivalent:*

(i) $\succeq_f^j \subseteq \succeq^j$*, i.e. f recovers the treatment ranking of subject j;*

(ii) $r^j(f) = 0$*;*

(iii) $\pi_f(j; \mathcal{B}) \subseteq \operatorname{argmax}_{a \in \mathcal{B}} \mu^j(a)$ *for any treatment subset $\mathcal{B} \subseteq \mathcal{A}^j$.*

Finally, $r^j\big(f; \mathcal{A}^j\big) \leq \binom{|\mathcal{A}^j|}{2} r^j(f)$.

Proof. (i) $\Rightarrow$ (iii) follows by definition of the ordering. (iii) $\Rightarrow$ (ii) follows by application to all pairs of treatments. (ii) $\Rightarrow$ (i): if $a \succeq_f^j a'$ then $a \in \pi_f(j; \{a, a'\})$, and $r^j(f; \{a, a'\}) = 0$ implies $\mu^j(a) = \max\{\mu^j(a), \mu^j(a')\}$, and thus $a \succeq^j a'$.

For the bound, define the worst surrogate-selected treatment $a^\dagger \in \operatorname{argmin}_{a \in \pi_f(j; \mathcal{A}^j)} \mu^j(a)$ and the ground-truth best treatment $a^\star \in \operatorname{argmax}_{a \in \mathcal{A}^j} \mu^j(a)$. Observe that $r^j(f; \{a^\star, a^\dagger\}) \geq \mu^j(a^\star) - \mu^j(a^\dagger) = r^j(f; \mathcal{A}^j)$. The claim follows. □

We conclude that if the pairwise regret of a surrogate map is zero, its "classical" regret will be zero as well.

We choose the pairwise version of regret for multiple reasons: first, the classical regret $r^j(f; \mathcal{A}^j)$ is a function of the whole set of treatments: in particular, if the set contains one clearly best treatment and multiple inferior, but similar other treatments, the regret of many maps will be zero. However, if we then expose the subject to only the inferior treatments and ask the surrogate map to rank them, the result could be completely wrong even though the regret of the map on the same subject and full $\mathcal{A}^j$ is zero. As an example (cf. Section 5), a subject could try marathon running and two mild walking routines. It might be clear via many maps that marathon running is better for the subject's endurance, however, ultimately, the subject might want to stick to walking for other reasons, which necessitates correctly ranking the walking routines as well.

Second, as we show in Section 3.2, the pairwise definition of regret allows us to generalize evaluation results to new N-of-1 trials containing previously unseen treatments under the *mild diversity condition* on the distribution of surrogate response pairs $(S^j(a), S^j(a'))$: if instead we used the classical definition, we would have to impose the diversity condition on the distribution of arbitrarily large tuples of surrogate responses, which could be much harder to satisfy in a small evaluation-time population of N-of-1 subjects.

Choice of outcome labels. We comment on the choice of labels $\mu^j(a)$ in (3). One could instead use other seemingly more informative labels $L^j(a)$, such as the realized outcome $Y^j(a)$ or $\mathbb{E}^j[Y^j(a) \mid S^j(a), X^j, U^j]$, and define the "more refined" regret $r_L^j(f; \mathcal{B}) := \mathbb{E}^j\left[\max_{a \in \mathcal{B}} L^j(a)\right] - \min_{a' \in \pi_f(j; \mathcal{B})} \mathbb{E}^j\left[L^j(a')\right]$. However, for any such more refined label with $\mathbb{E}^j[L^j(a)] = \mu^j(a)$, the tower property gives

$$r_L^j(f; \mathcal{B}) = r^j(f; \mathcal{B}) + c^j(\mathcal{B}), \qquad c^j(\mathcal{B}) := \mathbb{E}^j\left[\max_{a \in \mathcal{B}} L^j(a)\right] - \max_{a \in \mathcal{B}} \mu^j(a) \;\geq\; 0,$$

where the offset $c^j(\mathcal{B})$ does not depend on the surrogate map f. In particular, although the regrets are not equal, they *rank all surrogate maps equally*. Thus, importantly, a minimizer of (2) also minimizes the regret based on oracle realized subject-level outcomes $Y^j(a)$.

Treatment ranking error. Besides the regret, we could also consider the *treatment ranking error* of a map f, i.e. the total fraction of misranked treatment pairs. For a surrogate map f and a subject $j \in \mathcal{J}$, we define the treatment ranking error and its population average as

$$\ell^j(f) \coloneqq \binom{|\mathcal{A}^j|}{2}^{-1} \sum_{\{a,a'\} \subseteq \mathcal{A}^j, a \neq a'} \mathbb{1}\{r^j(f;\{a,a'\}) > 0\} \in [0,1], \qquad L_{\mathcal{J}}(f) \coloneqq \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \ell^j(f), \tag{8}$$

and analogously, the pseudo-ranking error $\tilde{\ell}^j(f)$. The ranking error is thus the 0-1 counterpart of the regret, and by Lemma 1, $\ell^j(f) = 0$ if and only if $r^j(f) = 0$. Unlike the regret, $\ell^j(f)$ depends on the outcome means only through their ranking, and is therefore invariant under strictly increasing transformations of μ^j, such as a change of the outcome scale. A random ranking of every pair has expected ranking error $1/2$.

A.3 POPULATION VS. INDIVIDUAL-LEVEL SURROGACY

We illustrate with a simple example that population-level surrogacy neither implies nor is implied by individual-level surrogacy. As population-level criterion, we take the meta-analytic framework (Buyse et al., 2000), the recommended standard for surrogate endpoint validation (Bujkiewicz et al., 2026), which has also been applied to the surrogate index evaluation (Stijven & Gilbert, 2025). For a surrogate map f and a randomized controlled trial with treatments $A \in \{a, a'\}$, the trial-level effects on the surrogate map f and on the outcome are

$$\alpha_f(a, a') := \mathbb{E}\big[f(S(a)) - f(S(a'))\big], \qquad \beta(a, a') := \mathbb{E}\big[Y(a) - Y(a')\big].$$

Across a collection of trials, the meta-analytic framework measures how well β is correlated with α_f through the coefficient of determination $R^2_{\text{trial}}(f)$. In the following example, we show how perfect meta-analytic evaluation does not imply and is not implied by individual-level surrogacy in the sense of Section 2.

Example. Consider two latent types $U \in \{-1, +1\}$ with $P(U = 1) = \frac{1}{2}$, and treatments $a = (a_1, a_2)$ with intensity $a_1 \in [-\frac{1}{2}, \frac{1}{2}]$ and a class $a_2 \in \{-1, +1\}$. Assume the following structural equations for the surrogate and outcome:

$$S_1 := A_1, \qquad S_2 := A_2\, U, \qquad Y := \theta_1 S_1 + \theta_2 S_2 + \gamma\, U + \epsilon_Y, \qquad \theta_1, \theta_2 > 0.$$

Biomarker S_1 has a homogeneous response to A_1 across all subjects, whereas the response of marker S_2 to the treatment class A_2 depends on the type: it improves for type $U = +1$ under class $A_2 = +1$ and for type $U = -1$ under class $A_2 = -1$. In every population trial, and every N-of-1 trial, we compare $a = (a_1, +1)$ and $a' = (a_1', -1)$, with a_1, a_1' i.i.d. Unif$[-\frac{1}{2}, \frac{1}{2}]$.

Writing $\Delta := a_1 - a_1'$, the individual-level and trial-level effects on the outcome are

$$\mu^j(a) - \mu^j(a') = \theta_1 \Delta + 2\theta_2 U^j, \qquad \beta(a, a') = \theta_1 \Delta.$$

If $\theta_1 < 2\theta_2$, confounding-induced heterogeneity dominates the individual treatment effect $\mu^j(a) - \mu^j(a')$, whereas the average treatment effect is driven by Δ only. Analogously, any linear surrogate map $f(s) = w_1 s_1 + w_2 s_2$ induces a within-subject contrast $w_1\Delta + 2w_2 U^j$ and computes the trial-level effect $\alpha_f = w_1\Delta$. Thus, a meta-analytic evaluation of f can only evaluate its first coefficient w_1, and ignores w_2. In contrast, the goodness of the f-induced individual-level treatment ranking is determined by both w_1 and w_2.

Consider the surrogate index estimator learned on a separate trial in which $A_1 \sim$ Unif$[-\frac{1}{2}, \frac{1}{2}]$ and $A_2 = +1$ with probability $q > \frac{1}{2}$, independently of U, i.e., treatment class $+1$ is assigned more often. We compute the surrogate index to be

$$f_{\text{SI}}(s) = \theta_1 s_1 + w_{\text{SI}}\, s_2, \qquad w_{\text{SI}} := \theta_2 + (2q - 1)\gamma.$$

Population-level surrogacy does not imply individual-level surrogacy. For the surrogate index, we have $\alpha_{f_{\text{SI}}} = \beta$ on every trial and any q and γ. Thus, the surrogate index is a perfect trial-level surrogate that exactly predicts the average treatment effect. However, if $w_{\text{SI}} < -\theta_1/2$, it selects the worst treatment for every N-of-1 subject and incurs the largest possible regret.

Individual-level surrogacy does not imply population-level surrogacy. Conversely, consider the map $f_{\text{het}}(s) := s_2$ which produces a within-subject contrast $2U^j$ and ranks the two treatments correctly for every subject. However, its trial-level effect vanishes in every trial, $\alpha_{f_{\text{het}}} = 0$, so that $R^2_{\text{trial}}(f_{\text{het}}) = 0$. Thus, a meta-analysis would discard f_{het}, although it achieves the lowest possible individual-level regret. Another example is the negative surrogate index $-f_{\text{SI}}$, which achieves zero regret (since f_{SI} achieves maximal regret), but $\alpha_{-f_{\text{SI}}} = -\beta$, i.e., it exhibits the surrogate paradox (Chen et al., 2007; VanderWeele, 2013) in every trial.

A.4 Empirical estimation of individual-level regret

A.4.1 Empirical evaluation pipeline

We now describe the concrete empirical procedure Algorithm 1 to evaluate surrogate maps via individual-level pseudo-regret based on our identifiability result Theorem 1. We consider access to a sample $D^{\text{obs}} = \{(A_i, X_i, S_i, Y_i)\}_{i=1}^{n_{\text{obs}}}$ from the observational study, drawn i.i.d. from $\mathbb{P}^{\text{obs}}$. This dataset contains one treatment and one long-term outcome per subject. Further, we consider access to a collection of evaluation N-of-1 subjects $\mathcal{J}^{\text{eval}}$, where for each N-of-1 subject, we observe a sequence of exposures followed by intermediate measurements only,

$$D^j = \{(A_i^j, X_i^j, S_i^j)\}_{i=1}^{n_j}, \quad \mathcal{A}^j = \{A_1^j, ..., A_{n_j}^j\},$$

where $S_i^j = S^j(A_i^j)$ by consistency. We assume $|\mathcal{A}^j| \geq 2$. For simplicity, we assume that all treatments in $\mathcal{A}^j$ are distinct.

The empirical procedure for estimating individual-level regret and treatment ranking error (8) of a surrogate map f on a population of N-of-1 subjects $\mathcal{J}^{\text{eval}}$ is described in Algorithm 1. In Appendix A.4.2, we show that empirical individual-level pseudo-regret is a consistent estimator of the true regret (combined with Theorem 1) and provide the corresponding error bound which holds uniformly over all surrogate maps f. Note that this assumes that the surrogate map f is learned on a separate sample independent from the evaluation sample, as we do in Section 4 and Section 5. We empirically validate the finite-sample behavior of our evaluation pipeline for different conditional mean estimators in Section 4.

Algorithm 1: Empirical individual-level pseudo-regret via outcome imputation

Input: Sample D^{obs} from the observational study; collection $\{D^j\}_{j \in \mathcal{J}^{\text{eval}}}$ of N-of-1 samples; surrogate map f

Output: Estimated average individual-level pseudo-regret $\widehat{R}(f)$ and ranking error $\widehat{\mathcal{L}}(f)$

Fit the outcome model $\widehat{m}(a, x, s)$ of $\mathbb{E}^{\text{obs}}[\, Y \mid A = a, X = x, S = s \,]$ on D^{obs}

for $j \in \mathcal{J}^{\text{eval}}$ **do**

  for $A_i^j \in \mathcal{A}^j$ **do**

    impute the pseudo-outcome $\widehat{\mu}^j(A_i^j) \leftarrow \widehat{m}\big(A_i^j, X_i^j, S_i^j\big)$

    compute the surrogate value $\widehat{v}_f^j(A_i^j) \leftarrow f\big(S_i^j, X_i^j\big)$

  for *each pair* $\{a, a'\} \subseteq \mathcal{A}^j$ **do**

    $\widehat{\pi}^j \leftarrow \operatorname{argmax}_{b \in \{a,a'\}} \widehat{v}_f^j(b)$

    $\widehat{r}_{aa'}^j \leftarrow \max_{b \in \{a,a'\}} \widehat{\mu}^j(b) \; - \; \min_{b \in \widehat{\pi}^j} \widehat{\mu}^j(b)$

  $\widehat{r}^j(f) \leftarrow \binom{|\mathcal{A}^j|}{2}^{-1} \sum_{\{a,a'\}} \widehat{r}_{aa'}^j, \quad \widehat{\ell}^j(f) \leftarrow \binom{|\mathcal{A}^j|}{2}^{-1} \sum_{\{a,a'\}} \mathbb{1}\{\widehat{r}_{aa'}^j > 0\}$

return $\widehat{R}(f) \leftarrow \dfrac{1}{|\mathcal{J}^{\text{eval}}|} \displaystyle\sum_{j \in \mathcal{J}^{\text{eval}}} \widehat{r}^j(f), \quad \widehat{\mathcal{L}}(f) \leftarrow \dfrac{1}{|\mathcal{J}^{\text{eval}}|} \displaystyle\sum_{j \in \mathcal{J}^{\text{eval}}} \widehat{\ell}^j(f)$

A.4.2 Consistency of the estimator

We show that the empirical pseudo-regret computed in Algorithm 1 consistently estimates the individual-level regret, and bound its error by the error of the outcome model. We fix an N-of-1 subject j with constant covariates X^j (as in Section 3) for which Assumptions 1 and 2 hold. We write $m^{\text{obs}}(a, x, s) := \mathbb{E}^{\text{obs}}[Y \mid A = a, X = x, S = s]$ for the population-level outcome regression function, and $\widehat{m}$ for its empirical estimate learned on the sample D^{obs} from the observational study. For every treatment $a \in \mathcal{A}^j$, we denote by

$$\widehat{\mu}^j(a) := \widehat{m}\big(a, X^j, S^j(a)\big), \qquad e^j(a) := \widehat{\mu}^j(a) - \tilde{\mu}^j(a)$$

the empirical imputed pseudo-outcomes and their errors, where $\tilde{\mu}^j(a) = m^{\text{obs}}(a, X^j, S^j(a))$ is the population-level pseudo-outcome (5). Since $S^j(a)$ is constant by Assumption 2, the surrogate induced score $\mathbb{E}^j[f(S^j(a), X^j)] = f(S^j(a), X^j)$ does not depend on $\hat{m}$, and thus, the estimation error of individual-level pseudo-regret can be attributed to imputed outcome labels only.

Proposition 1 (Plug-in error bound)**.** *Denote by $\mathcal{P}^j$ the set of distinct pairs in $\mathcal{A}^j$. For every outcome model $\widehat{m}$ and every surrogate map f,*

$$\left|\widehat{r}^{\,j}(f) - r^j(f)\right| \leq \frac{1}{|\mathcal{P}^j|} \sum_{\{a,a'\}\in\mathcal{P}^j} \left|e^j(a) - e^j(a')\right| \leq 2 \max_{a\in\mathcal{A}^j} \left|e^j(a)\right|. \tag{9}$$

Proof. Fix a pair of treatments $\{a, a'\} \in \mathcal{P}^j$ and consider the surrogate-induced policy $\pi := \pi_f(j; \{a, a'\})$. The regret of the pair equals either $[\tilde{\mu}^j(a) - \tilde{\mu}^j(a')]_+$, $[-\tilde{\mu}^j(a) + \tilde{\mu}^j(a')]_+$ or $|-\tilde{\mu}^j(a) + \tilde{\mu}^j(a')|$ depending on the value of π, and can be thus written as a function $\phi_f(\Delta_{a,a'})$ with $\Delta_{a,a'} = \tilde{\mu}^j(a) - \tilde{\mu}^j(a')$. Similarly, the empirical pseudo-regret can be written as $\phi_f(\hat{\Delta}_{a,a'})$, where $\hat{\Delta}_{a,a'} = \hat{\mu}^j(a) - \hat{\mu}^j(a')$. As we see, ϕ_f is 1-Lipschitz, and thus

$$|\hat{r}^j(f; \{a, a'\}) - \tilde{r}^j(f; \{a, a'\})| = |\phi_f(\hat{\Delta}_{a,a'}) - \phi_f(\Delta_{a,a'})| =$$
$$|\phi_f(\Delta_{a,a'} + e^j(a) - e^j(a')) - \phi_f(\Delta_{a,a'})| \leq |e^j(a) - e^j(a')|.$$

Via Theorem 1, we have $\tilde{r}^j = r^j$, and the claim follows. □

Corollary 1 (Consistency)**.** *Denote by $\widehat{m}_n$ the regression function fitted on D^{obs} with $|D^{\text{obs}}| = n$. If $\widehat{m}_n(a, X^j, S^j(a)) \xrightarrow{p} m^{\text{obs}}(a, X^j, S^j(a))$ for all $a \in \mathcal{A}^j$, then $\sup_f |\widehat{r}^{\,j}(f) - r^j(f)| \xrightarrow{p} 0$.*

The corollary follows directly from Proposition 1 by continuous mapping theorem.

We note that both the error bound and the consistency result hold uniformly over surrogate maps. Further, empirical regret inherits the convergence rate of the corresponding regression function, as illustrated in Figure 2a.

A.5 Generalization to new treatments

No generalization to arbitrary treatments. We note that no evaluation procedure and no surrogate map $f(S, X)$ can generalize to an *arbitrary* new treatment: it suffices to consider a pair $a, a' \in \mathcal{A}^{\text{tst}}$ of treatments which completely bypass the surrogate vector, but strongly act on Y through the *direct* causal mechanism f_0. In this case, the surrogate responses are equal $g_0(a, \cdot, \cdot) = g_0(a', \cdot, \cdot)$, but the outcome treatment effect $\mu^j(a) - \mu^j(a')$ can be made arbitrarily large, and so can the individual-level regret of any surrogate map.

For an N-of-1 population $\mathcal{J}$, let $\mathbb{Q}$ be the surrogate contrast law on $\mathcal{X} \times \mathcal{U} \times \mathcal{S} \times \mathcal{S}$ obtained by sampling a subject $j \in \mathcal{J}$ uniformly and then a pair of its surrogate responses $(S^j(a), S^j(a'))$ uniformly without replacement. We write $\mathbb{Q}^{\text{eval}}$ and $\mathbb{Q}^{\text{tst}}$ for the contrast laws of the populations $\mathcal{J}^{\text{eval}}$ and $\mathcal{J}^{\text{tst}}$, and abbreviate the population-level regrets $R^{\text{tst}}(f) := \mathbb{E}_{j \sim \text{Unif}(\mathcal{J}^{\text{tst}})}[r^j(f)]$ and $\tilde{R}^{\text{eval}}(f) := \mathbb{E}_{j \sim \text{Unif}(\mathcal{J}^{\text{eval}})}[\tilde{r}^j(f)]$. We state the following generalization result:

Theorem 2 (Generalization to new treatments)**.** *Let f be a valid surrogate map. Assume that Assumptions 1 and 2 hold for every evaluation subject $j \in \mathcal{J}^{\text{eval}}$, Assumption 3 holds, U^j is $\mathbb{P}^j$-a.s. constant for every $j \in \mathcal{J}^{\text{tst}}$, and $\mathbb{Q}^{\text{tst}} \ll \mathbb{Q}^{\text{eval}}$. Then perfect evaluation-time pseudo-regret implies perfect test-time regret, i.e., $\tilde{R}^{\text{eval}}(f) = 0 \implies R^{\text{tst}}(f) = 0$. If moreover the density $q := d\mathbb{Q}^{\text{tst}}/d\mathbb{Q}^{\text{eval}}$ is bounded by some $0 \leq B < \infty$, then*

$$R^{\text{tst}}(f) \;\leq\; B \cdot \tilde{R}^{\text{eval}}(f). \tag{10}$$

Note that our generalization condition is *treatment-independent* and reduces to a diversity condition on the covariates and pairs of surrogate responses *only*. In particular, the causal mechanism $g_0(a, \cdot)$ of a new treatment $a \in \mathcal{A}^{\text{tst}}$ is not constrained.

We note that the generalization result is partially made possible by the pairwise definition of the regret (3) (cf. Appendix A.2). If instead we defined our objective plainly as "identifying the best treatment for each subject", generalization to new treatments would require substantially stricter conditions on the diversity of the population of N-of-1 subjects and their sets of treatments $\mathcal{A}^j$, which become exponentially harder to satisfy as $|\mathcal{A}^j|$ grows.

A.6 Extension to within-subject varying covariates

In Section 2 and Section 3, for simplicity of presentation, we have assumed that the covariates X^j of an N-of-1 subject are constant. Examples of such covariates may include biological sex, baseline health status, socioeconomic status, education, approximate age bracket etc. In practice, some of the observed covariates can change between exposures within the same subject. Such covariates could be, e.g., calendar time, season, current age, current health, length of current employment, current income and more.

In this section, we explore generalization of our surrogate map evaluation to a setting which includes such within-subject varying covariates. In particular, this generalization allows us to relax our N-of-1 setting to a *within-subject observational study*. In practice, this means that we can observe subjects over time (as they perform self-assigned treatments, influenced by their changing environment) and treat them as N-of-1 study subjects in our framework.

We write $X = (X_{\text{fix}}, X_{\text{rand}})$, where X_{fix} (fixed covariates) are constant within a subject (e.g., baseline age, sex, baseline health status) and the *context* X_{rand} (random covariates) may vary within the subject-specific distribution. We generally allow X_{rand} to affect the treatment, the surrogates and the long-term outcome, resulting in edges $X_{\text{rand}} \to A$, $X_{\text{rand}} \to S$ and $X_{\text{rand}} \to Y$. The fixed covariates, as before, are allowed to affect A, S, Y as well.

A.6.1 Generalized setting

The structural equations $\mathcal{C}^{\text{obs}}$ of the population observational study read

$$X_{\text{fix}} \sim \mathbb{P}_{X_{\text{fix}}}, \qquad U \sim \mathbb{P}_{U|X_{\text{fix}}}, \qquad X_{\text{rand}} \sim \mathbb{P}^{\text{obs}}_{X_{\text{rand}}|X_{\text{fix}},U},$$
$$A := \pi_0(X_{\text{fix}}, X_{\text{rand}}, \epsilon_A), \qquad S := g_0(A, X_{\text{fix}}, X_{\text{rand}}, U), \qquad Y := f_0(A, S, X_{\text{fix}}, X_{\text{rand}}, U, \epsilon_Y),$$

with ϵ_A and ϵ_Y exogenous noise variables independent of each other and of $(X_{\text{fix}}, U, X_{\text{rand}})$.

For any N-of-1 subject j, the fixed covariates X^j_{fix} and the latent type U^j are constant. The structural equations $\mathcal{C}^j$ then read:

$$X^j_{\text{rand}} \sim \mathbb{P}^j_{X_{\text{rand}}}, \qquad A^j := \pi_0(X^j_{\text{fix}}, X^j_{\text{rand}}, \epsilon_A), \qquad S^j := g_0(A^j, X^j_{\text{fix}}, X^j_{\text{rand}}, U^j). \tag{11}$$

Our generalized subject-specific SCM (11) allows for context covariates X^j_{rand} such as current age, season or current health status. The latent type U^j of a subject may depend on the fixed covariates, but is, in total, fixed. We note, however, that (11) still allows a latent *state* that varies with the context, $\tilde{U} := \psi(X_{\text{fix}}, X_{\text{rand}}, U)$, which becomes apparent when one substitutes it into g_0 and f_0.

For any treatment $a \in \mathcal{A}^j$ and a context X_{rand}, we can define context-specific potential outcomes and the context-specific outcome label

$$S^j(a; X_{\text{rand}}) := g_0(a, X^j_{\text{fix}}, X_{\text{rand}}, U^j), \qquad Y^j(a; X_{\text{rand}}) := f_0\big(a, S^j(a; X_{\text{rand}}), X^j_{\text{fix}}, X_{\text{rand}}, U^j, \epsilon_Y\big),$$
$$\mu^j(a; X_{\text{rand}}) := \mathbb{E}\left[Y^j(a; X_{\text{rand}})\right].$$

In particular, the outcome labels $\mu^j(a)$ in Section 3 can be written as averages over all possible contexts:

$$\mu^j(a) = \mathbb{E}^j\left[Y^j(a)\right] = \int \mu^j(a; X_{\text{rand}})\, d\mathbb{P}^j_{X_{\text{rand}}}. \tag{12}$$

The main difficulty of the generalization is that we are still interested in correctly ranking the expected outcomes $\mu^j(a)$, and not the context-dependent outcomes $\mu^j(a; X_{\text{rand}})$. To see this, imagine an N-of-1 subject who tries two exercise routines a and a', one at a younger age and lower weight, and another at an older age and higher weight. If we, instead, proceed to rank $\mu^j(a; X_{\text{rand}})$, we might rate an inferior treatment in a *favorable* context higher than a preferred treatment in a disadvantageous context. If we recommend treatment a to a subject, they might repeat it in a completely different context, losing its perceived advantage. Thus, our objective is to perform *context-independent* within-subject ranking of treatments, which presents a technical difficulty, since we normally only observe one context per treatment.

A.6.2 Regret identifiability under additional assumptions

We impose Assumptions 1 and 2 with $X^j = (X^j_{\text{fix}}, X^j_{\text{rand}})$. We impose the following additional assumptions:

Assumption 4 (Shared conditional distribution for the context covariates)**.** *Under* $\mathbb{P}^{\text{obs}}$, $X_{\text{rand}} \perp\!\!\!\perp U \mid X_{\text{fix}}$. *Further, write* $\kappa(\cdot \mid X_{\text{fix}}) := \mathbb{P}^{\text{obs}}_{X_{\text{rand}} \mid X_{\text{fix}} = x_{\text{fix}}}$ *for the conditional distribution of the context covariates given the fixed covariates in the population study. For every N-of-1 subject* j, $\mathbb{P}^j_{X_{\text{rand}}} = \kappa(\cdot \mid X^j_{\text{fix}})$.

Assumption 4 states that, given the subject-level fixed covariates, the context covariates reveal no additional information about the subject's latent type. As described in Appendix A.6.1, our setting still allows for a within-subject varying latent state which changes with X_{rand}. Further, Assumption 4 states that the distribution of context covariates given the fixed covariates is shared across all subjects. Examples of covariates satisfying this assumption include weather (X_{rand}) given location (X_{fix}), current age (X_{rand}) given baseline age (X_{fix}), or time/season (X_{rand}) given date of enrollment in the study (X_{fix}).

Assumption 5 (Positivity)**.** *For every N-of-1 subject* j *and every treatment* $a \in \mathcal{A}^j$, $\operatorname{supp} \kappa(\cdot \mid X^j_{\text{fix}}) \subseteq \operatorname{supp} \mathbb{P}^{\text{obs}}_{X_{\text{rand}} \mid A = a, X_{\text{fix}} = X^j_{\text{fix}}}$.

Assumption 5 is a standard positivity assumption that requires that, in the population study, every treatment has a non-zero-probability of being performed in every context.

Assumption 6 (Additive shift by context covariates)**.** *There are measurable functions* g° *and* h *such that*

$$g_0(a, X_{\text{fix}}, X_{\text{rand}}, u) = g^\circ(a, X_{\text{fix}}, u) + h(X_{\text{fix}}, X_{\text{rand}}).$$

Thus, we require that the effect of the treatment on the surrogate vector is separable from the drift introduced by the context covariates X_{rand}. Assumption 6 allows us to compute the surrogate response in a different context using the surrogate response from the observed context via $S^j(a; X'_{\text{rand}}) = S^j(a; X_{\text{rand}}) - h(X^j_{\text{fix}}, X_{\text{rand}}) + h(X^j_{\text{fix}}, X'_{\text{rand}})$. Other structural assumptions on g_0 which allow for identification of outcomes in counterfactual contexts could be used instead of Assumption 6, such as transport via quantile functions in Athey & Imbens (2006).

Below, we present a generalization of Theorem 1 to the setting of within-subject varying context:

Theorem 3 (Identifiability of individual-level regret)**.** *Let* j *be an N-of-1 subject for which Assumptions 1, 2 and 4 to 6 hold. Then, for all* $a \in \mathcal{A}^j$ *and all* $X_{\text{rand}} \in \operatorname{supp} \kappa(\cdot \mid X^j_{\text{fix}})$*, the counterfactual surrogate responses* $S^j(a; X_{\text{rand}})$ *are identified, and the outcome labels* $\mu^j(a)$ *are identified via*

$$\mu^j(a) = \int \mathbb{E}^{\text{obs}}\left[Y \,\middle|\, A = a,\ X_{\text{fix}} = X^j_{\text{fix}},\ X_{\text{rand}}, S = S^j(a; X_{\text{rand}}) \right] d\kappa(X_{\text{rand}} \mid X^j_{\text{fix}}). \tag{13}$$

Theorem 3 gives rise to the generalization of the surrogate map evaluation pipeline in Appendix A.4.1: first, we regress the surrogate response to obtain the decomposition $S = g^\circ(A, X_{\text{fix}}, U) + h(X_{\text{fix}}, X_{\text{rand}})$. This allows us to impute the context-counterfactual surrogate $S^j(a; X'_{\text{rand}}) = S^j(a; X_{\text{rand}}) - h(X^j_{\text{fix}}, X_{\text{rand}}) + h(X^j_{\text{fix}}, X'_{\text{rand}})$. Second, as before, we regress the outcome on the treatment, fixed and varying covariates, and the surrogate. Finally, we average the resulting regression function over the context variable distribution of subject j. For a constant context, (13) reduces to the pseudo-outcome (5).

Proof. We show that for every possible context X_{rand} of subject j, i.e., every context X_{rand} in the support of $\kappa(\cdot \mid X^j_{\text{fix}})$, the context-dependent outcome mean $\mu^j(a; X_{\text{rand}})$ equals the population regression of Y evaluated at the subject's counterfactual surrogate response $S^j(a; X_{\text{rand}})$, and that this surrogate response is identified. Averaging over contexts then gives the claim. Write $m(a, s, X_{\text{fix}}, X_{\text{rand}}, u) := \mathbb{E}[f_0(a, s, X_{\text{fix}}, X_{\text{rand}}, u, \epsilon_Y)]$, so that

$$\mu^j(a; X_{\text{rand}}) = m\big(a, S^j(a; X_{\text{rand}}), X^j_{\text{fix}}, X_{\text{rand}}, U^j\big).$$

Let $(a, X^j_{\text{fix}}, X_{\text{rand}}, s)$ lie in the support of $(A, X_{\text{fix}}, X_{\text{rand}}, S)$ under $\mathbb{P}^{\text{obs}}$. On the event $\{A = a, X_{\text{fix}} = X^j_{\text{fix}}, X_{\text{rand}} = X_{\text{rand}}, S = s\}$, the latent type U of a population-study

participant satisfies $g_0(a, X^j_{\text{fix}}, X_{\text{rand}}, U) = s$, so by Assumption 2, we have $U = u_s := g_0(a, X^j_{\text{fix}}, X_{\text{rand}}, \cdot)^{-1}(s)$ and $Y = f_0(a, s, X^j_{\text{fix}}, X_{\text{rand}}, u_s, \epsilon_Y)$. Since ϵ_Y is exogenous, the conditional distribution of Y given this event is the distribution of $f_0(a, s, X^j_{\text{fix}}, X_{\text{rand}}, u_s, \epsilon_Y)$, and the mean is $m(a, s, X^j_{\text{fix}}, X_{\text{rand}}, u_s)$. For $s = S^j(a; X_{\text{rand}}) = g_0(a, X^j_{\text{fix}}, X_{\text{rand}}, U^j)$, Assumption 2 gives $u_s = U^j$. Thus, the conditional distribution of Y given the event equals the distribution of $Y^j(a; X_{\text{rand}})$, and its mean is $\mu^j(a; X_{\text{rand}})$.

Let X^a_{rand} be the context in which subject j received treatment a, and let X'_{rand} be any context in the support of $\kappa(\cdot \mid X^j_{\text{fix}})$. By Assumption 6,

$$S^j(a; X'_{\text{rand}}) - S^j(a; X^a_{\text{rand}}) = g_0\big(a, X^j_{\text{fix}}, X'_{\text{rand}}, U^j\big) - g_0\big(a, X^j_{\text{fix}}, X^a_{\text{rand}}, U^j\big) = h\big(X^j_{\text{fix}}, X'_{\text{rand}}\big) - h\big(X^j_{\text{fix}}, X^a_{\text{rand}}\big). \tag{14}$$

In the population study, $A = \pi_0(X_{\text{fix}}, X_{\text{rand}}, \epsilon_A)$ with ϵ_A exogenous, so by Assumption 4, we have $U \perp\!\!\!\perp (A, X_{\text{rand}}) \mid X_{\text{fix}}$. In other words, among participants with treatment a and covariates X_{fix}, the latent types have the same distribution in every context. Thus, we have

$$\mathbb{E}^{\text{obs}}[S \mid A = a, X_{\text{fix}} = X_{\text{fix}}, X_{\text{rand}} = X_{\text{rand}}] = G(a, X_{\text{fix}}) + h(X_{\text{fix}}, X_{\text{rand}}),$$

where $G(a, X_{\text{fix}}) := \mathbb{E}^{\text{obs}}[g^\circ(a, X_{\text{fix}}, U) \mid X_{\text{fix}} = X_{\text{fix}}]$ does not depend on the context, and both sides are defined for every X_{rand} in the support of $\kappa(\cdot \mid X_{\text{fix}})$ by Assumption 5. The difference of the conditional means at X'_{rand} and X^a_{rand}, with $X_{\text{fix}} = X^j_{\text{fix}}$, is therefore the right-hand side of (14), so $S^j(a; X'_{\text{rand}})$ is identified.

By Assumption 1, the observed in-context surrogate response $(a, X^j_{\text{fix}}, X^a_{\text{rand}}, S^j(a; X^a_{\text{rand}}))$ lies in the support of $(A, X_{\text{fix}}, X_{\text{rand}}, S)$, so U^j lies in the support of U given $(A, X_{\text{fix}}) = (a, X^j_{\text{fix}})$. This conditional distribution is the same in every context, and treatment a occurs in context X'_{rand} by Assumption 5. Hence $(a, X^j_{\text{fix}}, X'_{\text{rand}}, S^j(a; X'_{\text{rand}}))$ also lies in the support of $(A, X_{\text{fix}}, X_{\text{rand}}, S)$.

Hence, at $\big(a, X^j_{\text{fix}}, X'_{\text{rand}}, S^j(a; X'_{\text{rand}})\big)$ for every X'_{rand} in the support of $\kappa(\cdot \mid X^j_{\text{fix}})$, we have

$$\mu^j(a; X'_{\text{rand}}) = \mathbb{E}^{\text{obs}}\big[Y \mid A = a, X_{\text{fix}} = X^j_{\text{fix}}, X_{\text{rand}} = X'_{\text{rand}}, S = S^j(a; X'_{\text{rand}})\big], \tag{15}$$

where $S^j(a; X'_{\text{rand}})$ is identified. Finally, since X_{rand} is a pretreatment variable and by Assumption 4, we have

$$\mu^j(a) = \int \mu^j(a; X_{\text{rand}})\, d\mathbb{P}^j_{X_{\text{rand}}}(X_{\text{rand}}) = \int \mu^j(a; X_{\text{rand}})\, d\kappa(X_{\text{rand}} \mid X^j_{\text{fix}}),$$

which together with (15) gives the claim. □

A.6.3 Simplified regret identifiability in the linear case

In general, generalized individual-level regret computation as given by Theorem 3 requires averaging the imputed outcome over the distribution of context covariates X_{rand}. In this section, we show that if the outcome model is linear in the surrogate and additive in the context variables, this computation can be simplified: the individual-level regret can be instead obtained by *adjusting* the surrogate w.r.t. the context variables, and then plugging in the adjusted value into the outcome model, dropping all terms that involve the context variables.

Assumption 7 (Linear outcome model)**.** *There are measurable functions α and ω and a vector θ such that*

$$Y = f_0(A, S, X_{\text{fix}}, X_{\text{rand}}, U, \epsilon_Y) = \alpha(A, X_{\text{fix}}, U) + \theta^\top S + \omega(X_{\text{fix}}, X_{\text{rand}}) + \epsilon_Y, \qquad \mathrm{E}[\epsilon_Y] = 0.$$

The treatment and the fixed covariates may enter α arbitrarily, and the context may affect the outcome, both through ω and through the surrogate shift. Denote by $S^{\text{adj},j}(a) = S^j(a) - h(X^j_{\text{fix}}, X^a_{\text{rand}})$ the surrogate potential outcome of the subject j from which the additive influence of the context variables X^a_{rand} has been removed. We call $S^{\text{adj},j}(a)$ the *adjusted* surrogate response. We have the following corollary in the linear case:

Corollary 2 (Regret identifiability in the linear case)**.** *Let the assumptions of Theorem 3 and Assumption 7 hold. For an N-of-1 subject j, define*

$$\bar{\mu}^j(a) := \alpha(a, X^j_{\text{fix}}, U^j) + \theta^\top S^{\text{adj},j}(a), \qquad a \in \mathcal{A}^j. \tag{16}$$

Then, for all $a \in \mathcal{A}^j$,

$$\mu^j(a) = \bar{\mu}^j(a) + c^j, \quad \textit{where} \quad c^j := \int \left[\theta^\top h(X^j_{\text{fix}}, X_{\text{rand}}) + \omega(X^j_{\text{fix}}, X_{\text{rand}})\right] d\mathbb{P}^j_{X_{\text{rand}}}(X_{\text{rand}}).$$

Further, $\bar{\mu}^j$ is identified up to a treatment-independent constant from a single context: for any reference context $X^0_{\text{rand}} \in \operatorname{supp} \kappa(\cdot \mid X^j_{\text{fix}})$,

$$\mathrm{E}^{\text{obs}}\left[Y \mid A = a,\ X_{\text{fix}} = X^j_{\text{fix}},\ X_{\text{rand}} = X^0_{\text{rand}},\ S = S^{\text{adj},j}(a) + h(X^j_{\text{fix}}, X^0_{\text{rand}})\right] = \bar{\mu}^j(a) + d^j, \tag{17}$$

where $d^j := \theta^\top h(X^j_{\text{fix}}, X^0_{\text{rand}}) + \omega(X^j_{\text{fix}}, X^0_{\text{rand}})$. Since c^j and d^j do not depend on the treatment, and the individual-level regret $r^j(f)$ and ranking error $\ell^j(f)$ (Equation (8)) are invariant to treatment-independent shifts, both are identified by the left-hand side of (17), without averaging over contexts.

Proof. Fix $a \in \mathcal{A}^j$ and write $s^{\text{adj}} := S^{\text{adj},j}(a)$. By consistency, $S^j(a) = S^j(a; X^a_{\text{rand}})$, so Assumption 6 gives $s^{\text{adj}} = g^\circ(a, X^j_{\text{fix}}, U^j)$ and $S^j(a; X_{\text{rand}}) = s^{\text{adj}} + h(X^j_{\text{fix}}, X_{\text{rand}})$ for every context X_{rand}. By Assumption 7,

$$Y^j(a; X_{\text{rand}}) = \alpha(a, X^j_{\text{fix}}, U^j) + \theta^\top s^{\text{adj}} + \theta^\top h(X^j_{\text{fix}}, X_{\text{rand}}) + \omega(X^j_{\text{fix}}, X_{\text{rand}}) + \epsilon_Y,$$

so $\mu^j(a; X_{\text{rand}}) = \bar{\mu}^j(a) + \theta^\top h(X^j_{\text{fix}}, X_{\text{rand}}) + \omega(X^j_{\text{fix}}, X_{\text{rand}})$, and integrating over $\mathbb{P}^j_{X_{\text{rand}}}$ gives $\mu^j(a) = \bar{\mu}^j(a) + c^j$.

For (17), let $s^0 := s^{\text{adj}} + h(X^j_{\text{fix}}, X^0_{\text{rand}}) = S^j(a; X^0_{\text{rand}})$. As shown in the proof of Theorem 3, $(a, X^j_{\text{fix}}, X^0_{\text{rand}}, s^0)$ lies in the support of $(A, X_{\text{fix}}, X_{\text{rand}}, S)$ under $\mathbb{P}^{\text{obs}}$, and on the event $\{A = a, X_{\text{fix}} = X^j_{\text{fix}}, X_{\text{rand}} = X^0_{\text{rand}}, S = s^0\}$ the latent type of a population participant equals U^j by Assumption 2. Since ϵ_Y is exogenous with mean zero, Assumption 7 gives

$$\mathrm{E}^{\text{obs}}\left[Y \mid A = a, X_{\text{fix}} = X^j_{\text{fix}}, X_{\text{rand}} = X^0_{\text{rand}}, S = s^0\right] = \alpha(a, X^j_{\text{fix}}, U^j) + \theta^\top s^0 + \omega(X^j_{\text{fix}}, X^0_{\text{rand}}) = \bar{\mu}^j(a) + d^j.$$

By the same argument, h is identified up to an additive function of X_{fix}. Finally, c^j and d^j do not depend on a, and $r^j(f)$ and $\ell^j(f)$ are invariant under within-subject constant shifts of the outcome means, which proves the claim. □

We note that in the linear setting, a similar simplification applies to the policy π_f induced by a surrogate map f. If a surrogate map f is applied to the adjusted surrogate, i.e., it assigns $(s, X_{\text{fix}}, X_{\text{rand}}) \mapsto f(s - h(X_{\text{fix}}, X_{\text{rand}}), X_{\text{fix}})$, the surrogate value $\mathbb{E}^j[f(S^j(a), X^j)]$ equals $f(S^{\text{adj},j}(a), X^j_{\text{fix}})$ and the averaging over X_{rand} can be skipped. In the special case that f is linear in s, i.e., $f(s, x) = \gamma^\top s + \phi(x)$, the surrogate value $\mathbb{E}^j[f(S^j(a), X^j)]$ equals $\gamma^\top S^{\text{adj},j}(a)$ plus a treatment-independent term. Here, f induces the same policy as $\gamma^\top S^{\text{adj},j}(a)$, and again no averaging is needed.

Finally, we note that since $\theta^\top s + \omega(X_{\text{fix}}, X_{\text{rand}}) = \theta^\top\{s - h(X_{\text{fix}}, X_{\text{rand}})\} + \tilde{\omega}(X_{\text{fix}}, X_{\text{rand}})$, Assumption 7 also holds with S replaced by S^{adj}. The outcome model can therefore be equivalently fitted on the adjusted surrogate values directly.

A.6.4 Application to the case study in Section 5 and the Cox model

In our real-world case study in Section 5, we consider a time-to-event outcome and model it by a Cox model. We measure the outcome labels on the negative log-hazard scale (Section 5). We note that although we have formulated Appendix A.6.3 in the case of a partially linear outcome structural equation, the same result applies to any functional of the distribution of Y which is linear in S, since Theorem 3 shows equality of the full distribution. In particular, we assume partially linear structure of the following transformation of Y, where Λ_0 is a cumulative baseline hazard:

$$\log \Lambda_0(Y) := -\eta + \epsilon_Y, \qquad \eta := \theta^\top S + \rho(A, X_{\text{fix}}, U) + \omega(X_{\text{fix}}, X_{\text{rand}}).$$

Equivalently, given its parents, the time-to-event outcome Y has hazard $\lambda_0(t)\exp\{\eta\}$ linear in S. Then, the conditional hazard of Y in the population study is

$$\lambda_0(t)\exp\left\{\eta_1\big(a, X_{\text{fix}}, s - h(X_{\text{fix}}, X_{\text{rand}})\big) + \theta^\top h(X_{\text{fix}}, X_{\text{rand}}) + \omega(X_{\text{fix}}, X_{\text{rand}})\right\},$$

which is a Cox model in the adjusted surrogate. It is identified from right-censored outcomes if censoring is independent of Y given A, X_{fix}, X_{rand} and S. We define the target as the negative log relative hazard averaged over contexts,

$$\mu^j(a) := -\int \left[\theta^\top S^j(a; X_{\text{rand}}) + \rho(a, X^j_{\text{fix}}, U^j) + \omega(X^j_{\text{fix}}, X_{\text{rand}})\right] d\mathbb{P}^j_{X_{\text{rand}}}(X_{\text{rand}}).$$

An analogous computation to the proof of Corollary 2 then gives $\mu^j(a) = -\bar{\mu}^j(a) - c^j$. Hence the individual-level regret on the negative log-hazard scale is identified by the pseudo-outcome $-\eta_1(a, X^j_{\text{fix}}, S^{\text{adj},j}(a))$, without averaging over the context variables.

In the case study Section 5, we consider context variables season (encoded as $\sin(2\pi\text{month}/12)$), calendar (absolute time in months) and tenure (time in months since j enrolled in the study). By interacting the context variables with subject-fixed variables such as baseline age, or location, resulting in $\zeta = (X_{\text{rand}}, X_{\text{rand}} \otimes X_{\text{fix}})$, we are able to capture variables such as current age and current weather. We outline our practical adjustment pipeline for the linear survival setting below:

(1) Regress S on A, X_{fix}, X_{rand} and $X_{\text{rand}} \otimes X_{\text{fix}}$, and set $S^{\text{adj}} := S - \hat{B}(X_{\text{rand}}, X_{\text{rand}} \otimes X_{\text{fix}})$. Here, $h(X_{\text{fix}}, X_{\text{rand}}) = B(X_{\text{rand}}, X_{\text{rand}} \otimes X_{\text{fix}})$ as in the proof of Theorem 3, provided that the regression is correctly specified.

(2) Fit a Cox model with linear predictor $\beta_A + \beta_S^\top S^{\text{adj}} + \beta_f^\top X_{\text{fix}} + \beta_\zeta^\top \zeta$.

(3) Impute pseudo-outcomes $-(\beta_{A_i^j} + \beta_S^\top S_i^{\text{adj},j})$, while ignoring the terms that do not involve the treatment.

By Corollary 2, the resulting pseudo-regret is therefore the individual-level regret on the negative log-hazard scale. All surrogate maps, including the single biomarkers, are evaluated on S^{adj}.

A.7 Learning surrogate maps

In Sections 2 and 3, we take a surrogate map as given and concentrate on the evaluation question, i.e., how can we, from data, find which surrogate map is better in the individual sense. In this section, we discuss the logical next step: how would one *learn* a good individual-level surrogate map?

Throughout this section, we impose Assumption 3 (no direct effect): by the same argument as in Section 3.2, if we allow arbitrary direct effect of new treatments, no map $f(S, X)$ can be guaranteed to rank even a single pair of new treatments correctly. For simplicity, as in Section 3, we assume that the subjects' covariates X^j are $\mathbb{P}^j$-a.s. constant. Generalization to within-subject varying covariates proceeds similarly to Appendix A.6.

Unless stated otherwise, in this section, we do *not* require the latent type U^j to be constant within a subject and do not impose Assumption 2.

A.7.1 Learning with N-of-1 data: empirical regret minimization

The simplest setting for learning an individual-level surrogate map is the following: Suppose that at learning time, we have access to *both* a sample $\mathcal{D}^{\text{obs}}$ from the observational study and a collection $\mathcal{J}^{\text{tr}}$ of N-of-1 subjects with repeated surrogate measurements. Then, following Algorithm 1, we can directly compute and minimize the corresponding training-time pseudo-regret:

$$\hat{f} \in \operatorname*{argmin}_f \widehat{R}_{\mathcal{J}^{\text{tr}}}(f), \tag{18}$$

where $\widehat{R}_{\mathcal{J}^{\text{tr}}}$ is the empirical pseudo-regret computed on the training N-of-1 subjects $\mathcal{J}^{\text{tr}}$ following Algorithm 1. Instead of directly minimizing the regret or the corresponding ranking error (8), in practice, we could replace the 0-1 loss by, e.g., a pairwise logistic loss on $f(S^j(a), X^j) - f(S^j(a'), X^j)$ (finite-sample properties of such ranking losses are studied in, e.g., Clémençon et al. (2008)). We refer to (18) as the *regret-minimizing map*.

Minimizing (18) comes with theoretical guarantees: if Assumptions 1 and 2 hold for the training-time N-of-1 subjects, the empirical pseudo-regret is a consistent estimator of the true regret on the training subjects via Theorem 1, and under sufficient training data variability, the surrogate map generalizes to arbitrary mediated new treatments via Theorem 2.

However, computing the regret-minimizing map requires N-of-1 data *at training time*, which is arguably not satisfied in most applications: More commonly, one aims to obtain a surrogate map *before* starting any N-of-1 experiments, to be able to give individualized recommendations as the subjects receive treatments. Hence, we devote the rest of this section to the more realistic scenario in which only data from an observational study is available to learn a surrogate map.

A.7.2 Learning with only observational data

Assume now that at learning time we only have access to $\mathcal{D}^{\text{obs}}$, a sample from the observational study $\mathbb{P}^{\text{obs}}(A, X, S, Y)$. We thus want to learn and fix a surrogate map before observing any N-of-1 subjects. The most common such map, which has been extensively studied in the literature and applied in practice, is the so-called *surrogate index* (Athey et al., 2026; Kallus & Mao, 2025; Chen & Ritzwoller, 2023),

$$f_{\text{SI}}(s, x) := \mathbb{E}^{\text{obs}}\left[Y \mid S = s, X = x\right], \tag{19}$$

given by the conditional expectation of the long-term outcome given the covariates and the surrogate measurements. Further, we define the *causal surrogate map*

$$f_{\text{causal}}(s, x) := \mathbb{E}^{\text{obs}}\left[Y \mid X = x, \text{do}(S = s)\right], \tag{20}$$

given by the interventional mean of the outcome when we cut all edges pointing to S and set the surrogate vector to s in the population study $\mathcal{C}^{\text{obs}}$.

We now consider the following toy example which satisfies Assumptions 2 and 3:

Example 2 (Running example)**.** *Consider an observational study $\mathcal{C}^{\text{obs}}$, in which a treatment $A \in \{1, 2\}$ is given (two different doses), and subjects have a latent type $U \in \{1, 2\}$ (degree of responsiveness). Let A and U be independent Bernoulli, and*

$$S := A \cdot U, \qquad Y := (1 - \kappa U)\, S - c\, U + \epsilon_Y, \tag{21}$$

with (not yet chosen) parameters $\kappa \geq 0$ and $c \geq 0$. Additionally, consider an N-of-1 subject j with latent type $u_j \in \{1, 2\}$ and a set of doses $\mathcal{A}^j \subset (0, \infty)$, with surrogate responses $S^j(a) = a\, u_j$. We thus have $S^j(2) - S^j(1) = u_j$. In words, strong responders ($u_j = 2$) show twice as strong surrogate response to the treatment compared to weak responders ($u_j = 1$).

Observe that the surrogate response $S = 1$ can only be observed in a weak responder with $A = 1$, and the response $S = 4$ can only be observed in a strong responder given $A = 2$. However, the response $S = 2$ could arise from $(A, U) = (2, 1)$ and from $(A, U) = (1, 2)$ with equal probability, and thus, on its own, carries no information about the latent type. Hence $\mathbb{E}^{\text{obs}}[U \mid S = s] = 1,\ \frac{3}{2},\ 2$ for $s = 1, 2, 4$, and $\mathbb{E}^{\text{obs}}[U] = \frac{3}{2}$. We explicitly compute both maps and obtain

$$f_{\text{SI}}(s) = s - \kappa\, s\ \mathbb{E}^{\text{obs}}[U \mid s] - c\ \mathbb{E}^{\text{obs}}[U \mid s]; \tag{22}$$
$$f_{\text{causal}}(s) = \left(1 - \tfrac{3}{2}\kappa\right)s - \tfrac{3}{2}c.$$

We further compute the true outcome contrasts within each type:

$$\mu^j(2) - \mu^j(1) = (1 - \kappa u_j)\, u_j = \begin{cases} 1 - \kappa, & u_j = 1, \\ 2(1 - 2\kappa), & u_j = 2. \end{cases} \tag{23}$$

The surrogate index can fail at the individual level while valid at the population level. Consider an instantiation of Example 2 with $\kappa = 0$, i.e. no surrogate-confounding interactions. By (23), the higher dose of the treatment is better for both latent types. The causal map $f_{\text{causal}}(s) = s - \frac{3}{2}c$ ranks the pair of treatments correctly within both latent types. Thus, it has zero regret and ranking error $r^j(f_{\text{causal}}) = \ell^j(f_{\text{causal}}) = 0$ for every N-of-1 subject. In contrast, the surrogate index in (22) gives

$$f_{\text{SI}}(2) - f_{\text{SI}}(1) = 1 - \tfrac{c}{2}, \qquad f_{\text{SI}}(4) - f_{\text{SI}}(2) = 2 - \tfrac{c}{2}.$$

Thus, for $2 < c < 4$, the index misranks the treatments for every low responder $u_j = 1$, i.e., $\ell^j(f_{\text{SI}}) = 1$ and $r^j(f_{\text{SI}}) = 1$ for $u_j = 1$.

We show that this failure of the surrogate index only becomes apparent on the individual level: in fact, the surrogate index can still correctly estimate the sign of the average treatment effect. If the two doses are randomized in an experiment on subjects drawn from $\mathcal{C}^{\text{obs}}$, the true average treatment effect is $\mathbb{E}^{\text{obs}}[Y(2) - Y(1)] = \mathbb{E}^{\text{obs}}[U] = \frac{3}{2} > 0$, while the surrogate index estimates

$$\mathbb{E}^{\text{obs}}\left[f_{\text{SI}}(S(2)) - f_{\text{SI}}(S(1))\right] = \tfrac{3}{2} - \tfrac{c}{2},$$

which has the correct sign if and only if $c < 3$. For $c \in (2, 3)$, the surrogate index therefore orders the two doses correctly by their average effect, and is in this sense a valid population-level surrogate, yet misranks the treatments for every N-of-1 low responder subject. Note that this is not in contradiction with the fact that f_{SI} is the best predictor of Y from (S, X), but it illustrates once more that the goal of correctly ranking treatments on an individual level does not correspond to, and can go against predictiveness of a surrogate map, as empirically demonstrated in Figure 5 in our case study Section 5.

A.7.3 Optimality of the causal map under additive confounding

Example 2 suggests that in a range of settings, removing the influence of confounding from the surrogate map might improve individual-level surrogacy. In the following assumption, we effectively characterize such settings:

Assumption 8 (Additive confounding)**.** *In the outcome structural equation, S and U are separable. In other words, there exist measurable functions h and k such that*

$$Y = h(S, X, \epsilon_Y) + k(X, U, \epsilon_Y).$$

Intuitively, Assumption 8 restricts heterogeneity of the treatment effect to the level of surrogate responses: subjects with different latent types might have different surrogate responses to the same treatment, but once these responses are observed, the remaining heterogeneity of the long-term outcome is exogenous, i.e. not introduced by U. This disallows hypothetical situations, where, for instance, a type of subjects tends to quickly increase their VO2max after exercise, but a subtype of these subjects sees no benefit of this improvement for their long-term health, whereas another, indistinguishable subtype greatly benefits from the VO2max increase.

Theorem 4 (Optimality of the causal map under additive confounding)**.** *Let Assumption 3 and Assumption 8 hold. Let* $\bar{h}(s,x) = \mathbb{E}^{\text{obs}}_{\epsilon_Y}[h(s,x,\epsilon_Y)]$ *and* $\bar{k}(x) = \mathbb{E}^{\text{obs}}[k(x,U,\epsilon_Y) \mid X = x]$*. Then* $f_{\text{causal}}(s,x) = \bar{h}(s,x) + \bar{k}(x)$*, and for every N-of-1 subject* j *with a finite treatment set* $\mathcal{A}^j$*, the causal map has zero individual-level regret:*

$$r^j(f_{\text{causal}}) = \ell^j(f_{\text{causal}}) = 0,$$

where ℓ^j *is the treatment ranking error* (8)*.*

Proof. The expression for f_{causal} follows from its definition (20) and Assumption 8. Fix a subject j with $X^j = x$. Let $\bar{m}$ be as in Assumption 3. We have for every treatment $a \in \mathcal{A}^j$:

$$\mu^j(a) = \mathbb{E}^j\left[\bar{m}\big(S^j(a), X^j, U^j\big)\right]. \tag{24}$$

From Assumption 8, we obtain

$$\begin{aligned}\mu^j(a) &= \mathbb{E}^j\left[h(S^j(a), x, \epsilon_Y)\right] + \mathbb{E}^j\left[k(x, U^j, \epsilon_Y)\right] \\ &= \mathbb{E}^{\text{obs}}_{\epsilon_Y}[h(S^j(a), x, \epsilon_Y)] + c^j(x) = \mathbb{E}^j[f_{\text{causal}}(S^j(a), x)] + c^j(x),\end{aligned}$$

where $c^j(x)$ is an error term which does not depend on a. Hence, the causal map estimates $\mu^j(a)$ up to a treatment-independent offset, and thus $r^j(f_{\text{causal}}) = \ell^j(f_{\text{causal}}) = 0$. □

We note that optimality of the causal map requires neither stability nor invertibility of the latent type U^j as in Assumption 2. Further, the guarantee is uniform over treatments and N-of-1 subjects: it applies to (fully mediated) treatments that never appear in the observational study.

Learning the causal surrogate map. We now give a brief overview of methods via which the causal surrogate map can be learned in our setting. Since Assumption 3 is equivalent to exclusion restriction in instrumental variable literature, it makes the treatments valid instruments for the causal effect of S on Y (Angrist et al., 1996). Under Assumption 8, estimating the causal map is equivalent to estimating $\bar{h}(s,x)$, and thus solving nonparametric instrumental-variable equations in the unknown $\bar{h}$. Under further structural assumptions, one can use two-stage least squares, nonparametric or neural IV estimators (Newey & Powell, 2003; Hartford et al., 2017), control-function estimators (Newey et al., 1999; Imbens & Newey, 2009), anchor regression (Rothenhäusler et al., 2021), and proximal methods (Tchetgen Tchetgen et al., 2024). In Section 5, we evaluate a survival version of the control-function map as a causal surrogate map.

A.7.4 Beyond additive confounding

A natural question is whether the causal surrogate map is also optimal in case when the confounding is not additive, i.e., one allows for interactions between S and U in the outcome structural equation. However, it becomes apparent quite quickly that if the latent type influences the long-term outcome beyond the surrogate responses, the within-subject ordering of two surrogate values may itself depend on the latent type, and thus, no fixed function of (S, X) can be optimal for all subjects. More concretely, let j and j' be two N-of-1 subjects with the same covariates x, and consider two pairs of treatments $a, b \in \mathcal{A}^j$, $a', b' \in \mathcal{A}^{j'}$ such that $S^j(a) = S^{j'}(a') = s$, $S^j(b) = S^{j'}(b') = s'$. If the confounding mechanism is unrestricted, one can easily construct f_0 such that $\mu^j(a) > \mu^j(b)$, but $\mu^{j'}(a') < \mu^{j'}(b')$, and thus, for every surrogate map evaluated on the two subjects, its regret is given by

$$r^j\big(f; \{a,b\}\big) + r^{j'}\big(f; \{a',b'\}\big) \;\geq\; \min\left\{\mu^j(a) - \mu^j(b),\; \mu^{j'}(b') - \mu^{j'}(a')\right\} > 0. \tag{25}$$

Thus, no fixed map $f(S,X)$ can uniformly achieve zero regret if one allows for non-additive confounding. Consequently, one might ask: if no surrogate map achieves perfect regret, can we at least show that one of the considered surrogate maps is always better than the other? We return to our running example to provide a negative answer:

Surrogate index vs. causal map under general confounding. We return to Example 2 and set $c = 0$ and $\kappa \in (\frac{1}{2}, 1)$, i.e. $Y = (1 - \kappa U)S + \epsilon_Y$. By (23), for low responders, the higher dose is the better treatment, whereas for high responders, the lower dose is better. In other words, increasing the surrogate is beneficial for the non-responsive type, and harmful for the responsive type. Since the causal map (22) cannot differentiate between types with the same surrogate response, it prefers the larger surrogate value for both types as long as $\kappa < \frac{2}{3}$. In contrast, the surrogate index satisfies $f_{\mathrm{SI}}(2) - f_{\mathrm{SI}}(1) < 0,\ f_{\mathrm{SI}}(4) - f_{\mathrm{SI}}(2) < 0$, and always prefers the smaller surrogate value.

Therefore, for $\kappa \in (\frac{1}{2}, \frac{2}{3})$:

(i) The surrogate index is wrong for low responders, and the causal map is right. We have $r^j(f_{\mathrm{causal}}) = 0$, $r^j(f_{\mathrm{SI}}) = 1 - \kappa$.

(ii) The surrogate index is right for high responders, the causal map is wrong: $r^j(f_{\mathrm{causal}}) = 2(2\kappa - 1)$, $r^j(f_{\mathrm{SI}}) = 0$.

Hence, depending on the proportion of both types in the population, both surrogate maps can incur large regret. Indeed, our example allows us to show that, more generally, observational data are not sufficient to determine the best surrogate map, whatever learning rule is used:

Proposition 2 (Observational data do not determine the best map)**.** *Let $\mathbb{P}^{\mathrm{obs}}$ be given by Example 2 with $c = 0$ and $\kappa \in (\frac{1}{2}, 1)$. There exist two N-of-1 populations $\mathcal{J}_1$ and $\mathcal{J}_2$, both generated by the mechanisms of Example 2 and satisfying Assumption 3, and a constant $c_0 > 0$, such that*

$$\inf_f R_{\mathcal{J}_1}(f) = \inf_f R_{\mathcal{J}_2}(f) = 0 \qquad \textit{but} \qquad \max\left\{R_{\mathcal{J}_1}(f),\, R_{\mathcal{J}_2}(f)\right\} \geq c_0 \quad \textit{for every surrogate map } f.$$

In particular, any map $\hat{f} = \Phi(\mathcal{D}^{\mathrm{obs}})$ that only depends on the observational data alone and does not have knowledge of the N-of-1 population has worst-case excess regret at least c_0.

Proof. Consider a population $\mathcal{J}_1$ of low responders exposed to treatments $\mathcal{A}^{\mathcal{J}_1} = \{1, 2\}$ and a population $\mathcal{J}_2$ of high responders exposed to $\mathcal{A}^{\mathcal{J}_2} = \{\frac{1}{2}, 1\}$. The surrogate responses of both populations are $(1, 2)$. However, for low responders, the larger surrogate value is better, with outcome gap $1 - \kappa$, and high responders, the smaller one is better, with outcome gap $(2\kappa - 1) \cdot 2 \cdot \frac{1}{2} = 2\kappa - 1$. In particular, $f(s) = s$ is a perfect surrogate map on $\mathcal{J}_1$ and $f(s) = -s$ is a perfect surrogate map on $\mathcal{J}_2$. By (25), for every surrogate map f, the sum of the two regrets is at least $\min\{1 - \kappa,\, 2\kappa - 1\}$, so the maximum (which is paid by the surrogate map with no knowledge of the N-of-1 population) is at least $c_0 := \frac{1}{2}\min\{1 - \kappa,\, 2\kappa - 1\} > 0$. □

In Appendix A.7.3, we have seen that under additive confounding, a single surrogate map (the causal map) is optimal for every population of N-of-1 subjects, and such a map can be learned from the observational study only. In contrast, in the general case, no such surrogate map exists.

A.8 Additional details on the synthetic experiments

Below, we present additional details for the synthetic experiments in Section 4, including the data-generating process and supplementary ablation studies.

A.8.1 Data-generating process

We generate synthetic data for learning and evaluating surrogate maps from the following subject-level structural causal model on treatments A, covariates X, latent type U, surrogates S and outcome Y:

$$\begin{aligned}
U^j &= \mu_0 + \Lambda X^j + \sigma_U\, \xi^j, && \xi^j \sim \mathcal{N}(0, I_d),\\
A_i^j &\sim \mathrm{Cat}\big(\mathrm{softmax}_a\, \tau\, w_a^\top X^j\big),\\
S_i^j &= M_{A_i^j}\big(U^j + \sigma_S\, \epsilon_i^j\big) + BX^j, && \epsilon_i^j \sim \mathcal{N}(0, I_d),\\
\eta_i^j &= \underbrace{(\theta + \Theta X^j)^\top S_i^j}_{\text{via surrogate}} + \underbrace{\gamma^\top U^j}_{\text{additive confounding}} + \underbrace{\kappa\,(S_i^j)^\top \Omega\, U^j}_{\text{interactive confounding}} + \lambda^\top X^j,\\
Y_i^j &\sim \mathrm{Bernoulli}\big(\Phi(\eta_i^j/\sigma_Y)\big),
\end{aligned} \tag{26}$$

where Φ denotes the standard normal CDF. The latent type U^j is drawn once per subject; ϵ_i^j however is redrawn at each measurement. Treatment assignment depends on the covariates X^j. We set $d = \dim U = \dim S = 6$ and $m = 4$ for covariates (modelled as age, sex, region, education). For each categorical treatment a, an a-specific surrogate response function is sampled via

$$M_a = \exp(\rho\, \Delta_a), \qquad \Delta_a = \tfrac{1}{2}(G_a - G_a^\top)/\|\tfrac{1}{2}(G_a - G_a^\top)\|_F, \qquad G_a \sim \mathcal{N}(0, I_{d\times d}),$$

so that every M_a is orthogonal. ρ is the rotation magnitude and controls how close all M_a remain to identity. If ρ is very small, surrogate responses to all treatments are relatively similar. Large ρ allows for arbitrary directions of surrogate responses. We produce disjoint treatment sets $\mathcal{A}^{\mathrm{tr}}$ (8 treatments) and $\mathcal{A}^{\mathrm{eval}}$ (6 treatments), so that during evaluation subjects are exposed to treatments never seen during training. The remaining structural parameters are drawn once and shared by every subject and every run: $\mu_0 = \mu\, u$ and $\theta = \alpha\, u'$ with u, u' uniform on the unit sphere of $\mathbb{R}^d$; $\gamma = \gamma_0\, u''$ as well; $\Lambda, B \in \mathbb{R}^{d\times m}$ and $\lambda \in \mathbb{R}^m$ with i.i.d. $\mathcal{N}(0, \cdot^2/m)$ entries; $\Theta \in \mathbb{R}^{d\times m}$ with i.i.d. $\mathcal{N}(0, 1/(dm))$ entries; $\Omega \in \mathbb{R}^{d\times d}$ with i.i.d. $\mathcal{N}(0, 1/d^2)$ entries; and propensity weights $w_a \in \mathbb{R}^m$ with i.i.d. $\mathcal{N}(0, 1/m)$ entries. Unless specified otherwise, we use the following default values:

$d = 6$	$\rho = 1$	$\sigma_U = 1$	$\sigma_Y = 1$
$\|\mu_0\| = 0.2$	$\alpha = \|\theta\| = 1$	$\gamma_0 = \|\gamma\| = 1$	$\tau = 1$
$\|\Lambda\|$-scale $= 0$	B-scale $= 0.4$	Θ-scale $= 0.2$	λ-scale $= 0.3$
$\sigma_S = 0$	$\kappa = 0$		

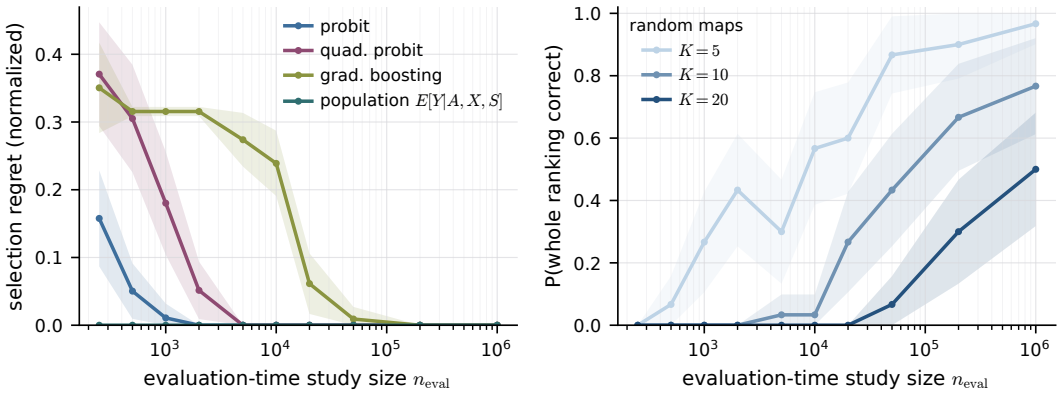
In particular, at default values of parameters, (26) satisfies Assumption 2. We comment on three coefficients of (26): $\|\gamma\|$ is the strength of additive confounding; κ controls the strength of interactive confounding, i.e. the magnitude of the $S \times U$ interaction. Increasing κ from zero thus violates Assumption 8 and makes the causal surrogate map an imperfect individual-level surrogate. Further, $\sigma_S \neq 0$ violates the stability and invertibility of the latent type in Assumption 2, biasing the pseudo-regret away from the true regret.

For the DGP (26) all conditional expectations are available in closed form. We can therefore report both finite-sample and population-level pseudo-regret.

A.8.2 Evaluation setup

For each subject j, we draw covariates $X^j \in \mathbb{R}^4$ to mimic real subject-level covariates: age in decade bins from approximate ACS 2022 marginals over 18-80, sex ($\mathrm{Ber}(0.5)$), region ($\mathrm{Ber}(0.5)$), and educational attainment (six levels with approximate ACS 2022 probabilities $(0.02, 0.05, 0.28, 0.37, 0.20, 0.08)$). Covariates are standardized before entering (26).

For learning and evaluating surrogate maps, we sample three datasets with treatments in either $\mathcal{A}^{\mathrm{tr}}$ or $\mathcal{A}^{\mathrm{eval}}$:



(a) Regret of surrogate map selection

(b) Probability of correct surrogate map ranking

Figure 6: **Ranking a family of candidate surrogate maps. (a)** Map selection regret, normalized by the spread of the candidate surrogate maps' true regrets, as a function of the evaluation-study size n_{eval} for $K = 20$ candidate maps. **(b)** Probability that the estimated pseudo-regret correctly orders the whole collection of K surrogate maps, for $K = 5, 10, 20$ (pooled probit). Bands are 95% CIs over 30 runs.

dataset	role	subjects	samples/subj.	treatments
D^{tr}	learn the surrogate maps	$n_{\text{tr}} = 20\,000$	1, with Y	$\mathcal{A}^{\text{tr}}$
D^{obs}	learn pseudo-label	$n_{\text{eval}} = 10\,000$	1, with Y	$\mathcal{A}^{\text{eval}}$
$\{D^j\}_{j\in\mathcal{J}^{\text{eval}}}$	the N-of-1 subjects scored	$N_{\text{eval}} = 5\,000$	2–3, w.o. Y	$\mathcal{A}^{\text{eval}}$

For N-of-1 subjects, we draw treatments without replacement. Sample sizes in the table represent default sample sizes that we use unless we conduct a sample size experiment. Each reported quantity is a mean over independent resamples of the three datasets with the structural parameters of (26) held fixed. We average over 5 runs by default. Bands and error bars are 95% CIs across runs.

We consider the following pseudo-outcome models: a pooled probit with treatment-specific coefficients for S, X and pooled coefficients for $X \otimes S$, which is correctly specified for additive confounding, i.e., when $\kappa = 0$ in (26); a quadratic probit model, which additionally has treatment-specific coefficients for $X \otimes S$ and the quadratic term $S \otimes S$, which is correctly specified even when $\kappa \neq 0$, and finally, a gradient boosting model.

Every figure reports one of the following, in each case averaged over the N-of-1 subjects $\mathcal{J}^{\text{eval}}$ and then over runs:

- the mean absolute error (MAE) of regret estimation $\frac{1}{|\mathcal{J}^{\text{eval}}|}\sum_j \left|r^j(f) - \hat{r}^j(f)\right|$, which compares the true individual-level regret (3) against the empirical pseudo-regret returned by Algorithm 1;
- When our object of interest is correctness of the ranking of *surrogate maps* by the pseudo-regret, we compare the true ordering of a family of candidate maps $f_1, \ldots, f_K$ by $R_{\mathcal{J}^{\text{eval}}}(f_k)$ against the estimated ordering by $\hat{R}(f_k)$ via
  - the *map selection regret* $R_{\mathcal{J}^{\text{eval}}}\left(f_{\hat{k}}\right) - \min_k R_{\mathcal{J}^{\text{eval}}}(f_k)$ with $\hat{k} = \operatorname{argmin}_k \hat{R}(f_k)$, i.e. the cost of selecting the best surrogate map via pseudo-regret instead of the true regret;
  - $\mathbb{P}$(full map ranking correct), the probability over runs that no pair of maps is misranked by the pseudo-regret;
  - $\mathbb{P}$(best map selected), the probability over runs that $\operatorname{argmin}_k \hat{R}(f_k)$ selects the truly best surrogate map.

A.8.3 Additional finite-sample experiments

In Section 4, we empirically validate how accurately the pseudo-regret recovers the true individual-level regret of a *fixed* surrogate map as a function of the size n_{eval} of the observational study. In practice, we might be more concerned with the question: does choosing the surrogate map with the lowest pseudo-regret actually result in the best surrogate map? In other words, does pseudo-regret allow us to rank surrogate maps correctly? In Figure 2b, we observe that although pseudo-regret itself might be biased as n_{eval} grows, e.g., when the outcome model is misspecified, the three surrogate maps are almost always ranked correctly. In this section, we aim to study this phenomenon more concretely.

We draw a family of K candidate maps uniformly with unit norm in the $(S, X, X \otimes S)$ coefficients. Figure 6a reports the map selection regret for $K = 20$, normalized by the maximum and minimum regret of the maps. All three outcome models eventually identify the best surrogate map, and do so faster than they reasonably approximate regret: for instance, pooled probit selects the best map without error starting at $n_{\text{eval}} = 2000$. In contrast, its MAE of regret estimation is still roughly 2.5% of the available regret. In Figure 6b, we show the probability of correctly ranking the full collection of surrogate maps as a function of the sample size n_{eval} and the number K of candidate surrogate maps. We observe that this is a substantially harder problem. In particular, we conclude that selecting the best surrogate maps out of a collection of surrogate maps requires a smaller evaluation dataset than correctly ranking the full candidate collection.

A.8.4 Learning surrogate maps

In Section 4, we evaluate fixed surrogate maps. In particular, we take analytical expressions of the surrogate index $\mathbb{E}[Y \mid S = s, X = x]$ and the causal surrogate map $\mathbb{E}[Y \mid \text{do}(S = s), X = x]$ (cf. also Appendix A.7) to isolate effects of empirical evaluation. In this section, we study aspects of finite-sample behavior of learning these maps. Below, we outline learning details for each map:

Surrogate index. We fit pooled probit of Y with features $(S,\ X,\ X \otimes S)$ with a ridge penalty to estimate $f_{\text{SI}}(s, x) = \mathbb{P}^{\text{tr}}(Y = 1 \mid S = s, X = x)$ directly.

Control function (Newey et al., 1999). Note that (26) satisfies Assumption 3, and thus the treatments are valid instruments for S. We estimate f_{causal} by two-stage residual inclusion. In the first stage we regress S on $A, A \otimes X$, and compute the residual $v = S - \widehat{\mathrm{E}}[S \mid A, X]$. In the second stage, we perform a probit regression of Y on $(S,\ X,\ X \otimes S,\ A \otimes v)$, and take the coefficients of the first three blocks as the control function surrogate map.

Two-stage least squares with a linear probability model. We fit a linear-probability two-stage least squares estimator with endogenous variables $(S,\ X \otimes S)$ and instruments $(A, A \otimes X)$. We note that this map is misspecified by construction, and cannot correctly estimate the causal surrogate map in the setting of (26).

In Figure 7, we plot the true individual-level regret of the three maps as we increase the number of treatments available at training time.

A.8.5 Ablation: violation of invertibility

We study violations of Assumption 2 by injecting within-subject random noise into the latent type via $S_i^j = M_{A_i^j}\left(U^j + \sigma_S \epsilon_i^j\right) + BX^j$, where the noise $\epsilon_i^j \sim \mathcal{N}(0, I_d)$ is drawn separately for each within-subject measurement. We increase σ_S from 0. At $\sigma_S = 0$ the type is recovered exactly; for non-zero σ_S, we can only recover a noisy version $U^j + \sigma_S \epsilon_i^j$. Invertibility therefore fails as soon as ϵ_S has non-zero variance. In Figure 8, we show how the error of the population-level pseudo-regret increases with the noise level. In the top panel ($\kappa = 0$, additive confounding) the relationship is non-monotonic, as there are two opposite effects of the injected noise: ϵ_S harms recovery of the latent type, biasing the pseudo-outcome, but it also adds exogenous treatment-invariant signal to S, which debiases the pseudo-outcome across treatments. In the bottom panel ($\kappa = 1$), the latent type influences how the surrogate affects the outcome, and the relationship is monotone: as invertibility breaks, the error of regret estimation grows. Additionally, in the top panel, the pooled

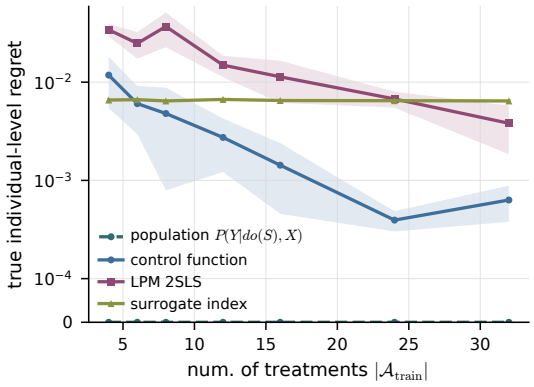


Figure 7: **Learning surrogate maps at an increasing number of training-time treatments.** True individual-level regret of each learned map against the number of training treatments $|\mathcal{A}^{\text{tr}}|$, with $20'000$ samples per treatment. The population causal map has zero regret (cf. Theorem 4). Bands are 95% CIs over 5 runs.

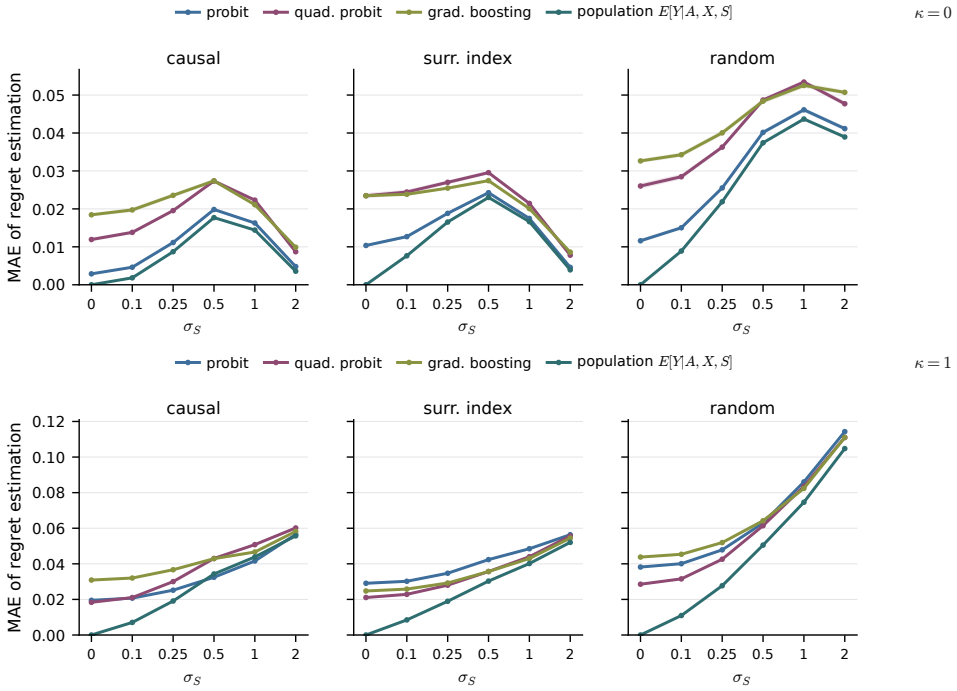


Figure 8: **Error of regret estimation under violation of Assumption 2.** We increase σ_S in (26) from 0 to 2, so that a progressively noisier copy of the latent type drives the surrogate. Top: additive confounding ($\kappa = 0$). Bottom: non-zero surrogate–type interaction ($\kappa = 1$). The population pseudo-label (teal) is the estimand's own error, free of any finite-sample component; $n_{\text{eval}} = 10\,000$, 30 runs.

probit regression is correctly specified and attains the lowest MAE (0.010 at $\sigma_S = 0$). However, in the bottom panel, the quadratic model is correctly specified and has the lowest MAE (0.021).

A.9 Additional details on the AHMS case study in Section 5

A.9.1 Study and cohort

The Apple Heart and Movement Study. We use data collected under informed consent from participants in the Apple Heart and Movement Study (AHMS) (MacRae, 2021; Truslow et al., 2024), a large-scale longitudinal cohort study sponsored by Apple Inc. in partnership with the American Heart Association and Brigham and Women's Hospital. AHMS enrolls U.S. residents aged 18 or older who use an Apple Watch paired with an iPhone. Participants enroll and provide informed consent through the Apple Research app (Truslow et al., 2024), control which data types they share, and may withdraw at any time. At the time of our analysis, the study included over 278,000 participants with up to five years of longitudinal data.

The data comprise wearable-derived metrics from the Apple Watch and iPhone, including heart rate, heart-rate variability, VO_2max, step count, workouts and walking metrics, together with photoplethysmography (PPG) and electrocardiogram (ECG) signals; demographics and medical history from onboarding surveys; and longitudinal outcome information from periodic health surveys and, where available, linked clinical health records.

Cohort filtering. Starting from all AHMS participants, we apply the following criteria: (i) age between 18 and 85 years at enrollment; (ii) BMI between 15 and 50 kg/m^2; (iii) non-null records of all fixed covariates X_{fix}; (iv) no prevalent cardiovascular disease or diabetes at enrollment; (v) at least one valid change window. A valid change window (cf. Figure 3b) consists of a baseline period (3 months), a change period (varied in our experiments) and a followup period (3 months), such that (a) at least one of the activity metrics (cf. Appendix A.9.2) has measurements in both the baseline and the followup period, and (b) every surrogate metric (cf. Appendix A.9.2) has measurements in both the baseline and the follow-up period. Participants with at least two disjoint valid change windows whose treatment labels differ are eligible as *N-of-1 participants* in our evaluation, cf. Appendix A.9.3. We discard all participants whose first cardiometabolic event occurs before or during either of the change windows.

A.9.2 Change windows, treatments, surrogates and outcome

Change windows. For the change window computation, we index time in months since a participant's enrollment. As introduced above and illustrated in Figure 3b, a change window starting at month m consists of a three-month baseline period $\mathcal{B} = [m, m+3)$, a change period $\mathcal{C} = [m{+}3, m{+}3{+}L)$ of L months, and a three-month follow-up period $\mathcal{F} = [m{+}3{+}L, m{+}6{+}L)$. Unless stated otherwise, we use $L = 9$, i.e., change windows of 15 months total duration, as depicted in Figure 3b. In Figure 4, we vary $L \in \{0, 3, 9, 12\}$. For a metric Z and a period $\mathcal{P}$, we write $\bar{Z}_{\mathcal{P}}$ for the average of the daily values of Z over the days in $\mathcal{P}$ with available measurements. We note that in a valid change window, daily recordings might be missing in the whole change period. We only require presence of recordings in the baseline and followup windows. For participants with long recordings, we select as many disjoint windows as possible by greedy search starting at enrollment month.

Fixed covariates. As fixed covariates for each subject, we include age, biological sex, BMI, race/ethnicity (binned), socioeconomic index, medical history, smoking status as well as the baselines of all activity and surrogate metrics.

Treatments. We consider the six activity metrics: step count, active energy burned, exercise time, stand time, walking/running distance and flights climbed. For a change window and every activity metric Z that is valid in it, we compute the within-window change $\Delta Z := \bar{Z}_{\mathcal{F}} - \bar{Z}_{\mathcal{B}}$ and discretize it using the lower and upper quartiles $q^Z_{0.25}$ and $q^Z_{0.75}$ of ΔZ across all valid windows of the full eligible cohort. We set $a = 1$ if $\Delta Z \geq q^Z_{0.75}$, $a = -1$ if $\Delta Z \leq q^Z_{0.25}$ and $a = 0$ otherwise (corresponding to no meaningful change over time). Since we consider six activity metrics, this results in up to 18 distinct labels. This way, most change windows could receive more than one treatment label. In each fold of the experiment (see below) we randomly assign each window one of the valid treatment labels. Table 2 reports the values of the quartile thresholds.

	Eligible cohort ($n = 31065$)	N-of-1 ($n = 14637$)
Demographics		
Age, years	42.8 (13.4)	43.9 (13.3)
Female, %	33.8	30.5
BMI, kg/m^2	27.6 (5.4)	27.5 (5.2)
Race/ethnicity, %		
White	78.7	79.9
Black	3.1	3.0
Asian	4.6	4.5
Hispanic	5.6	5.1
Other	8.0	7.5
Socioeconomic index, %		
Low	53.9	57.0
Medium	21.8	21.0
High	24.2	21.9
Unknown	0.1	0.1
Medical history		
Hypertension, %	19.6	20.1
Hypercholesterolaemia, %	24.9	26.1
Smoking (current / former / never), %	4.3 / 24.8 / 71.0	3.8 / 25.2 / 71.0
Baseline activity		
Daily step count	7301.3 (2960.5)	7426.1 (2982.3)
Active energy burned, kcal/day	606.4 (226.3)	625.3 (227.9)
Exercise time, min/day	37.3 (26.3)	38.2 (26.5)
Stand time, min/day	118.6 (43.8)	120.0 (43.2)
Walking/running distance, km/day	5.6 (2.4)	5.7 (2.4)
Flights climbed, per day	8.5 (7.6)	8.9 (8.1)
Baseline wearable biomarkers		
Resting heart rate, bpm	62.4 (8.2)	61.8 (8.1)
VO_2max, mL/kg/min	33.6 (7.7)	34.1 (7.6)
HRV (SDNN), ms	42.2 (15.1)	41.9 (14.7)
Six-minute-walk distance, m	481.7 (33.8)	483.6 (32.0)
Walking heart rate, bpm	98.9 (11.1)	98.2 (10.9)
Walking speed, m/s	1.2 (0.1)	1.2 (0.1)

Table 1: **Baseline characteristics** of the eligible cohort and of the eligible N-of-1 participants, measured at enrollment. Mean (SD) for continuous variables, % for categorical variables.

		Thresholds	
Activity metric	Unit	$q_{0.25}$	$q_{0.75}$
Step count	steps/day	-1178.9	770.4
Active energy burned	kcal/day	-98.1	46.7
Exercise time	min/day	-7.5	8.6
Stand time	min/day	-17.7	13.8
Walking/running distance	km/day	-0.99	0.55
Flights climbed	flights/day	-1.58	1.24

Table 2: **Treatments.** Activity metrics and the quartile thresholds of their within-window change.

N-of-1 subjects. For a participant j with multiple change windows, the treatment set $\mathcal{A}^j$ consists of the labels assigned to their valid windows. We only consider windows whose assigned labels differ, including pairs of labels defined through different activity metrics (e.g., *steps*$+1$ vs. *exercise time* $+1$); pairs of windows with identical labels are discarded. The individual-level pseudo-regret of j is the average over these pairs, as in (3). Since treatment labels are drawn per fold, the set of treatment pairs of a participant can vary across folds.

Surrogates. The surrogate space consists of the within-window changes $\bar{S}_{\mathcal{F}} - \bar{S}_{\mathcal{B}}$ of the six common wearable biomarkers (resting heart rate/RHR, heart rate variability/HRV, VO2max, Six-

minute-walk distance, walking heart rate, walking speed) (MacRae, 2021). We additionally extend the surrogate space by within-window changes in embeddings of a PPG foundation model (below). Before fitting pseudo-outcome models, we perform adjustment of the surrogate variables for 3 within-subject varying covariates (current month, $\sin(2\pi \cdot \text{month}/12)$, months since enrollment) as described in Appendix A.6.4. Via interactions with baseline covariates, this allows adjustment for seasonal and aging effects. Further, we equalize the scale of the surrogates by z-scoring the change vectors using the cohort-wide mean and standard deviation. Whenever we use wearable markers on their own as a surrogate map, we optionally flip their sign (for RHR and walking heart rate), such that a higher surrogate value always corresponds to a presumed physiological improvement.

PPG embeddings. We use the embeddings of the PPG foundation model of Abbaspourazad et al. (2024), which maps PPG segments to a 256-dimensional embedding vector. For each period, we average the embeddings of all PPG segments recorded in it and take the difference between the follow-up and baseline averages. Additionally, we perform dimensionality reduction of the PPG change vectors via PCA and reduce dimensionality from 256 to 20.

Outcome. As the long-term outcome, we consider time to the first onset of a *cardiometabolic event*, defined as the earliest new diagnosis of any of the following: coronary artery disease, myocardial infarction, coronary revascularization, heart failure, stroke or transient ischemic attack, peripheral artery disease, Type-2 diabetes. Events are based on participants' responses to the periodic study surveys, supplemented by clinical health records (CHR) where available. Time is measured from the beginning of the change window. We use a composite outcome since first diagnoses are relatively rare in our dataset. We choose the outlined conditions since healthy lifestyle and physical activity is consistently associated with lower incidence of cardiometabolic disease (Kyu et al., 2016; Pandey et al., 2015; Paluch et al., 2023; Master et al., 2022). This makes the considered outcome a natural target for activity-based interventions. A participant is classified as having an event if any component occurs after the end of the change window, with the event time taken as the earliest component date; participants without an event are censored at their last date of active follow-up. We note that the Cox pseudo-outcome model in Appendix A.9.3 relies on random censoring given (A, X, S).

A.9.3 Evaluation protocol

Following Section 3 and Appendix A.7, our pipeline uses three disjoint datasets with the following roles: a training-time study D^{tr} on which surrogate maps are fit (cf. Appendix A.9.4), an evaluation-time study D^{obs}, on which we fit the pseudo-outcome model; and a set of N-of-1 participants $\mathcal{J}^{\text{eval}}$ on which pseudo-regret is computed. For participants in D^{tr} and D^{obs}, we only consider one change window, and thus only one treatment: the first valid window starting from enrollment month. For N-of-1 participants, we consider the maximal collection of valid windows.

Pseudo-labels via survival analysis. Since the outcome in the AHMS case study is time-to-event, it is best approximated with a survival model (Clark et al., 2003). We thus instantiate the outcome model of Algorithm 1 as an L_2-penalized Cox proportional-hazards model fitted on D^{obs}, with linear predictor

$$\eta(a, x, s) = \beta_a + \beta_S^\top s + \beta_X^\top x, \tag{27}$$

where s is the context-adjusted (Appendix A.6), z-scored surrogate change S^{adj} and x are the subject's covariates. The penalty is selected by 5-fold cross-validation of the partial likelihood within D^{obs}.

For an N-of-1 participant j and a treatment $a \in \mathcal{A}^j$, we impute the pseudo-label on the negative log-hazard scale,

$$\hat{\mu}^j(a) := -\Big(\hat{\beta}_a + \hat{\beta}_S^\top S^{\text{adj},j}(a)\Big), \tag{28}$$

while ignoring treatment-independent terms. Larger values of $\hat{\mu}^j(a)$ correspond to a lower hazard of a cardiometabolic event, and are thus "healthier". Under proportional hazards, the survival probability $\exp\{-\Lambda_0(t)e^\eta\}$ is decreasing in η at every horizon t, so ranking treatments by (28) is equivalent to ranking them by their survival probability at any fixed horizon; in particular, the ranking error does not depend on a choice of horizon. The regret on a pair of treatments is then a log hazard ratio and thus allows us to interpret the scale of the regret: a pair regret of 0.05 means that the window

selected by the surrogate map has an approximately 5% higher hazard than the better window of the pair.

To keep the N-of-1 participants disjoint from the data used to fit the pseudo-outcome model and to learn the surrogate maps, we use $K = 20$ rounds of cross-fitting. We partition the full cohort into $K = 20$ folds, and assign the first 9 folds to map learning, the next 9 folds to pseudo-outcome fitting, and the last 2 folds to the N-of-1 participants. We then rotate the selection 20 times. After the selection, each cohort is filtered according to its purpose: for the learning datasets, only subjects with outcome data are retained. For the N-of-1 cohort, only subjects with data on multiple treatments are retained. All hyperparameters for learning maps and the pseudo-label are selected within their respective cohorts D^{obs} and D^{tr}. As alluded in Appendix A.9.2, in every fold, treatment labels for a window are redrawn, such that one subject can contribute one window with multiple treatment labels.

Aggregation of results. In fold k, we compute for each N-of-1 participant the pseudo-regret and the ranking error averaged over all their window pairs, as in Algorithm 1. We then average over the N-of-1 population $\mathcal{J}^{\text{eval}}$ to obtain $\hat{R}(f)$ and $\hat{L}(f)$. In the figures, we report the mean and 95% CIs or standard deviation of $\hat{R}_(f)$ across the $K = 20$ cross-fitting folds. Regret relative to VO2max is computed in each fold as $\hat{R}(f)/\hat{R}(f_{\text{VO}_2\text{max}})$ before averaging.

A.9.4 Surrogate maps

We fit all surrogate maps on D^{tr} using as input the context-adjusted, z-scored within-window surrogate change S^{adj} (Appendix A.9.2) and, possibly, the fixed covariates x_{fix}.

We consider three surrogate spaces: the six wearable biomarkers (metrics), the PPG embedding changes, and their union (full surrogate space). Learned maps are fit within the map-learning D^{tr} of each cross-fitting round k (Appendix A.9.3), and the hyperparameters are selected by 5-fold cross-validation within D^{tr}. Under our outcome model, the covariate terms cancel out in within-subject contrasts, and thus, the treatment ranking is fully specified by the S and A coefficients of the pseudo-outcome. We consider linear maps of form $f(S, X)$, and by the same argument, only their S coefficients decide the maps' regret.

First, we note that each of the biomarkers S_i (RHR, HRV, VO2max, walking heart rate, 6min walk, walking speed) can be evaluated as a fixed surrogate map. For RHR and walking HR, lower values are preferred for health. We thus flip the sign of changes in RHR and walking HR before evaluating.

As our main class of surrogate maps, we fit unadjusted and adjusted Cox models. For the unadjusted map, we fit an L_2-penalized Cox model for the outcome with linear predictor $\gamma^\top s$ on D^{tr} (we do not use the covariates and thus do not adjust for their introduced confounding). For the adjusted map, the fitted Cox model additionally includes the covariates, $\gamma^\top s + \phi^\top x_{\text{fix}}$. This map removes the confounding from S coefficients introduced by the observed confounders x_{fix} (age, baseline health etc.) In both cases, we take as the map value the negative log-hazard of the surrogates, $f(s, x_{\text{fix}}) = -\hat{\gamma}^\top s$. Depending on whether or not the covariates are observed at deployment, the two maps can be seen as survival versions of the surrogate index (19). If x_{fix} are observed at learning-time only (since study participants normally contribute more data than users at deployment), then the adjusted map can be seen as an approximation of the causal surrogate map (cf. Appendix A.7.3). If x_{fix} are observed at both learning and deployment, the adjusted map corresponds to the surrogate index.

Next, we fit the control-function map with the aim to estimate the causal surrogate map (20) even in presence of unobserved confounding. We use the treatment labels as instruments for the surrogate change. Under full mediation (Assumption 3), the treatments affect the outcome only through the surrogate, and we can apply instrumental-variable methods (Appendix A.7.3). We use a two-stage control function approach (Newey et al., 1999; Imbens & Newey, 2009). In the first stage, we regress S on A, x_{fix} and compute the residual $\hat{v}$. In the second stage, we fit an L_2-penalized Cox model with linear predictor $\gamma^\top s + \phi^\top x_{\text{fix}} + \rho^\top \hat{v}$. The *control-function surrogate map* is then $f(s) = -\hat{\gamma}^\top s$.

Further, we attempt to learn the regret-minimizing surrogate map of Appendix A.7.1 by splitting D^{tr} into an observational component $\{(A_i, S_i, X_i, Y_i)\}_i$ with outcome data, and a collection of N-of-1 subjects. We then fit a training-time pseudo-outcome model on the observational component,

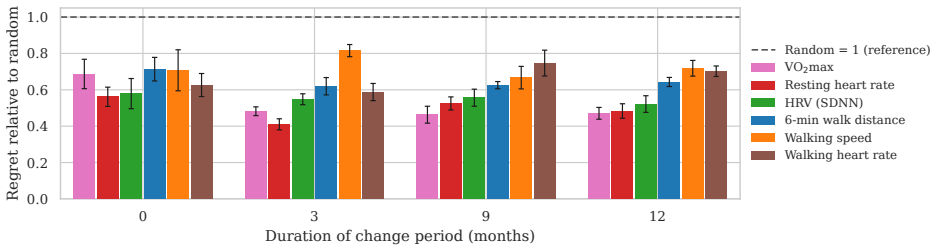


Figure 9: Average individual-level regret of the fixed wearable-metric surrogates, relative to the random map (dashed line at 1), as the change-window length increases. Values below 1 indicate lower regret than chance. Each bar is the mean over 20 cross-fitting folds; error bars are 95% confidence intervals.

impute pseudo-regret on N-of-1 subjects and minimize the resulting logistic loss as described in Appendix A.7.1.

Finally, as a reference, we include the random map which randomly assigns the best score to one of two treatment windows with probability $1/2$.

A.9.5 Additional results

In Figure 9, we additionally display a comparison of fixed wearable markers as individual-level surrogates across change windows of different durations. We divide the regret values by the regret of the random map, thus, the relative regret of 1.0 corresponds to random guessing. We observe that VO2max and resting heart rate are consistently the best surrogates, and surrogacy of VO2max improves as the change window gets longer.

A.10 Deferred proofs

A.10.1 Proof of Theorem 1

We fix the N-of-1 subject j and write x^j for its covariates and u^j for its latent type. Write

$$m(a, s, x, u) := \mathbb{E}_{\epsilon_Y}\left[f_0(a, s, x, u, \epsilon_Y)\right],$$

which is well defined by the standing integrability assumption of Section 2.

Since X^j and U^j are $\mathbb{P}^j$-a.s. constant, the potential surrogate outcome $S^j(a) = g_0(a, x^j, u^j)$ is $\mathbb{P}^j$-a.s. constant for every $a \in \mathcal{A}^j$, and consistency of the subject-level SCM $\mathcal{C}^j$ gives $S_i^j = S^j(A_i^j)$. Since ϵ_Y is exogenous, we have

$$\mu^j(a) = \mathbb{E}^j\left[Y^j(a)\right] = \mathbb{E}_{\epsilon_Y}\left[f_0\big(a, S^j(a), x^j, u^j, \epsilon_Y\big)\right] = m\big(a, S^j(a), x^j, u^j\big), \qquad a \in \mathcal{A}^j. \tag{29}$$

In the population study $\mathcal{C}^{\text{obs}}$ we have $S = g_0(A, X, U)$, so by injectivity of $g_0(a, x, \cdot)$ on $\operatorname{supp} \mathbb{P}^{\text{obs}}_{A,X,S}$ (Assumption 2), it holds that

$$U = g_0^{-1}(A, X, S) \qquad \mathbb{P}^{\text{obs}}\text{-a.s.}, \tag{30}$$

that is, the latent type is a function of (A, X, S). We thus obtain

$$\begin{aligned}\mathbb{E}^{\text{obs}}[Y \mid A, X, S] &= \mathbb{E}^{\text{obs}}[f_0(A, S, X, U, \epsilon_Y) \mid A, X, S] \\ &= \mathbb{E}^{\text{obs}}\big[f_0\big(A, S, X, g_0^{-1}(A, X, S), \epsilon_Y\big) \mid A, X, S\big] = m\big(A, S, X, g_0^{-1}(A, X, S)\big)\end{aligned}$$

$\mathbb{P}^{\text{obs}}$-a.s. By Assumption 1, $\big(a, x^j, S^j(a)\big) \in \operatorname{supp} \mathbb{P}^{\text{obs}}_{A,X,S}$ for every $a \in \mathcal{A}^j$, so the pseudo-outcome (5) is well defined and

$$\tilde{\mu}^j(a) = m\big(a, S^j(a), x^j, g_0^{-1}\big(a, x^j, S^j(a)\big)\big), \qquad a \in \mathcal{A}^j. \tag{31}$$

For the subject j, invertibility in Assumption 2 gives $g_0^{-1}\big(a, x^j, S^j(a)\big) = u^j$ for every $a \in \mathcal{A}^j$. Substituting into (31) and using (29), we obtain

$$\tilde{\mu}^j(a) = m\big(a, S^j(a), x^j, u^j\big) = \mu^j(a), \qquad a \in \mathcal{A}^j,$$

and thus equality of outcome labels and pseudo-labels.

Finally, let f be a valid surrogate map and $B \subseteq \mathcal{A}^j$. The induced policy $\pi_f(j; B)$ depends on the subject's surrogate values $\mathbb{E}^j[f(S^j(a), X^j)]$ only. Thus, we have $r^j(f; B) = \tilde{r}^j(f; B)$ for every $B \subseteq \mathcal{A}^j$, and averaging over pairs $\{a, a'\} \subseteq \mathcal{A}^j$ gives $r^j(f) = \tilde{r}^j(f)$. Similarly, we obtain equality of the ranking errors $\ell^j(f) = \tilde{\ell}^j(f)$.

A.10.2 Proof of Theorem 2

Throughout, $\bar{m}$ is as in Assumption 3, so that

$$\mu^j(a) \;=\; \bar{m}\big(S^j(a), X^j, U^j\big) \qquad \text{for every } a \in \mathcal{A}^j \subseteq \mathcal{A}^{\text{obs}} \cup \mathcal{A}^{\text{tst}}. \tag{32}$$

Step 1: Regret is a function of surrogate responses. For $x \in \mathcal{X}$, $u \in \mathcal{U}$ and surrogate values $s, s' \in \mathcal{S}$, define

$$\rho_f(x, u, s, s') := \max_{t \in \{s, s'\}} \bar{m}(t, x, u) \;-\; \min_{t \in \Pi_f(x; s, s')} \bar{m}(t, x, u), \qquad \Pi_f(x; s, s') := \operatorname*{argmax}_{t \in \{s, s'\}} f(t, x). \tag{33}$$

That is, $\rho_f(x, u, s, s')$ is the regret of a subject with covariates x and latent type u who uses f to choose between two treatments with surrogate responses s and s'. It is non-negative, measurable, and, importantly, does not depend on the treatments. Now fix a subject j and a pair $\{a, a'\} \subseteq \mathcal{A}^j$. By (32), $\mu^j(a) = \bar{m}(S^j(a), X^j, U^j)$, and thus $\pi_f(j; \{a, a'\})$ contains exactly the treatments whose surrogate responses lie in $\Pi_f(X^j; S^j(a), S^j(a'))$. The definition of the regret (2) therefore gives

$$r^j\big(f; \{a, a'\}\big) = \rho_f\big(X^j, U^j, S^j(a), S^j(a')\big). \tag{34}$$

Step 2: Population regret of f as an integral of ρ_f. Consider a collection of N-of-1 subjects $\mathcal{J}$ and its corresponding surrogate contrast distribution $\mathbb{Q}$. Let $\{A, A'\}$ denote a pair drawn uniformly among distinct treatment pairs in $\mathcal{A}^j$. First, for the total individual-level regret of subject $j \in \mathcal{J}$, we obtain:

$$r^j(f) \;=\; \binom{|\mathcal{A}^j|}{2}^{-1} \sum_{\{a,a'\}\subseteq\mathcal{A}^j} \rho_f\big(X^j, U^j, S^j(a), S^j(a')\big) \;=\; \mathbb{E}\Big[\,\rho_f\big(X^j, U^j, S^j(A), S^j(A')\big)\,\Big|\, j\,\Big].$$

Finally, we take the expectation over the N-of-1 subjects $j \in \mathcal{J}$, use the tower property over j and the definition of the surrogate contrast distribution to obtain

$$\mathbb{E}_{j\sim\mathrm{Unif}(\mathcal{J})}\left[r^j(f)\right] \;=\; \mathbb{E}_{\mathbb{Q}}\left[\rho_f\right]. \tag{35}$$

Applying (35) to the evaluation and test populations yields the following regret identities:

$$R^{\mathrm{eval}}(f) = \mathbb{E}_{\mathbb{Q}^{\mathrm{eval}}}\left[\rho_f\right], \qquad R^{\mathrm{tst}}(f) = \mathbb{E}_{\mathbb{Q}^{\mathrm{tst}}}\left[\rho_f\right]. \tag{36}$$

In particular, Steps 1 and 2 allowed us to express the evaluation-time and test-time regret of a surrogate map f as an expectation of the same function ρ_f under two different measures ($d\mathbb{Q}^{\mathrm{eval}}$ and $d\mathbb{Q}^{\mathrm{tst}}$). The remaining steps are to relate the evaluation-time regret to its corresponding pseudo-regret via Theorem 1 and to change measure.

Step 3: From evaluation-time pseudo-regret to regret. Assumptions 1 and 2 hold for $\mathbb{P}^{\mathrm{obs}}$ and every evaluation-time N-of-1 subject $j \in \mathcal{J}^{\mathrm{eval}}$, so Theorem 1 gives $\tilde{\mu}^j(a) = \mu^j(a)$ for every $a \in \mathcal{A}^j$ and $\tilde{r}^j(f) = r^j(f)$. Averaging over the evaluation subjects and using (36), we obtain

$$\tilde{R}^{\mathrm{eval}}(f) \;=\; R^{\mathrm{eval}}(f) \;=\; \mathbb{E}_{\mathbb{Q}^{\mathrm{eval}}}\left[\rho_f\right]. \tag{37}$$

Step 4: Change of measure. First, assume $\tilde{R}^{\mathrm{eval}}(f) = 0$. Using (37), non-negativity of ρ_f, and $\mathbb{Q}^{\mathrm{tst}} \ll \mathbb{Q}^{\mathrm{eval}}$, we obtain that it must also hold that $\mathbb{E}_{\mathbb{Q}^{\mathrm{tst}}}[\rho_f] = 0$, and thus $R^{\mathrm{tst}}(f) = \mathbb{E}_{\mathbb{Q}^{\mathrm{tst}}}[\rho_f] = 0$.

Now assume that in addition, the corresponding density $q := d\mathbb{Q}^{\mathrm{tst}}/d\mathbb{Q}^{\mathrm{eval}}$ is bounded by $0 \le B < \infty$. Then non-negativity of ρ_f gives

$$R^{\mathrm{tst}}(f) \;=\; \int \rho_f \,\frac{\mathrm{d}\mathbb{Q}^{\mathrm{tst}}}{\mathrm{d}\mathbb{Q}^{\mathrm{eval}}}\,\mathrm{d}\mathbb{Q}^{\mathrm{eval}} \;\le\; B\int \rho_f \,\mathrm{d}\mathbb{Q}^{\mathrm{eval}} \;=\; B\,R^{\mathrm{eval}}(f) \;=\; B\,\tilde{R}^{\mathrm{eval}}(f),$$

by (37).