

Certified Decisions

Isaiah Andrews and Jiafeng Chen

ABSTRACT. Hypothesis tests and confidence sets are ubiquitous in empirical research, yet their connection to subsequent decision-making is often unclear. We view statistical inference as pairing recommended decisions with guarantees of their quality—or *certificates*—and develop a theory of *certified decisions*. Specifically, we attach P-certificates—upper bounds on loss that hold with probability at least $1 - \alpha$ —to recommended actions. We prove that it is without loss to limit attention to P-certificates generated by minimax decisions over confidence sets, or what [Manski \(2021\)](#) terms “as-if decisions with a set estimate.” We further show that such certificates allow safe adoption decisions for ambiguity-averse downstream decision-makers. A parallel argument applies to E-certified decisions obtained from e-values in settings with unbounded loss. We illustrate this view of statistical inference through an application to economic mobility and one to AI model routing.

Date: This version: October 23, 2026; first version February 25, 2025. Andrews: Department of Economics, Massachusetts Institute of Technology, iandrews@mit.edu; Chen: Department of Economics, Stanford University, jiafeng@stanford.edu. We thank Yeon-Koo Che, Roberto Corrao, José Ignacio Cuesta, Guido Imbens, Lihua Lei, José Montiel Olea, Aaron Roth, Jonathan Roth, Suproteem Sarkar, Chris Walker, Davide Viviano, and seminar and workshop participants at NYU, Columbia, University of Pennsylvania, Stanford, and SEA 2025 for useful discussions. We thank Giacomo Fraccaroli for superb research assistance and valuable insights. *AI Disclosure:* AI tools (GPT, Claude) were used throughout the research process, including writing code and drafting proofs. The paper itself is entirely written by the authors, and we take full responsibility for its content.

1. Introduction

Inferential tools, such as hypothesis tests and confidence intervals, are ubiquitous in empirical practice. At the same time, many empirical papers aim, directly or indirectly, to improve decision-making by potential consumers of research, such as policymakers, firms, or households. As highlighted by [Manski \(2021\)](#), the connection between inference and decision-making is unclear: outside of special cases, optimal decisions in the tradition of [Wald \(1949\)](#) rarely depend on the data through the results of conventional inference procedures.¹ There is thus an apparent tension between the goal of informing decision-making and conventional, inference-focused, methods for empirical research.

Despite this tension—or perhaps because of it—some procedures in the literature bridge inference and decision-making. In particular, when researchers report confidence sets, one natural recipe for decisions is to proceed *as if* the confidence set enumerates all plausible parameter values, and choose actions that minimize the worst-case loss over the confidence set. That is, for $\mathcal{A}$ an action space, $L(a, \theta)$ a loss function, and $\hat{\Theta}(Y) \subset \Theta$ a data-dependent confidence set with $P\{\theta \in \hat{\Theta}(Y)\} \geq 1 - \alpha$, one may choose actions

$$\delta(Y) \in \arg \min_{a \in \mathcal{A}} \sup_{\theta \in \hat{\Theta}(Y)} L(a, \theta). \quad (1)$$

[Manski \(2021\)](#) terms decision rules of this form “as-if decisions with set estimates.”² Making decisions based on confidence sets is implicitly envisioned by Neyman, who writes ([Neyman, 1977](#), emphasis ours)

The problem of confidence intervals consists in determining two functions of the observables $[Y_1(\cdot) \leq Y_2(\cdot)]$ to be used in the following manner: Whenever the observable variables X assume some values $[x]$, we shall calculate the corresponding values of Y_1 and Y_2 , say $Y_1(x) < Y_2(x)$, and **then assert (or act on the assumption) that**

$$Y_1(x) \leq \theta \leq Y_2(x).$$

¹See, however, [Tetenov \(2016\)](#); [Bates et al. \(2024\)](#); [Viviano et al. \(2025\)](#); [Epanomeritakis and Viviano \(2026\)](#) for some interesting exceptions. Another exception is Bayesian inference, where both decisions and inferential conclusions are generated by the posterior; see [Erol and Vohra \(2026\)](#) for recent work that aims to reconcile frequentist confidence sets with Bayesian decision-making.

²Early instances of such decisions appear in [Henry and Onatski \(2012\)](#); [Goldfarb and Iyengar \(2003\)](#).

More recently, Ben-Michael *et al.* (2021); Chernozhukov *et al.* (2026) and Yamin (2026) select policies by minimizing the worst-case loss over a confidence set for a social welfare function.

Several recent papers have also considered the other direction, first making a statistical decision and then constructing confidence sets for the resulting loss or welfare. Kitagawa and Tetenov (2018) select a policy by maximizing an empirical welfare function, and in their supplementary materials discuss how one may construct a confidence set for the true welfare via simultaneous inference and projection (see also Ponomarev and Semenova, 2024). The recent literature studying inference on winners (Benjamini *et al.*, 2019; Andrews *et al.*, 2024; Zrnic and Fithian, 2025, 2024; Cheng *et al.*, 2025) generalizes this approach, considering selection of the “best” of several options based on noisy estimates of loss or welfare, and then constructing bounds for the true loss associated with the estimated “best” choice. That is, for $\delta(Y)$ a decision rule chosen by optimizing some empirical criterion, this literature (in the one-sided case) constructs a loss bound $R(Y)$ with

$$P\{L(\delta(Y), \theta) \leq R(Y)\} \geq 1 - \alpha. \quad (2)$$

These exercises—choosing actions based on confidence sets and constructing confidence bounds for the losses of recommended actions—can be unified by viewing statistical inference as a means of *certifying* the performance of proposed decisions. Imagine a data analyst who provides a decision-maker with a recommended action and an estimated loss bound $(\delta(Y), R(Y))$. If this pair satisfies (2), we call $R(Y)$ a P-certificate for $L(\delta(Y), \theta)$, and $(\delta(Y), R(Y))$ a *P-certified decision*. The loss bounds $R(Y)$ derived in the literature studying inference on winners are thus P-certificates by construction. For as-if decision-making, on the other hand, note that if we define

$$R(Y) = \sup_{\theta \in \hat{\Theta}(Y)} L(\delta(Y), \theta) \quad (3)$$

then $\theta \in \hat{\Theta}(Y)$ implies $R(Y) \geq L(\delta(Y), \theta)$, so

$$P\{L(\delta(Y), \theta) \leq R(Y)\} \geq P(\theta \in \hat{\Theta}(Y)) \geq 1 - \alpha.$$

Hence, both procedures discussed above generate P-certificates. This paper develops—and advocates for—a view in which the purpose of statistical inference is to provide guarantees for decisions.

We first study the optimal provision of P-certified decisions. We find that the class of as-if decisions à la (1) and (3) is *essentially complete* under any preference ordering

that favors tighter certificates.³ That is, any P-certified decision $(\tilde{\delta}, \tilde{R})$ can be weakly improved by *some as-if decision* (δ, R) corresponding to some confidence set, in the sense that $R(\cdot) \leq \tilde{R}(\cdot)$ almost surely and the loss bound $R(\cdot)$ is thus weakly tighter than the loss bound $\tilde{R}(\cdot)$. This establishes that as-if decisions are *universal*: all reasonable P-certified decisions can be cast as as-if optimization (1). To be sure, the result does not distinguish certain confidence sets from others. Nevertheless, it has some bite, and implies for instance that the studentized projection approach discussed in e.g. Andrews *et al.* (2024) is dominated by the approaches of Chernozhukov *et al.* (2026).

With more structure on the decision problem, we further show that some optimality results for inference translate to optimality for P-certificates. In particular, if the parameter is scalar and the loss function prefers high parameter values (e.g., implementing treatments with higher effects yields lower loss), then as-if decisions based on uniformly most accurate confidence lower bounds for θ yield optimal P-certificates.

Next, to connect P-certified decisions to standard decision-theoretic notions such as risk and expected utility, we study an ambiguity-averse decision-maker with limited capacity for data analysis who delegates that analysis to an analyst. The analyst provides the decision-maker with a P-certified decision. The decision-maker then needs to decide whether to adopt the action recommended by the analyst, or to default to a safe action with known (but potentially large) loss. This decision-maker believes that the analyst’s certificate $R(Y)$ contains information about the loss of the recommended action, but has limited knowledge of the data-generating process and wishes to protect themselves against a large increase in risk uniformly over a potentially large class of data-generating processes.

P-certificates are useful to such a decision-maker. Specifically, when losses lie in $[0, 1]$ and the default action costs $C \in [0, 1)$, any rule that adopts the recommended action $\delta(Y)$ only when $R(Y) \leq C$ yields expected loss no greater than $C + \alpha(1 - C)$, meaning that the decision-maker faces a risk increase of at most $\alpha(1 - C)$. Moreover, we characterize optimal adoption decisions for decision-makers who balance subjective and worst-case risk. We find that an optimal adoption rule either never adopts or interpolates between the threshold rule $\mathbb{1}(R(Y) \leq C)$ and the Bayes rule that adopts whenever subjective expected loss is below C .

³This result echoes—though it is developed independently of—a recent result by Kiyani *et al.* (2025), in the context of conformal prediction.

We then illustrate how this perspective—viewing inference as providing certificates for decisions—helps interpret empirical findings and construct recommendations in two applications. First, we revisit Creating Moves to Opportunity (CMTO, [Bergman et al., 2024](#)), an experiment where housing voucher holders are recommended putatively high-economic-mobility Census tracts. Since these Census tracts are selected based on noisy estimates of their quality, there is concern about whether the voucher holders were indeed induced to move to higher-quality places—especially with additional ambiguity over their previous locations and their location choices.

[Andrews et al. \(2024\)](#) and [Mogstad et al. \(2024\)](#) report inference results with starkly different qualitative conclusions on this question: [Andrews et al. \(2024\)](#) report that targeted tracts have significantly higher average mobility than non-targeted tracts, while [Mogstad et al. \(2024\)](#) report that the tract-level mobility ranking is highly uncertain. We show that their inference results are relevant for different decision problems, with different ambiguity sets over movers’ relocation patterns, and introduce a family of decision problems that interpolates between them. We then construct certified decisions for this spectrum of decision problems. This exercise quantifies how much robustness to relocation patterns can be accommodated while still certifying positive mobility gains.

Our second application considers routing queries to large language models (LLMs) using noisy estimates of model quality. We view different inference-on-winner-type procedures ([Zrnic and Fithian, 2025, 2024](#); [Andrews et al., 2024](#); [Chernozhukov et al., 2026](#)) as providing both recommended actions and resulting certificates. The tightness of the resulting certificate can be thought of as a notion of decision-relevant power for a given procedure, which we thus compare across procedures. We find cases in which recommending a model with a lower estimated payoff but greater precision yields a more favorable guarantee than certifying the model with the highest estimated payoff. This illustrates how choosing actions in a certificate-aware way—as [Chernozhukov et al. \(2026\)](#) advocate—can be beneficial.

Finally, we extend our analysis to certified decisions that correspond to an alternative approach to frequentist inference. A rapidly growing literature proposes *e-values* as alternatives to p-values and confidence sets (see [Ramdas and Wang, 2024](#), for a review). We show that e-variables generate as-if decisions—called E-posterior decisions by [Grünwald \(2023\)](#)—that provide an *E-certificate* of the form $\mathbb{E}_P[L(\delta(Y), \theta)/R(Y)] \leq 1$. Similarly, we show that the class of such as-if decisions is essentially complete among E-certified decisions: for any E-certified $(\tilde{\delta}, \tilde{R})$, there

exists a collection of e-variables and a corresponding E-posterior decision $(\delta(\cdot), R(\cdot))$ such that $R(\cdot) \leq \tilde{R}(\cdot)$ almost surely. Such decisions imply risk bounds for downstream decision-makers even in contexts with unbounded loss functions. We illustrate these certificates in the LLM routing application.

This paper is most similar to recent work by Kiyani *et al.* (2025), who study decision-making in the context of conformal inference. In work preceding ours, Kiyani *et al.* (2025) show that minimizing the expected value of a $1 - \alpha$ upper bound for loss is equivalent to choosing a conformal prediction set with coverage $1 - \alpha$ to minimize expected minimax loss (see (7)). The key to both arguments is the observation that a probabilistic upper bound on loss can be inverted to form a confidence set.

Section 2 defines P-certified decisions, shows how confidence sets generate them, studies their optimal provision, and analyzes their value to downstream ambiguity-averse decision-makers. Section 3 illustrates our P-certificates in application, while Section 4 discusses the extension to E-certificates.

2. P-certified decisions

Consider a statistical decision problem characterized by a loss function $L(a, \theta) : \mathcal{A} \times \Theta \rightarrow \mathbb{R}$ for an action $a \in \mathcal{A}$ and an unknown parameter $\theta \in \Theta$. Data $Y \in \mathcal{Y}$ is drawn from some distribution P . To model uncertainty, following Andrews *et al.* (2025), we assume that Nature may choose any (θ, P) pairs from a set $\mathcal{M}$. For instance, if (θ, P) satisfy a statistical model indexed by θ , then $\mathcal{M} = \{(\theta, P_\theta) : \theta \in \Theta\}$. More generally, we allow that θ may fail to fully describe P , for instance due to nuisance parameters or model incompleteness. This formulation also does not impose point-identification: we may have $\{(\theta_1, P), (\theta_2, P)\} \subset \mathcal{M}$ for $\theta_1 \neq \theta_2$. Throughout, we refer to L as a loss function, but it is possible to choose $L(a, \theta) = \tilde{L}(a, \theta) - \inf_{a' \in \mathcal{A}} \tilde{L}(a', \theta)$ so that L represents regret for some primitive loss function $\tilde{L}$ instead.

We start by defining *P-certified decisions* and establishing some basic properties.

Definition 1 (P-certified decisions). For some $\alpha \in (0, 1)$, the pair $(\delta(\cdot), R(\cdot)) : \mathcal{Y} \rightarrow \mathcal{A} \times \overline{\mathbb{R}}$ is called a $1 - \alpha$ *P-certified decision* if, for every $(\theta, P) \in \mathcal{M}$,

$$P(L(\delta(Y), \theta) \leq R(Y)) \geq 1 - \alpha. \quad (4)$$

When this holds, $R(Y)$ is called a $(1 - \alpha)$ *P-certificate* for $\delta(Y)$.

The guarantee (4) is familiar—demanding that $R(Y)$ be a $(1 - \alpha)$ -confidence upper bound for the random quantity $L(\delta(Y), \theta)$. This guarantee is a frequentist one: it holds ex ante over repeated draws of the data Y for fixed (θ, P) , uniformly

over the relevant econometric model $\mathcal{M}$. A realization of a low $R(Y)$ certifies that the corresponding action $\delta(Y)$ is attractive, in the sense that either its loss is low ($L(\delta(Y), \theta) \leq R(Y)$), or something improbable—specifically, an event with probability at most α —has occurred.

We next study procedures that (optimally) generate P-certified decisions. We then present a stylized delegation model in which P-certified decisions, generated by an analyst recommending a decision, allow subsequent risk bounds for an ambiguity-averse decision-maker who decides whether or not to accept analyst recommendations.

First, P-certified decisions are readily generated by as-if optimization with set estimates (Goldfarb and Iyengar, 2003; Henry and Onatski, 2012; Manski, 2021; Lim and Park, 2026).

Proposition 1. *Let $\hat{\Theta}(Y)$ be a confidence set: $P(\hat{\Theta}(Y) \ni \theta) \geq 1 - \alpha$ for all $(\theta, P) \in \mathcal{M}$. For any $\delta(Y)$, let $R(Y) = \sup_{\theta \in \hat{\Theta}(Y)} L(\delta(Y), \theta)$. Then (δ, R) is a $(1 - \alpha)$ P-certified decision. In particular, if the minimum $\delta^*(Y) := \arg \min_a \sup_{\theta \in \hat{\Theta}(Y)} L(a, \theta)$ exists, then the corresponding $R^*(Y) := \sup_{\theta \in \hat{\Theta}(Y)} L(\delta^*(Y), \theta) \leq \sup_{\theta \in \hat{\Theta}(Y)} L(\delta(Y), \theta)$ for any other δ .⁴*

Proof. Fix $(\theta, P) \in \mathcal{M}$. If $\theta \in \hat{\Theta}(Y)$ then $L(\delta(Y), \theta) \leq R(Y)$ by definition. Thus $1 - \alpha \leq P(\theta \in \hat{\Theta}(Y)) \leq P\{L(\delta(Y), \theta) \leq R(Y)\}$ and (δ, R) is a P-certified decision. For the second part, observe that $R^*(Y) = \min_a \sup_{\theta \in \hat{\Theta}(Y)} L(a, \theta) \leq \sup_{\theta \in \hat{\Theta}(Y)} L(\delta(Y), \theta)$. $\square$

Proposition 1 shows that the worst-case loss of any decision rule over a given confidence set is automatically a P-certificate of that decision rule. Correspondingly, by *optimizing* this worst-case loss over a confidence set—exactly as in as-if optimization—we obtain the tightest certificate among those constructed this way. In Section 2.1, we shall see that the converse is also essentially true: Any reasonable P-certified decision (δ, R) is (weakly) beaten by as-if optimization using *some* confidence set.

We illustrate these results with a running example.

Example 1 (Inference on winners). We consider an example based on the recent literature on inference on winners (Benjamini *et al.*, 2019; Andrews *et al.*, 2024; Zrnic and Fithian, 2025, 2024). Suppose we are interested in potentially implementing an action a from a finite set $\mathcal{A}$. For instance, we could be interested in whether to

⁴Throughout, the supremum over an empty set is equal to $-\infty$.

implement a program, as studied in [Bergman *et al.* \(2024\)](#), that encourages low-income households with children to move to high-opportunity neighborhoods. In this context, each $a \in \mathcal{A}$ could index a potential set of “recommended” neighborhoods.

Our goal may be to maximize some outcome, for instance the average household income rank in adulthood (relative to the adult income distribution) for children in the targeted households. Specifically, for $\theta(a) \in [0, 1]$ the (unknown) true outcome associated with selection a , we would like to choose actions a that maximize $\theta(a)$. To represent this preference, we can consider the loss function $L(a, \theta) = 1 - \theta(a)$.

Data (e.g. Census-tract level measures for economic opportunity together with a model for long-term outcomes as in [Bergman *et al.* 2024](#)) can be used to form an estimate $X(a)$ for the outcome associated with each possible action a , along with an associated standard error $\sigma(a)$, which for simplicity we treat as known. For $a \in \mathcal{A}$, suppose that $\frac{X(a) - \theta(a)}{\sigma(a)} \sim Z(a)$, for some random variables $\{Z(a)\}_{a \in \mathcal{A}}$ with a known distribution (e.g., multivariate Gaussian with known variance). In this setting, we can let $Y = (X(\cdot), \sigma(\cdot))$ collect the estimates and standard errors.

One reasonable decision rule is empirical welfare maximization, which takes $\delta(Y) = \arg \max_{a \in \mathcal{A}} X(a)$. This decision rule corresponds to as-if optimization against a joint confidence set for θ based on an unstudentized statistic $X(a) - \theta(a) = Z(a)\sigma(a)$:

$$\hat{\Theta}(Y) = \bigcap_{a \in \mathcal{A}} [X(a) - c_{1-\alpha}, \infty), \text{ where } P\left(\sup_{a \in \mathcal{A}} Z(a)\sigma(a) \leq c_{1-\alpha}\right) = 1 - \alpha \quad (5)$$

This confidence set corresponds to an unstudentized version of what [Andrews *et al.* \(2024\)](#) call the projection approach, which was suggested in the supplementary materials of [Kitagawa and Tetenov \(2018\)](#). Its corresponding certificate is

$$R(Y) = 1 - X(\delta(Y)) + c_{1-\alpha},$$

which inflates the estimated loss by $c_{1-\alpha}$. ■

2.1. Optimal provision of P-certificates. In settings where P-certificates are useful, it seems reasonable to prefer tighter bounds on the loss. That is, given the choice between level- $(1 - \alpha)$ P-certificates $R(Y)$ and $\tilde{R}(Y)$, if $R(Y) \leq \tilde{R}(Y)$ for all Y , it is natural to prefer R to $\tilde{R}$. More generally, if we consider any loss function $B(R, \theta)$ that is increasing in R for all θ , then $\mathbb{E}_P[B(R(Y), \theta)] \leq \mathbb{E}_P[B(\tilde{R}(Y), \theta)]$ for all (P, θ) , so R yields a lower expected loss than $\tilde{R}$.

When R and $\tilde{R}$ are ordered almost surely, or more generally in the sense of first-order stochastic dominance, we say R weakly dominates $\tilde{R}$.

Definition 2. We say that a $1 - \alpha$ P-certified decision (δ, R) *weakly dominates* another $1 - \alpha$ P-certified decision $(\tilde{\delta}, \tilde{R})$ if

$$P(R(Y) \geq r) \leq P(\tilde{R}(Y) \geq r)$$

for all r and all $(\theta, P) \in \mathcal{M}$.

We next show that the class of P-certified decisions from as-if optimization against a confidence set, à la [Proposition 1](#), is an *essentially complete* class with respect to dominance. That is, any P-certified decision rule $(\tilde{\delta}, \tilde{R})$ is weakly dominated by some decision rule (δ, R) formed by as-if optimization. As an implication, these decision rules are also essentially complete with respect to any risk function of the form $\mathbb{E}_P[B(R, \theta)]$, for any weakly increasing $B(\cdot, \theta)$.

To state this result, denote by $\mathcal{C}_{P, \alpha, \epsilon}$ the class of decision rules formed by as-if optimization against a confidence set $\hat{\Theta}(Y)$ in the following fashion: For $\epsilon > 0$, consider $(\delta(Y), R(Y))$ such that (i) $\hat{\Theta}$ is a $1 - \alpha$ confidence set under $\mathcal{M}$ and (ii)

$$\sup_{\theta \in \hat{\Theta}(Y)} L(\delta(Y), \theta) \leq \inf_{a \in \mathcal{A}} \sup_{\theta \in \hat{\Theta}(Y)} L(a, \theta) + \epsilon \quad \text{and} \quad R(Y) = \sup_{\theta \in \hat{\Theta}(Y)} L(\delta(Y), \theta) \quad (6)$$

The constant ϵ accounts for cases where the minimax loss over $\hat{\Theta}(Y)$ is not achieved by any element of $\mathcal{A}$; our result holds for any $\epsilon > 0$.

Theorem 1. Assume that $\inf_{a \in \mathcal{A}} L(a, \theta) > -\infty$ for every $\theta \in \Theta$. For any P-certified decision $(\tilde{\delta}, \tilde{R})$ and any $\epsilon > 0$, there exists some $(\delta, R) \in \mathcal{C}_{P, \alpha, \epsilon}$ that weakly dominates $(\tilde{\delta}, \tilde{R})$.

[Theorem 1](#) shows a sense in which it is without loss to limit attention to P-certificates derived from as-if optimization with confidence sets. Thus, not only is as-if optimization a convenient way to generate P-certified decisions, but there is, in a sense, no need to look beyond this recipe. [Theorem 1](#) follows from the simple observation that inverting the P-certification property yields a valid confidence set,

$$\tilde{\Theta}(Y) := \left\{ \theta : L(\tilde{\delta}(Y), \theta) \leq \tilde{R}(Y) \right\}, \quad P(\theta \in \tilde{\Theta}(Y)) = P(L(\tilde{\delta}(Y), \theta) \leq \tilde{R}(Y)) \geq 1 - \alpha.$$

As-if optimizing against this confidence set can only weakly improve the resulting certificate.

Remark 2 (Properties of as-if optimization). Since restricting to as-if optimization over confidence sets is, for the purpose of generating P-certified decisions, without loss, we highlight a few additional features of these decision rules. First, confidence sets allow for decision-makers with different decision problems—including ones unknown to

the analyst—to derive their own P-certificates by optimizing (6), making confidence sets attractive for communicating information to heterogeneous decision-makers. Second, in addition to generating P-certificates, as-if decisions with confidence sets are dynamically consistent in the sense axiomatized by [Lim and Park \(2026\)](#).

Lastly, we comment on set-identification of θ . The mechanics of as-if optimization, its guarantees of P-certification, and its essential completeness do not qualitatively depend on whether θ is point- or set-identified.⁵ This feature is distinct from some other problems considered in the partial identification literature: [Christensen *et al.* \(2026\)](#), for instance, argue that set-identification of payoffs meaningfully changes optimal decision rules in their setting. By contrast, so long as we consider decision problems for which θ —and not the identified set—is the decision-relevant parameter,⁶ the coverage property that matters for P-certification is coverage for θ as in [Imbens and Manski \(2004\)](#), as opposed to coverage for the identified set ([Chernozhukov *et al.*, \(2007\)](#)). ■

Remark 3 (Relation to [Kiyani *et al.* \(2025\)](#)). In the context of conformal prediction, recent work by [Kiyani *et al.* \(2025\)](#) (their Theorem 2.3) shows a result similar to [Theorem 1](#).⁷ To connect these results, let ϑ be a random prediction target and let H be a joint distribution over (ϑ, X) . In a prediction context, ϑ is the unknown true label of a unit and X is the observed features. A conformal prediction set $\hat{\Theta}(X)$ satisfies $H(\vartheta \in \hat{\Theta}(X)) \geq 1 - \alpha$, where the probability is taken over $(\vartheta, X) \sim H$. [Kiyani *et al.* \(2025\)](#) show that the following optimization problems are equivalent, with essentially the same argument as that underlying [Theorem 1](#):

$$\begin{aligned} & \min_{R(\cdot), \delta(\cdot)} \mathbb{E}_H[R(X)] \text{ subject to } H\{L(\delta(X), \vartheta) \leq R(X)\} \geq 1 - \alpha \\ & \min_{\hat{\Theta}(\cdot)} \mathbb{E}_H \left[\min_{a \in \mathcal{A}} \max_{\vartheta \in \hat{\Theta}(X)} L(a, \vartheta) \right] \text{ subject to } H\{\vartheta \in \hat{\Theta}(X)\} \geq 1 - \alpha. \end{aligned} \quad (7)$$

In contrast to [Kiyani *et al.* \(2025\)](#), we consider a parameter inference setting, where probabilities are solely taken over Y fixing θ , rather than jointly over (X, ϑ) . We also consider evaluations of the certificate $R(\cdot)$ beyond its expectation, using a weaker preference ordering ([Definition 2](#)). ■

⁵Of course, construction of confidence sets is often qualitatively different in point- versus set-identified regimes.

⁶One could instead define the loss function as relative to the identified set, with $L(a, \Theta_I(P)) := \sup_{\theta \in \Theta_I(P)} \tilde{L}(a, \theta)$ as in [Christensen *et al.* \(2026\)](#) and [Henry and Onatski \(2012\)](#). P-certification for such loss functions then requires confidence sets covering the identified set.

⁷We developed [Theorem 1](#) independently of their result.

Example 2 (Inference on winners, continued). **Theorem 1** establishes a strong sense in which as-if optimization is reasonable. The dominance ordering in **Definition 2** is demanding, however, and there are many (R, δ) , $(\tilde{R}, \tilde{\delta})$ pairs that cannot be ordered, in the sense that neither $R(Y)$ nor $\tilde{R}(Y)$ (weakly) stochastically dominates the other. Consequently, weak dominance in the sense of **Definition 2** provides only a partial order over the class of certified decisions, and does not generally provide clear recommendations for “optimal” procedures. This partial order is nevertheless sufficient to rule out some methods discussed in the literature.

To illustrate, let us return to **Example 1**. There, we select a recommended action via empirical welfare maximization, $\delta(Y) = \arg \max_{a \in \mathcal{A}} X(a)$. We may P-certify this decision rule using unstudentized projection as in (5), or we may use studentized projection instead:

$$\hat{\Theta}^*(Y) = \bigtimes_{a \in \mathcal{A}} [X(a) - c_{1-\alpha}^* \sigma(a), \infty), \text{ where } P\left(\max_a Z(a) \leq c_{1-\alpha}^*\right) = 1 - \alpha,$$

forming a P-certificate $R^*(Y) = \sup_{\theta \in \hat{\Theta}^*} L(\delta(Y), \theta) = 1 - X(\delta(Y)) + c_{1-\alpha}^* \sigma(\delta(Y))$ (with R^* clipped to $[0, 1]$). Studentization ensures that the width of the interval reflects the standard error for the recommended policy $\delta(Y)$, and further ensures that e.g. the inclusion of a single, very noisily estimated option a in the choice set does not greatly influence the critical value. For this reason, the theoretical discussion of projection in [Andrews et al. \(2024\)](#) focuses primarily on the studentized approach.

As discussed in **Example 1**, $R(Y)$ can be understood via as-if optimization. By contrast, (δ, R^*) is less naturally understood via as-if optimization. If the analyst conducts as-if optimization over $\hat{\Theta}^*(Y)$, their recommended decision is $\tilde{\delta}(Y) = \arg \max_{a \in \mathcal{A}} \{X(a) - \sigma(a)c_{1-\alpha}^*\}$, which is a special case of the risk-aware empirical maximization approach proposed by [Chernozhukov et al. \(2026\)](#) and can differ from $\delta(Y)$ when $\sigma(\cdot)$ is non-constant. The corresponding risk bound $\tilde{R}(Y) = 1 - \max_{a \in \mathcal{A}} \{X(a) - \sigma(a)c_{1-\alpha}^*\}$ is no larger than $R^*(Y)$ in every realization of Y , so $(\tilde{\delta}, \tilde{R})$ weakly dominates (δ, R^*) .⁸ ■

2.1.1. Optimality under ordered θ . The non-uniqueness of optimal certified decisions, in the sense of weak dominance, is intuitive: **Theorem 1** shows that P-certificates are closely linked to confidence sets, and in most problems the class of admissible

⁸Formally, however, if $\Theta = [0, 1]^{\mathcal{A}}$ then (δ, R^*) still belongs to the essentially complete class we study, since it results from as-if optimization against the confidence set $\hat{\Theta}^{**}(Y) = \{\theta \in \Theta : L(\delta(Y), \theta) \leq 1 - (X(\delta(Y)) - \sigma(\delta(Y))c_{1-\alpha}^*)\}$. Note, however, that $\hat{\Theta}^{**}(Y)$ is a superset of $\hat{\Theta}^*(Y)$, which helps explain the superior performance of $(\tilde{\delta}, \tilde{R})$ relative to (δ, R^*) .

(i.e. undominated) confidence sets is large. If we narrow attention to settings where optimal (uniformly most accurate) confidence sets exist, one might hope that optimal P-certified decisions will exist as well and be found by as-if optimization. The next theorem shows that hope is borne out, provided the loss is monotone in the parameter and we limit our comparison of stochastic dominance to risk values exceeding $\inf_{a \in \mathcal{A}} L(a, \theta)$.⁹

To state the result, we recall the definition of uniformly most accurate confidence bounds:

Definition 3 (Lehmann and Romano (2024), pp. 79–80). Suppose $\Theta \subseteq \mathbb{R}$. A random variable $\hat{\theta}(Y)$ is a $1 - \alpha$ *confidence lower bound* if $P(\hat{\theta}(Y) \leq \theta) \geq 1 - \alpha$ for all $(\theta, P) \in \mathcal{M}$. Furthermore, it is *uniformly most accurate* if for any $1 - \alpha$ confidence lower bound $\tilde{\theta}(Y)$, any $t > 0$, and any $(\theta, P) \in \mathcal{M}$,

$$P\{\hat{\theta}(Y) \leq \theta - t\} \leq P\{\tilde{\theta}(Y) \leq \theta - t\}.$$

Theorem 4. Suppose $\Theta \subseteq \mathbb{R}$ is compact, $L(a, \theta)$ is weakly decreasing and right-continuous in θ for each a , and $R(\theta) = \inf_{a \in \mathcal{A}} L(a, \theta)$ is achieved by some $a \in \mathcal{A}$ for every $\theta \in \Theta$. Let $\hat{\theta}(Y)$ be a $1 - \alpha$ uniformly most accurate confidence lower bound, assumed to exist. Let (δ, R) be a certified decision that as-if optimizes against $[\hat{\theta}(Y), \infty) \cap \Theta$:

$$R(Y) \equiv L(\delta(Y), \hat{\theta}(Y)) = \inf_{a \in \mathcal{A}} L(a, \hat{\theta}(Y)).$$

Then for any other $1 - \alpha$ P-certified $(\tilde{\delta}, \tilde{R})$ and any $r > R(\theta)$,

$$P(R(Y) \geq r) \leq P(\tilde{R}(Y) \geq r).$$

Example 3 (Treatment choice). **Theorem 4** does not apply to the inference on winners problem (**Example 1**) since the parameter of interest θ is vector-valued, and there does not in general exist a uniformly most accurate confidence bound. Consider instead a setting where the decision-maker considers a single treatment, for instance enrollment in a job training program, and must decide whether to treat an entire population or not, $a \in \{0, 1\}$. Suppose the loss function is $L(a, \theta) = -a\theta$ and the analyst observes a normally distributed estimate $Y \sim \mathcal{N}(\theta, \sigma^2)$ for known σ^2 . Then the optimal certified decision in the sense of **Theorem 4** is $\delta(Y) = \mathbb{1}(Y > \sigma\Phi^{-1}(1 - \alpha))$ with certificate $R(Y) = -(Y - \sigma\Phi^{-1}(1 - \alpha))_+$ —i.e., treating the population when the

⁹Absent the latter restriction, it is impossible to dominate as-if optimization against the trivial confidence set which takes $\hat{\Theta}(Y)$ empty with probability α , and equal to Θ with probability $1 - \alpha$.

uniformly most accurate confidence lower bound exceeds zero, or equivalently when a *one-sided* test against $\theta \leq 0$ rejects.

On the other hand, if we instead consider *regret* our objective, $L(a, \theta) = -a\theta - \min(0, -\theta)$, then this loss function is no longer monotone in θ and [Theorem 4](#) no longer applies. Interestingly, as-if optimization using the standard two-sided confidence set $[Y - \Phi^{-1}(1 - \alpha/2)\sigma, Y + \Phi^{-1}(1 - \alpha/2)\sigma]$ yields the empirical success rule $\delta(Y) = \mathbb{1}(Y > 0)$, regardless of the level α . Its corresponding regret certificate is then $R(Y) = \max(0, \Phi^{-1}(1 - \alpha/2)\sigma - |Y|)$ —which is zero whenever $\theta = 0$ is rejected at level α . ■

2.2. Benefits to a downstream ambiguity-averse decision-maker. The guarantee (4) is arguably intuitive and may be imposed by fiat as a prerequisite for decisions—much like how confidence sets have become standard implementations of statistical inference. Indeed, recent work by [Galeotti et al. \(2024, Definition 1\)](#) imposes exactly such a notion. Yet P-certification is not immediately connected to standard decision-theoretic targets such as risk or expected utility. Towards establishing such a connection, this subsection analyzes how P-certified decisions may provide value to a downstream, ambiguity-averse decision-maker who delegates data analysis to an analyst while hoping to minimize expected loss (risk). The analyst provides the decision-maker with a P-certified decision, and the decision-maker must choose whether to accept the recommendation or to instead default to some safe action.

For this subsection, we assume that the loss function is bounded, and without further loss of generality assume $L(a, \theta) \in [0, 1]$ (correspondingly, we may assume certificates are also bounded $R(Y) \in [0, 1]$). The decision-maker has access to a default action $a_0 \notin \mathcal{A}$, which yields a loss $C \in [0, 1]$ which is known to the decision-maker and does not depend on θ . Our decision-maker decides whether to adopt the recommended action $\delta(Y)$ —incurring an uncertain loss of $L(\delta(Y), \theta)$ —or to implement the default action a_0 and incur loss C . We model the decision-maker as unfamiliar with the details of the relevant statistical decision problem. Correspondingly, the decision-maker would like to delegate data analysis to an analyst, while requiring worst-case assurances for the performance of the resulting decision.

Let $Q \in \{0, 1\}$ be an indicator for whether the decision-maker adopts the analyst’s recommendation. A simple adoption rule is $Q := \mathbb{1}(R(Y) \leq C)$ —adopting whenever the reported certificate is lower than the default cost. The corresponding loss is $QL(\delta(Y), \theta) + (1 - Q)C$. For such adoption rules, P-certified decisions provide a worst-case risk bound: the decision-maker incurs no more than $u\alpha(1 - C)$ excess risk if they never adopt when $R > C$ and adopt with probability at most u when $R \leq C$.

Proposition 2. For $u \in [0, 1]$ and any $(\theta, P) \in \mathcal{M}$, any adoption decision Q with $\mathbb{E}[Q \mid Y] \leq u \mathbb{1}(R(Y) \leq C)$ has bounded risk

$$\mathbb{E}_P[QL(\delta(Y), \theta) + (1 - Q)C] \leq C + u\alpha(1 - C).$$

Proposition 2 shows one way in which P-certified decisions provide value to ambiguity-averse decision makers: they ensure that the decision-maker does not have to tolerate much excess risk so long as they adopt $\delta(Y)$ only when $R(Y)$ suggests that this is preferable to the default action.¹⁰ To further control the excess risk, the decision-maker may either adopt less, in the sense of choosing a smaller u , or ask that the analyst provide recommendations with higher confidence, requesting a larger $1 - \alpha$. Since Proposition 2 applies to all P-certified decisions, including those derived from trivial confidence sets, naturally the worst-case bound exceeds C .

Of course, a decision-maker solely concerned with worst-case risk and facing sufficient ambiguity would always default to the safe action. We next consider optimal adoption for a particular form of ambiguity-averse preferences.

Consider a decision-maker who is ambiguity averse and uncertain of the precise loss function, but believes the P-certification property. The decision maker observes the recommended action and the certificate, $(\delta(Y), R(Y))$, drawn from a distribution G . The true value of G is the law of (δ, R) induced by the true distribution P and the analyst's choices, which in turn depend on the loss L . Let $\ell : \mathcal{A} \rightarrow [0, 1]$ denote a generic loss function, which may lie outside the class $\{L(\cdot, \theta) : \theta \in \Theta\}$ induced by the true L . Our next result considers a decision-maker who entertains *all* pairs $(\ell(\cdot), G)$ satisfying P-certification:

$$\mathcal{M}_{\text{DM}}^{(P, \alpha)} := \{(\ell(\cdot), G) : G(\ell(\delta) \leq R) \geq 1 - \alpha, G(\mathcal{A} \times [0, 1]) = 1\}.$$

¹⁰Let $\delta_Q(Y)$ denote the recommendation when $Q = 1$ and the default action otherwise. Since $P\{L(\delta_Q(Y), \theta) > QR(Y) + (1 - Q)C\} \leq \alpha$ and $L(\delta_Q(Y), \theta) \leq 1$, we also obtain the following P -specific bound, which is markedly better if $R(Y)$ can be much smaller than C :

$$\mathbb{E}_P[L(\delta_Q(Y), \theta)] \leq \alpha + \mathbb{E}_P[QR(Y) + (1 - Q)C].$$

If $Q = \mathbb{1}(R(Y) \leq C)$, this bound simplifies to $\alpha + \mathbb{E}_P[\min(R(Y), C)]$.

This model encodes that P-certification is essentially the only thing that the decision-maker knows.¹¹

Relative to this model, the decision-maker's *adoption problem* has the adoption loss

$$L_{\text{adopt}}(q; \ell(\cdot), \delta(Y), C) = q\ell(\delta(Y)) + (1 - q)C \quad q \in [0, 1]. \quad (8)$$

A decision rule for this problem is an adoption probability function $q(a, r) \in [0, 1]$, prescribing the probability of adoption as a function of recommended action $a = \delta(Y)$ and certificate $r = R(Y)$. The corresponding risk function is then

$$R_{\text{adopt}}(q(\cdot, \cdot); \ell(\cdot), C, G) := \mathbb{E}_G[q(\delta(Y), R(Y))\ell(\delta(Y)) + (1 - q(\delta(Y), R(Y)))C].$$

By [Proposition 2](#), if $q \leq u\mathbb{1}(R(Y) \leq C)$, then $R_{\text{adopt}}(q; \ell, C) \leq C + u\alpha(1 - C)$ for all $(\ell(\cdot), G) \in \mathcal{M}_{\text{DM}}^{(P, \alpha)}$.¹²

We suppose that the decision-maker balances between subjective Bayes risk minimization and worst-case risk over $\mathcal{M}_{\text{DM}}^{(P, \alpha)}$, following [Banerjee et al. \(2020\)](#). The decision-maker has prior π over (ℓ, G) and minimizes a weighted average of Bayes risk under π and worst-case risk over $\mathcal{M}_{\text{DM}}^{(P, \alpha)}$: For some weight $\omega \in (0, 1)$

$$\min_{q \in \mathcal{Q}} \left\{ (1 - \omega) \mathbb{E}_{(\ell, G) \sim \pi} [R_{\text{adopt}}(q(\cdot, \cdot); \ell(\cdot), C, G)] + \omega \sup_{(\ell, G) \in \mathcal{M}_{\text{DM}}^{(P, \alpha)}} R_{\text{adopt}}(q(\cdot, \cdot); \ell(\cdot), C, G) \right\} \quad (9)$$

Equation (9) encodes a trade-off between ambiguity-averse preferences ([Gilboa and Schmeidler, 1989](#)) and subjective expected utility maximization.

The next proposition characterizes optimal adoption decisions for such preferences. In particular, among adoption decision rules that solely depend on the reported certificate, the optimal adoption decision is either to never adopt, or to adopt with a smoothed version of the threshold rule $\mathbb{1}(R(Y) \leq C)$, where the exact smoothing depends on the prior π and the weight ω .

Proposition 3. *Assume:*

¹¹To prove that $(\delta(\cdot), R(\cdot))$ is a P-certified decision, when the loss function L is money-metric the analyst can in principle offer insurance policies that pay out 1 unit of utility if it turns out that $L(\delta(Y), \theta) > R(Y)$, in exchange for α units of insurance premium. The decision-maker can purchase any amount of such policies before seeing the data. If $(\delta(\cdot), R(\cdot))$ is indeed a P-certificate, offering such insurance has nonnegative expected value for the analyst. As [Neyman \(1937\)](#) writes, “The theoretical statistician constructing [a confidence interval] may be compared with the organizer of a game of chance in which the gambler has a certain range of possibilities to choose from while, whatever he actually chooses, the probability of his winning and thus the probability of the bank losing has permanently the same value, $[\alpha]$.”

¹²Note that this model rationalizes the simple threshold decision $\mathbb{1}(R(Y) \leq C)$ as itself as-if optimization under the confidence set $\{l : l \leq R(Y)\}$, valid under all (ℓ, G) , for $\ell(\delta(Y))$.

- (1) $C \in (0, 1)$ and $\mathcal{A}$ contains at least two elements.
- (2) The class of decision rules $\mathcal{Q} = \{q : \mathcal{A} \times [0, 1] \rightarrow [0, 1] : q(a, r) = q(r)\}$ consists of rules that do not depend on the reported action $\delta(Y)$.
- (3) The prior π in (9) is optimistic in the sense that

$$m(r) := \mathbb{E}[\ell(\delta(Y)) \mid R(Y) = r] \text{ is weakly increasing and } m(C) < C$$

where the expectation is over the joint distribution of $\ell(\cdot), \delta(Y), R(Y)$ implied by $(\ell, G) \sim \pi$ and $(\delta(Y), R(Y)) \sim G$. Suppose $m(r_B) = C$ at some $r_B > C$.

Then there exists a weakly decreasing function $t^* : (0, 1) \rightarrow [0, \frac{r_B - C}{1 - C}]$ with the following property: for every $\omega \in (0, 1)$, there is an optimal adoption rule q_ω^* for (9) that is either identically zero or has the form

$$q_\omega^*(r) = \begin{cases} 1, & \text{if } r \leq C \\ \min \left\{ 1, \frac{1-C}{r-C} t^*(\omega) \right\} & \text{if } C < r < r_B \\ 0 & \text{if } r \geq r_B \end{cases}.$$

Proposition 3 shows that the optimal adoption decision is a compromise between the simple threshold rule $q(r) = \mathbb{1}(r \leq C)$ and the Bayes decision $q_B(r) = \mathbb{1}(r \leq r_B)$, under beliefs such that subjective expected loss is increasing in the reported certificate. When the certificate $R(Y)$ falls below r_B , the subjective part of the decision-maker’s preferences would push for adoption. The presence of the ambiguity-averse preferences then reduces adoption when the certificate $R(Y)$ is below r_B but above the default cost C . In this region, even conditional on the certificate being valid, the worst-case loss could still exceed C but expected losses are below C . The optimal rule balances this possibility—along with the possibility that the certificate fails—with expected losses. As the weight ω on the pessimistic part increases, adoption in this “ambiguous” region $r \in [C, r_B]$ becomes rarer. Eventually, as ω increases, the pessimism completely dominates and no adoption decision beats the default cost C .

Together, **Proposition 2** and **Proposition 3** show that P-certified decision rules are approximately compatible with simple adoption decisions for an ambiguity-averse decision-maker willing to tolerate some increase in her worst-case risk relative to a safe option. Thus, P-certified decisions offer downstream decision-makers—including ones less familiar with the statistical environment and the mapping from parameters to loss—the opportunity to take data-driven decisions without sacrificing (much) safety.

Note that **Proposition 2** and **Proposition 3** apply to *any* P-certified decisions, including the trivial certificate where $R(Y) = 1$ with probability $1 - \alpha$, independent of

$\delta(Y)$. Thus, these results do not require that the analyst be particularly skilled in data analysis, nor that they necessarily share the same objective as the decision-maker, just that the analyst be able to credibly provide P-certificates.

Example 4 (Inference on winners, continued). We next return to [Example 1](#) but now consider a setting where a policymaker chooses whether to adopt an analyst’s recommendations on neighborhood choice. We assume that the policymaker faces a fixed cost κ of making a recommendation at all, for instance because such recommendations have not been made in the past and this new program needs to be explained to the public. To represent this decision-maker’s preferences we can consider the loss function $L(a, \theta) = 1 - \theta(a)$, and take $C = 1 - \theta(a_0) - \kappa$ where for simplicity we treat $\theta(a_0)$ as known.

Suppose the decision-maker is assisted by an analyst who has access to data and knows the joint distribution of estimation error, where for $a \in \mathcal{A}$, $\frac{X(a) - \theta(a)}{\sigma(a)} \sim Z(a)$. In this setting, we again take $Y = (X(\cdot), \sigma(\cdot))$. The analyst may select a recommended action via empirical welfare maximization and a corresponding P-certificate via either the unstudentized or studentized projections ($R(Y)$ or $R^*(Y)$ in [Example 2](#)).

Given either $(\delta(Y), R(Y))$ or $(\delta(Y), R^*(Y))$, the decision-maker must decide whether to implement $\delta(Y)$. The simple adoption rules $\mathbb{1}\{R(Y) \leq C\}$ and $\mathbb{1}\{R^*(Y) \leq C\}$ are size- α tests for the null $L(\delta(Y), \theta) \geq C$ —following them means that the decision-maker adopts the recommended action only when the analyst rejects the null that it increases loss relative to the status quo. [Proposition 2](#) shows that such adoption rules control worst-case excess loss, and [Proposition 3](#) shows that, under preferences that weight Bayes and worst-case risks, optimal R -dependent adoption rules are an interpolation between the simple rule and the Bayes optimal decision rule. ■

3. Illustrations of P-certified decisions

We next illustrate how the certified decisions perspective helps interpret empirical findings and construct recommendations in two applications.

3.1. Creating Moves to Opportunity. In an experiment known as Creating Moves to Opportunity (CMTO), [Bergman et al. \(2024\)](#) offered low-income Seattle families assistance to move to neighborhoods designated as *high-mobility*. These high-mobility tracts are roughly the top third of census tracts in the area, ranked by noisy estimates of upward mobility from the Opportunity Atlas ([Chetty et al., 2026](#)). [Bergman et al. \(2024\)](#), in selecting these Census tracts, face the challenge of only observing noisy

estimates of tract mobility. Different statistical inference analyses can thus reach different inferential conclusions about the extent to which the data support designating the selected tracts as high-mobility.

In one analysis, [Andrews *et al.* \(2024\)](#) compare the *average mobility* of the selected tracts against the average mobility of tracts where the voucher holders currently reside. They find evidence that, indeed, the select tracts are higher mobility—in the sense that confidence sets for this difference-of-averages exclude zero. In another analysis, however, [Mogstad *et al.* \(2024\)](#) instead ask which tracts can be confidently placed in the true top third. They find that, for the 397 tracts in Seattle and King County, *every tract* is included in the 95% confidence set for the true top-third (and also in the confidence set for the bottom-third) of tracts. Thus, at conventional levels, there is insufficient evidence to support that any given tract—even the one with the highest mobility estimate—truly belongs to the top third.

Of course, [Andrews *et al.* \(2024\)](#) and [Mogstad *et al.* \(2024\)](#) consider different inferential questions, which are in turn relevant for different decision problems. In particular, we show in this section that each inference result corresponds to a P-certificate for an appropriately formulated problem. The key distinction between these decision problems is the ambiguity set assumed for relocation patterns—that is, for a given selection of target tracts, where will voucher households move from, and where will they move to? If prospective movers are selected at random from the population of housing voucher holders with children and choose targets uniformly at random among recommended tracts, then the analysis of [Andrews *et al.* \(2024\)](#) is directly reassuring. Instead, if movers could systematically move from the best unselected tract to the worst selected tract, then the CMTO program might not actually improve economic mobility at all, corresponding to the conclusion of [Mogstad *et al.* \(2024\)](#). Consequently, as [Mogstad *et al.* \(2024\)](#) write, “Assumptions on location choices indeed are needed to be statistically confident [...] that CMTO would have a positive impact” (p.514).

To formalize this connection, consider an analyst who recommends a set $S \subset \{1, \dots, n\}$ of $m := |S| = \lfloor n/3 \rfloor$ target tracts. We let θ_i be the true economic mobility of tract i , collected in a vector $\theta \in \mathbb{R}^n$. Suppose the status-quo locations of prospective movers are summarized by a probability vector $p = (p_1, \dots, p_n)$, while conditional on S movers choose new locations according to a probability vector q supported on S . The average change in economic mobility for such a population of movers is then $(q - p)' \theta$. For a given class $W(S) \in \Delta^{n-1}$ of p and $V(S) \in \Delta^{n-1}$ of q —to be

specified—containing the plausible set of movement patterns, we will consider the worst-case mobility gain over $W(S)$ and $V(S)$:

$$U(S, \theta; W, V) := \min_{p \in W(S), q \in V(S)} (q - p)' \theta. \quad (10)$$

This then generates a large menu of decision problems indexed by $W(\cdot), V(\cdot)$.

We winnow down this menu by choosing some parametrizations for $W(\cdot), V(\cdot)$ that vary in how “spread out” movers’ current and future locations are. One intuitive parametrization for plausible targets $V(\cdot)$ is one where movers spread over at least a fraction $\rho \in [0, 1]$ of the selected tracts:

$$V_\rho(S) := \left\{ q \in \Delta^{n-1} \text{ supported on } S : q_i \leq \frac{1}{1 \vee \lceil \rho m \rceil} \right\}.$$

With $\rho = 0$, $V_{\rho=0}(S)$ allows all movers to settle in the same tract; with $\rho = 1$, $V_{\rho=1}(S)$ only contains the uniform distribution on S . We also consider two specifications for allowable distributions $W(\cdot)$ of where movers come from. The first specification fixes $W(S) = \{b\}$, where b_i is the share of housing-voucher households with children living in tract i in 2018, the year CMT0 began.¹³ The second specification mirrors $V_\rho(S)$ and considers adversarial choices of p :

$$W_\rho(S) := \left\{ p \in \Delta^{n-1} \text{ supported on } S^C : p_i \leq \frac{1}{1 \vee \lceil \rho(n - m) \rceil} \right\},$$

where we assume that in the status quo, movers are spread out across a fraction ρ of the non-selected tracts. We call these two regimes *voucher origins* and *adversarial origins*, respectively.¹⁴

Having parametrized V and W , we now have a menu of decision problems, indexed by ρ and by whether p obeys the voucher-origin or adversarial-origin regime. Over W_ρ and V_ρ , we can also directly compute the minimum in (10). Let $\underline{\theta}_\rho(S)$ be the average of the smallest ρ fraction¹⁵ of θ_i values for $i \in S$, and let $\bar{\theta}_\rho(S)$ be the analogous quantity over the largest values. We thus have:

$$U_{\text{voucher}}(S, \theta; \rho) := U(S, \theta; \{b\}, V_\rho) = \underline{\theta}_\rho(S) - b' \theta$$

¹³We follow Andrews *et al.* (2024) in choosing b .

¹⁴To give a little intuition, in the actual CMT0 experiment, the treated families spread across 46 distinct tracts out of 132. A back-of-the-envelope upper bound for ρ is thus $46/132 \approx 0.35$. Separately, for each CZ we can compute an estimate of ρ for how spread out voucher holders’ current locations are, $\lfloor 1/\max_i b_i \rfloor / (n - m)$. Its median across the 50 largest Commuting Zones is about 0.096 (0.143 for King County).

¹⁵Technically, the average of the smallest- $(\lceil \rho |S| \rceil \vee 1)$ values.

$$U_{\text{adver}}(S, \theta; \rho) := U(S, \theta; W_\rho, V_\rho) = \underline{\theta}_\rho(S) - \bar{\theta}_\rho(S^C). \quad (11)$$

We can interpret [Andrews et al. \(2024\)](#) and [Mogstad et al. \(2024\)](#) as targeting different decision problems in the two “corners” of this menu. [Andrews et al. \(2024\)](#) construct confidence intervals for $U_{\text{voucher}}(S, \theta; \rho = 1)$. Instead, asking if every selected tract is in the top third—parallel to [Mogstad et al. \(2024\)](#)—examines whether $U_{\text{adver}}(S, \theta; \rho = 0) \geq 0$. Different values of ρ then interpolate between these two corner cases.

We next construct P-certified recommendations for decision problems on this menu. The data $Y := (Y_1, \dots, Y_n)$ consists of Opportunity Atlas estimates of $\theta := (\theta_1, \dots, \theta_n)$, which we assume are independent and Gaussian with known variances: $Y_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$, as in [Andrews et al. \(2024\)](#), [Mogstad et al. \(2024\)](#) and [Chen \(2026\)](#). To P-certify decisions, it suffices to construct a confidence set for θ , which then implies P-certified decisions via as-if optimization.

Given W, V , a straightforward *fixed-width* confidence set is

$$\hat{\Theta}_{\text{FW}}(Y) = \left\{ \vartheta \in \mathbb{R}^n : \begin{array}{l} (q - p)' \vartheta \geq (q - p)' Y - c_\alpha \\ \text{for every } S \text{ with } |S| = m, p \in W(S), q \in V(S) \end{array} \right\}.$$

The critical value c_α can be chosen as a value satisfying

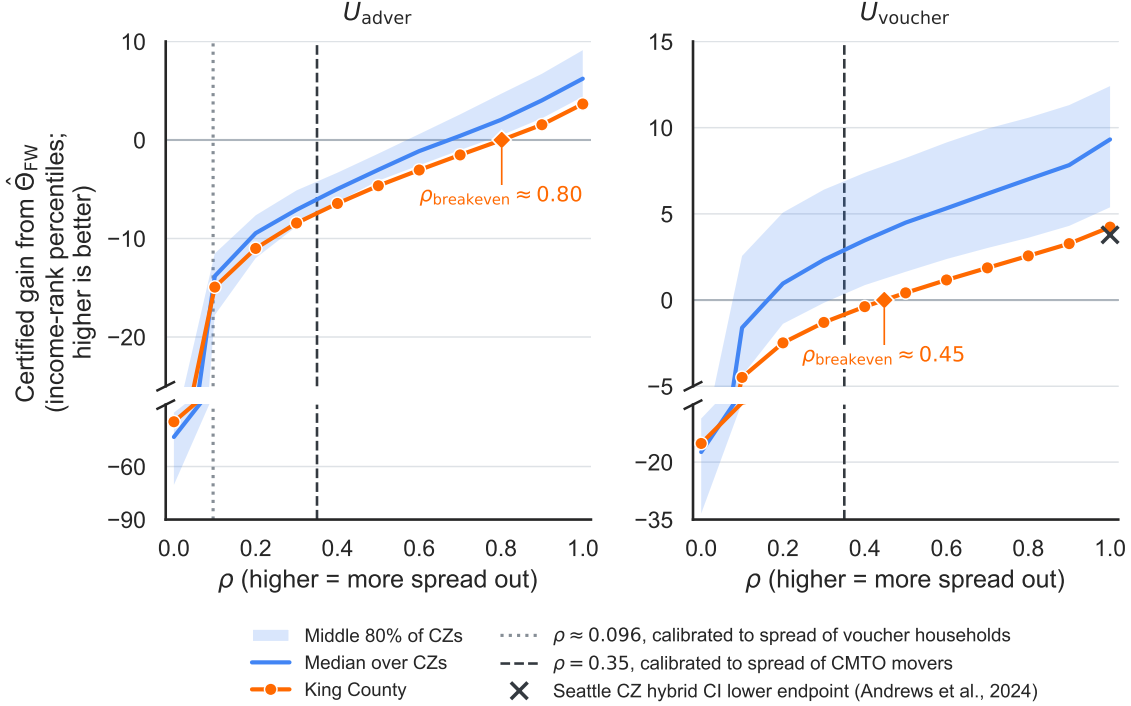
$$\mathbb{P}_{\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)} \underbrace{\left\{ \max_{\substack{S \subseteq \{1, \dots, n\} \\ |S| = m}} \max_{\substack{p \in W(S) \\ q \in V(S)}} (q - p)' \epsilon \leq c_\alpha \right\}}_{\text{efficiently computable when } (W, V) = (W_\rho, V_\rho)} \geq 1 - \alpha.$$

This confidence set corresponds to the following P-certified decisions for each of the decision problems in (11) (where c_α depends on the variant of the decision problem). Note that, since we consider utility rather than loss here, certificates are stochastic *lower bounds* of utility:

$$\delta_{\text{FW}}(Y) = \{\text{Largest } m \text{ entries of } Y\}, \quad R_{\text{FW}}(Y) = \underbrace{U(\delta_{\text{FW}}(Y), Y; \rho)}_{\text{plug-in estimate of mobility increase}} - c_\alpha. \quad (12)$$

Regardless of the decision problem and of the variances of Y , we always recommend the tracts with the point estimates in the top third. The certificate (12) then adjusts a plug-in estimate of mobility increase by a penalty for the uncertainty. As such, $\hat{\Theta}_{\text{FW}}$ has the drawback of ignoring precision of the estimates when choosing S . In [Appendix C.1](#), we consider a variable-width confidence set which penalizes imprecision when as-if optimizing the selected tracts; it gives different recommended tracts but similar certificates to those from this fixed-width confidence set.

FIGURE 1. Certificates for mobility gains across the largest Commuting Zones



Notes. We compute certificates R_{FW} for decision problems in (11) for each of the largest 50 CZs, following definitions in Andrews *et al.* (2024). For each decision problem, we plot the 10th, 50th, and 90th percentiles of these certificates across the 50 CZs. We also plot the certificates for King County, WA. We also plot the two spread references discussed in Footnote 14: the origin reference $\rho \approx 0.096$ in the adversarial-origins panel and the destination reference $\rho \approx 0.35$ in both panels. $\square$

We plot the certificates (12) in Figure 1 across the menu of decision problems in (11). We do so both for King County, WA—as in the actual CMT0 experiment—and for the 50 largest Commuting Zones (CZs). Unsurprisingly, we can certify higher mover mobility gains when movers are more spread out (higher ρ) or under the voucher-origins regime, as opposed to the adversarial-origins regime. Either direction takes one closer to Andrews *et al.* (2024)’s corner of the menu than to Mogstad *et al.* (2024)’s. Our exercise goes beyond this qualitative description and exactly quantifies how much robustness to location choices we can demand while still certifying beneficial selections.

At the 5% level, we can certify a beneficial selection for King County when $\rho \geq 0.45$, even under adversarial destination choices, provided movers’ origins follow the observed distribution of voucher holders. In particular, when $\rho = 1$, the fixed-width lower bound is 4.0 rank percentiles, similar to the (two-sided) lower bound of 3.8 for the hybrid inference method in Andrews *et al.* (2024, Figure V). At $\rho \approx 0.35$ —calibrated for how spread out the location choices are for those in the actual CMTO experiment—the certificate is slightly negative. Thus, if movers were to move adversarially in the CMTO—maintaining the spread of locations—we cannot certify that the recommendations are net-beneficial at the 5% level.

If we also require robustness to the movers’ current locations, we can only certify a beneficial selection if $\rho \geq 0.8$. This adversarial-origins regime turns out to be less favorable both because the plug-in estimate is lower, $U_{\text{adver}}(\delta, Y; \rho) \leq U_{\text{voucher}}(\delta, Y; \rho)$, and because the adversarial regime generally has a larger penalty c_α for uncertainty. Finally, King County turns out to be a comparatively difficult case: We can certify beneficial selections for a majority of the CZs at lower values of ρ for either of the two regimes. Of course, if movement patterns were not maximally adversarial given the permitted degree of concentration ρ , one would obtain more optimistic conclusions.

3.2. Routing large language models (LLMs). Next, we illustrate certified decisions in the context of LLM routing. Arena.ai publishes a leaderboard of AI models based on user preferences.¹⁶ Arena’s leaderboards are influential in the AI industry as measures of model quality. We study how a decision-maker should route queries across LLMs when model quality is unknown. This setting is similar to that of the empirical exercise in Zrnic and Fithian (2025), but more explicitly cast as a decision problem. A decision problem of this sort—at least after conditioning on sufficiently rich prompt and context characteristics—is faced by automatic routing services such as Cursor Router, GitHub Copilot Auto, and OpenRouter Auto.¹⁷

Arena’s measures are derived from pairwise user comparisons of LLM outputs (Chiang *et al.*, 2024). The resulting *Arena Scores* are scaled Bradley–Terry estimates displayed on an Elo-like scale; we refer to them as Elo below (see Hamilton *et al.* (2023) for background on the Bradley–Terry model).¹⁸ In this dataset, each LLM i

¹⁶<https://huggingface.co/datasets/lmarena-ai/leaderboard-dataset>

¹⁷See, e.g., <https://cursor.com/blog/router>, <https://docs.github.com/en/copilot/concepts/models/auto-model-selection>, and <https://openrouter.ai/docs/guides/routing/routers/auto-router>.

¹⁸Namely, the fitted probability of model A beating model B is $1/(1 + 10^{(Y_B - Y_A)/400})$, per <https://github.com/lmarena/arena-rank>.

has an estimated Elo score Y_i with standard error σ_i . We will treat $Y_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$ for true Elo score θ_i . In principle, the Y_i are correlated across i since the estimation noise in Y arises from randomly sampled pairwise comparisons. Without access to the variance-covariance matrix of the estimated Elo scores, we adopt the working model that Y_i are independent across i .¹⁹ Arena publishes Elo scores for many tasks; throughout, we use the “Overall” category in “Text Arena.” As of August 24, 2026, the top-five (out of all 394 models and, separately, out of the 218 non-proprietary models) ranked models in this category are shown in Figure 2.

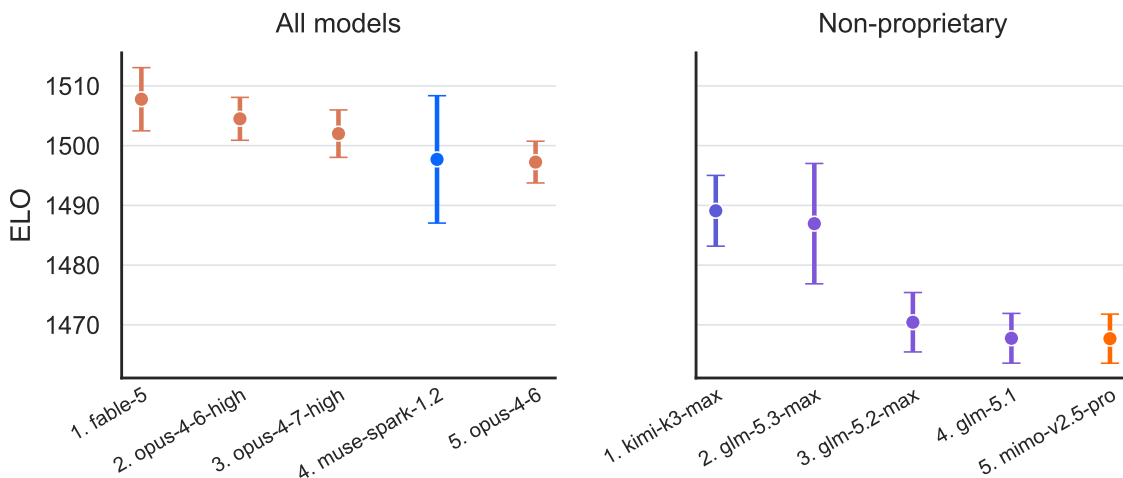


FIGURE 2. The five highest-ranked models among all models (left) and among non-proprietary models (right), as of August 24, 2026, in Text Arena’s Overall category. Bars show marginal 95% confidence intervals.

We first consider a decision maker choosing which LLMs to route queries to with utility function $u(a, \theta) = \theta_a$ (as with the CMTO exercise, certificates are now stochastic lower bounds for utility). We compare a suite of methods for generating P-certified decisions:

- (i) The *conditional* and *hybrid* methods of Andrews *et al.* (2024) take $\delta(Y) = \arg \max_a Y_a$ and generate confidence lower bounds for $\theta_{\delta(Y)}$.
- (ii) *Max-t confidence sets* for $\theta = (\theta_1, \dots, \theta_n)$ imply as-if decisions $\delta(Y)$ and corresponding certificates.

¹⁹This working model is accurate if comparisons between models i and j are small proportions of all comparisons for model i and for model j . Without this assumption, one could consider constructing confidence sets or, later, e-variables robust to dependence (e.g. Bonferroni instead of max-t confidence sets; averaging valid marginal e-variables). One could also consider a limited sensitivity analysis, e.g., by considering equicorrelated covariance matrices.

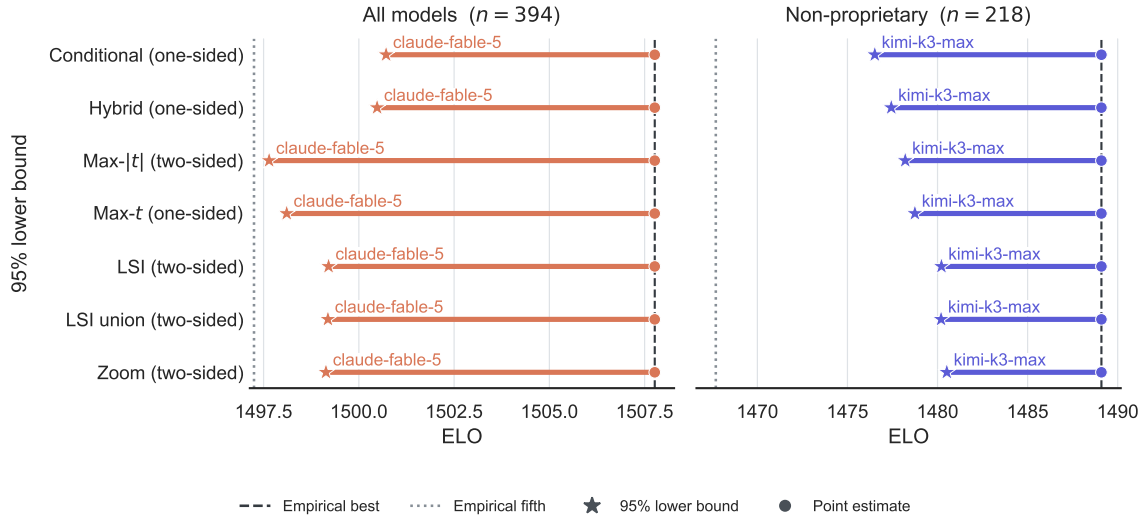


FIGURE 3. P-certificates for Elo

- (iii) *Locally simultaneous inference* (LSI), proposed by Zrnic and Fithian (2024), can generate a confidence set for $\theta_{\delta(Y)}$ given a user-supplied $\delta(\cdot)$. We apply it to certifying the empirical winner $\delta(Y) = \arg \max_a Y_a$. It does so by identifying a data-dependent set of plausible winners and conducting simultaneous inference over that set. We consider two variants of LSI, where one uses a max-t statistic and the other applies a Bonferroni/union-bound-type correction.
- (iv) Finally, the *zoom test*, proposed by Zrnic and Fithian (2025), similarly certifies the empirical winner $\delta(Y) = \arg \max_a Y_a$. Although related to LSI, it is designed specifically for inference on winners (Zrnic and Fithian, 2025).

All of these methods, viewed as P-certified decisions, are already as-if decisions. Thus they cannot be improved by Theorem 1. Since they are all valid P-certified decisions, comparing their corresponding P-certificates measures an analogue of power tailored to the decision problem (though, of course, on a single data draw).

We compare the P-certificates and decisions of these methods in Figure 3. Both among all models and among all non-proprietary models, every method selects the empirical winner (though this is by construction for methods (i), (iii), and (iv)): Claude Fable and Kimi K3, respectively. Every 95% certificate is above the point estimate of the fifth-ranked model. For choosing among all models, conditional and hybrid approaches in Andrews *et al.* (2024) appear to produce the tightest certificate. This is possibly due to the fact that Claude Fable is a clear winner—thus conditioning

on the identity of the empirical winner does not sacrifice too much information. By contrast, among non-proprietary models, Kimi K3 and GLM 5.3 have similar point estimates, which may make the conditional procedures’ certificates less favorable. In either case, various certificates differ by at most 4 Elo points, roughly the gap between the point estimates of Claude Fable and the runner-up Claude Opus 4.6 High.

It is also possible to consider mixed strategies, where an action is a probability vector and its payoff is $a'\theta$. This corresponds to a decision-maker who will make repeated LLM-routing decisions and values long-run average quality. Mixing among models can in principle considerably reduce uncertainty, and thus may be attractive for certified decisions (Chernozhukov *et al.*, 2026). One could consider a joint confidence set for $\{a'\theta : a \in \Delta([n])\}$ (Scheffé, 1953; Kudo, 1963) and choose $\delta(Y) \in \Delta([n])$ via as-if optimization. However, because there are a large number of models here, doing so yields a certificate that is about 10 Elo points *lower* for the setting in the left pane of Figure 3.²⁰ Restricting mixtures to the top-ranked models (selected on a holdout sample by data-fission) raises the resulting certificate, but not appreciably above the bounds shown in Figure 3.

So far, this exercise ignores LLM-inference costs. An alternative objective is to maximize *price-equivalent Elo* $u_\lambda(a, \theta) = \theta_a - \lambda p_a$, for p_i the price per million tokens for model i . We calibrate λ by estimating willingness-to-pay on OpenRouter data. Specifically, we take market share data published by OpenRouter for July 24–August 24, 2026,²¹ defining a model’s market share as the proportion of API requests routed to it. Under a logit random utility model in prices and Elo

$$u_j = \underbrace{\beta_E Y_j + \sigma \epsilon_j + \xi_j}_{\text{WTP for model } j} - p_j, \quad \epsilon_j \stackrel{\text{i.i.d.}}{\sim} \text{Gumbel}, \quad \mathbb{E}[\xi_j Y_j] = \mathbb{E}[\xi_j p_j] = 0,$$

we can estimate β_E by regressing log market shares on an intercept, Elo, and prices, and then normalizing the price coefficient to -1 . Doing so finds that the estimated willingness to pay for 100 Elo is $100\hat{\beta}_E = \$9.24$ per million tokens (95% CI: [\$5.63, \$14.53]). The reciprocal of $\hat{\beta}_E$ is $\hat{\lambda}$.

²⁰The Scheffé-type critical value scales as $\sqrt{n}$ when there are n models, whereas max-t critical values scale like $\sqrt{\log n}$.

²¹<https://openrouter.ai/rankings>, accessed August 24, 2026. This calibration dataset uses 176 paid standard OpenRouter product variants matched to Arena, each with at least 10,000 requests, and covers 83% of monthly chat tokens.

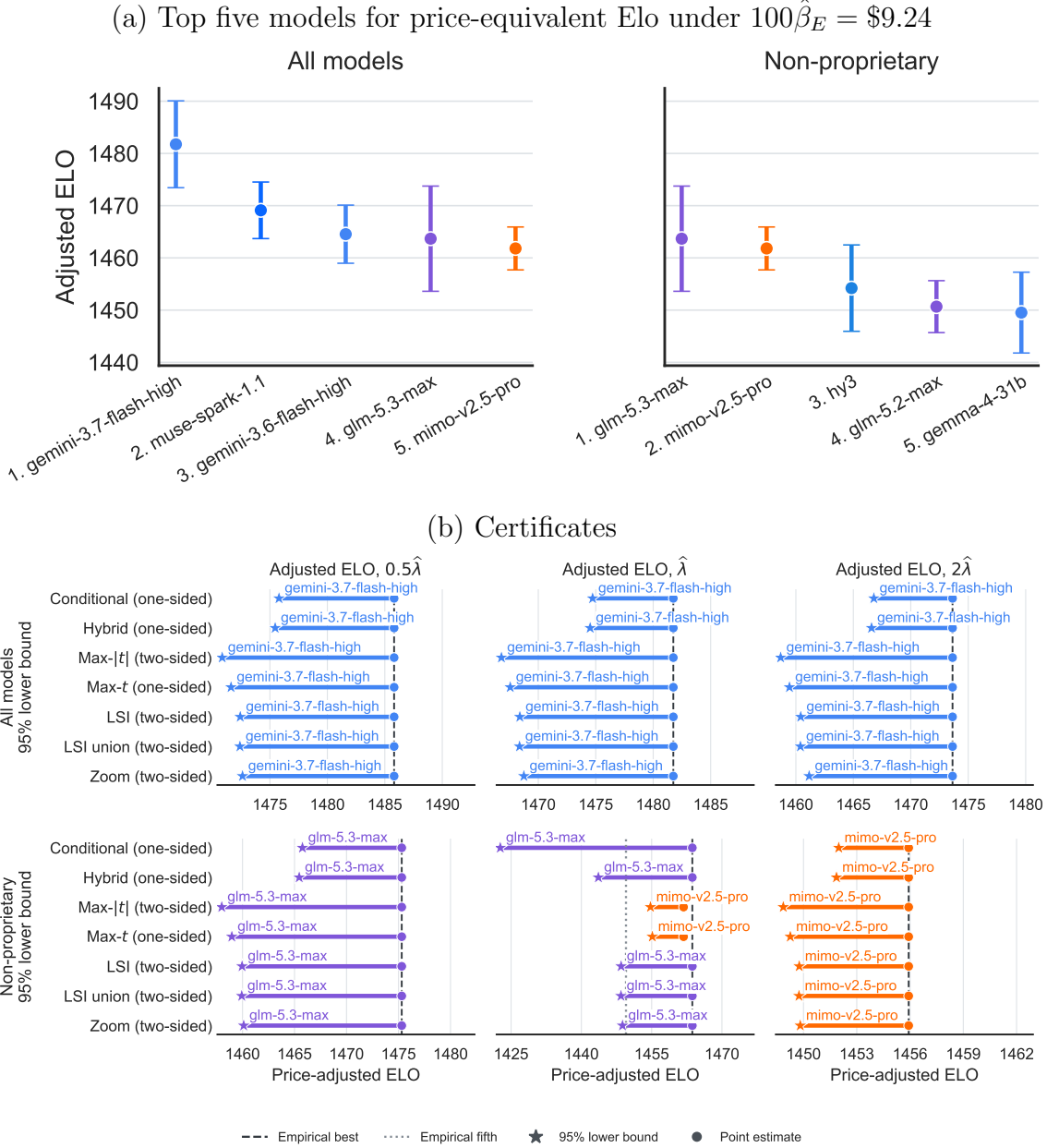


FIGURE 4. P-certificates for price-equivalent Elo

Using λ calibrated under this simple model (as well as under sensitivity analyses by doubling or halving the estimated λ —while ignoring λ 's estimation uncertainty),²² we

²²Arena and OpenRouter have somewhat different calculations of prices as different weighted averages of input and output token prices. We estimate λ and calculate price-equivalent Elo both using the OpenRouter prices. As a result, the price-equivalent exercise restricts to 160 models with matched Arena–OpenRouter data.

can repeat the exercise above in [Figure 4](#). Among all models, every method selects Gemini 3.7 Flash at each of the three values of $\hat{\lambda}$, where conditional and hybrid certificates appear tightest—again, seemingly due to the fact that Gemini 3.7 Flash clearly outperforms the runner-up.

On the other hand, for non-proprietary models at the estimated willingness to pay, the two leading models, GLM 5.3 and Mimo 2.5, have more similar point estimates: GLM 5.3 has the higher estimate but a substantially larger standard error. Consequently, the max- t -based procedures select Mimo 2.5 because its lower confidence bound is higher. This flexibility to optimize $\delta(Y)$ produces tighter certificates than the other methods designed to certify the empirical winner. At $0.5\hat{\lambda}$, every procedure selects the GLM-5.3, which has a higher point estimate; at $2\hat{\lambda}$, every procedure selects the less expensive Mimo 2.5, with the conditional and hybrid procedures producing the highest lower bounds in both cases.

This exercise illustrates how P-certificates allow us to compare inference procedures through the guarantees they provide for decisions. In most versions of this exercise, there is little disagreement over the selected LLM, and thus the tightness of the resulting guarantees reflects the power of inference-on-winner procedures against *one-sided* alternatives. Empirically, if there is a clear winner, conditional inference-type procedures provide tight certificates; otherwise—especially when there is a clear top-two but no clear winner—locally simultaneous inference procedures are more favorable, and almost always outperform max- t intervals. However, in the case where the empirical winner is very noisy ([Figure 4\(b\)](#), middle panel), as-if optimization against max- t intervals ([Chernozhukov et al., 2026](#)) no longer chooses the empirical winner, and potentially leads to much tighter certificates. This latter observation illustrates the advantages of viewing decision and inferential goals as linked and optimizing them jointly.

4. E-certified decisions

We have shown that P-certified decisions offer a natural route to combine (frequentist) statistical inference with decision-making. However, high-probability guarantees, as appearing in confidence intervals and hypothesis tests, are not the only type of frequentist guarantee available. A recent literature in statistics proposes *e-values* as an alternative and a complement to traditional statistical inference (see [Ramdas and Wang, 2024](#), for a review). An *e-variable* against the parameter value θ_0 is a

nonnegative random variable $E(Y, \theta_0)$ such that

$$\mathbb{E}_P[E(Y, \theta_0)] \leq 1 \text{ for all } P \text{ such that } (\theta_0, P) \in \mathcal{M}.$$

One can interpret $E(\cdot, \theta_0)$ as the realized dollar payoff of a bet, priced at \$1, against the hypothesis $H_0 : \theta = \theta_0$. Since the expected payoff under H_0 is less than \$1, this bet is (at most) fair. Correspondingly, a large realization of $E(Y, \theta_0)$ can be interpreted as evidence against H_0 —in the sense that it is an unlikely event under H_0 —much like a small p-value is evidence against H_0 .

As with confidence sets, we can use e-variables to certify the quality of decisions. Throughout this section, we assume $L(a, \theta) \geq 0$, but no longer assume $L(a, \theta) \leq 1$.

Definition 4. For fixed $\gamma > 0$, a pair $(\delta(\cdot), R(\cdot))$ with R non-negative is an *E-certified decision* at multiple γ if

$$\mathbb{E}_P \left[\frac{L(\delta(Y), \theta)}{R(Y)} \right] \leq \gamma \text{ for all } (\theta, P) \in \mathcal{M}.$$

We abbreviate E-certified decisions at multiple 1 as E-certified decisions.

As [Definition 4](#) highlights, $R(Y)$ is a stochastic upper bound for loss in the sense that the ratio $L(\delta(Y), \theta)/R(Y)$ has ex ante expectation at most a prespecified γ .²³ These decisions generalize worst-case expected loss considerations: Observe that ordinary worst-case risk for a given decision rule $R = \sup_{(\theta, P) \in \mathcal{M}} \mathbb{E}_P[L(\delta(Y), \theta)]$ is automatically a constant-in- Y E-certificate for $\delta(Y)$. From this perspective, minimax risk decisions are attractive because they minimize such data-independent E-certificates.

Similar to our previous results, one can derive E-certified decisions using a version of as-if optimization, now defined using e-variables. [Grünwald \(2023\)](#) proposes what he terms the E-posterior: if there exists an e-variable $E(\cdot, \theta)$ for every $\theta \in \Theta$, the E-posterior is defined as $1/E(Y, \theta)$, with the convention $1/0 = \infty$. [Grünwald \(2023\)](#) proposes computing the action that minimizes the worst-case E-posterior-weighted loss, along with the minimized value. Formally, because the infimum over $\mathcal{A}$ need not be attained, let $\mathcal{C}_{E, \epsilon}$ be the following class of (δ, R) such that, for some collection of e-variables $E(\cdot, \theta)$,

$$R(Y) \equiv \sup_{\theta \in \Theta} \frac{L(\delta(Y), \theta)}{E(Y, \theta)} \leq \inf_{a \in \mathcal{A}} \sup_{\theta \in \Theta} \frac{L(a, \theta)}{E(Y, \theta)} + \epsilon \quad (13)$$

Intuitively, for choosing $\delta(\cdot)$, states θ deemed implausible by the data—those with large $E(Y, \theta)$ —have their losses downweighted, while those with small $E(Y, \theta)$ have

²³We use the conventions that $0/0 = 0$ and $x/0 = \infty$ for $x > 0$.

their loss upweighted. Grünwald (2023) shows that $(\delta(Y), R(Y))$ is an E-certified decision at multiple 1.²⁴

We next show that, as with P-certificates in Theorem 1, it is without loss to limit attention to as-if optimization, since the set of E-certified decisions of the form (13) is an essentially complete class with respect to the dominance order.

Proposition 4. *For any E-certified decision $(\tilde{\delta}, \tilde{R})$ and any $\epsilon > 0$, there exists some collection of e-variables $E(Y, \theta)$ for which the corresponding E-certified decision $(\delta, R) \in \mathcal{C}_{E, \epsilon}$ chosen via (13) weakly dominates $(\tilde{\delta}, \tilde{R})$.*

As in our results for P-certificates, using E-certificates also provides guarantees when combined with downstream decision-making. A decision-maker who adopts only when $R(Y) \leq C$ incurs expected loss of at most $2C$. Unlike our results for P-certificates, such a guarantee does not require that the loss function be bounded above.

Proposition 5. *Given an E-certified decision rule (δ, R) , for any adoption decision $Q \leq \mathbb{1}(R(Y) \leq C)$ a.s.,*

$$\mathbb{E}_P [L(\delta_Q(Y), \theta)] \leq \mathbb{E}_P \left[\max \left(\frac{L(\delta(Y), \theta)}{R(Y)}, 1 \right) C \right] \leq 2C \text{ for all } (\theta, P) \in \mathcal{M}$$

Motivated by Proposition 5, if we only sought certificates with $\mathbb{E}_P \left[\max \left(\frac{L(\delta(Y), \theta)}{R(Y)}, 1 \right) \right] \leq 1 + \gamma$ to start with, $R(Y)$ can often be improved by slightly modifying the E-posterior. We detail this improvement in Appendix B.

4.1. Illustration with LLM routing (continued). We illustrate E-certificates in the context of the LLM routing exercise. Since Elo does not have a natural upper bound, loss (negative utility) does not have a natural lower bound. E-certification, however, requires a nonnegative loss function. As a result, we consider *regret* $L(a, \theta) = \max_{a'} u(a', \theta) - u(a, \theta) \geq 0$ as our certification target.

We next discuss one way to construct e-variables $E(Y, \theta)$ to form E-certificates of regret. Fix a θ and let $k = k_\theta$ be some element of $\arg \max \theta_s$ (or $\arg \max \theta_s - \lambda p_s$)—so that k is the best choice under θ . Intuitively, operating as if θ were the true parameter would be sensible, unless there is evidence that $\mathbb{E}[Y_j] > \mathbb{E}[Y_k]$. We thus should prefer e-variables that are large when $\mathbb{E}[Y_j] > \mathbb{E}[Y_k]$. For a fixed competitor $j \neq k$, a sensible

²⁴Here, our setup is slightly different since the minimization is up to a constant ϵ . Nevertheless, observe that since $R(Y) \geq L(\delta(Y), \theta)/E(Y, \theta)$,

$$\mathbb{E}_P [L(\delta(Y), \theta)/R(Y)] \leq \mathbb{E}_P [E(Y, \theta)] \leq 1.$$

TABLE 1. P- and E-certificates for regret

Score	Subset	$ \mathcal{A} $	Selected	$R_P(Y)$	$R_E(Y)$
Raw	All	394	claude-fable-5	20.45	7.29
Raw	Non-proprietary	218	kimik3-max	27.26	14.01
Adjusted $0.5\hat{\lambda}$	All	160	gemini-3.7-flash-high	22.58	7.26
Adjusted $0.5\hat{\lambda}$	Non-proprietary	87	glm-5.3-max	13.81	3.66
Adjusted $\hat{\lambda}$	All	160	gemini-3.7-flash-high	15.00	1.17
Adjusted $\hat{\lambda}$	Non-proprietary	87	glm-5.3-max	22.50	13.73
Adjusted $2\hat{\lambda}$	All	160	gemini-3.7-flash-high	7.91	0.11
Adjusted $2\hat{\lambda}$	Non-proprietary	87	mimo-v2.5-pro	17.06	6.24

Notes. In every case, P- and E-certified decisions select the empirical winner. The corresponding P- and E-certificates are reported as $R_P(Y)$ and $R_E(Y)$, respectively. The P-certificates are derived from as-if optimization with 95% max-t confidence sets. $\square$

choice is the likelihood ratio statistic

$$0 \leq e_j(Y, \theta) := \exp(ct_j(Y, \theta) - c^2/2)$$

where $t_j(Y, \theta) := \frac{Y_j - Y_k - (\theta_j - \theta_k)}{\sqrt{\sigma_j^2 + \sigma_k^2}}$. The variable e_j is an e-variable because $t_j(Y, \theta) \sim_{\theta} \mathcal{N}(0, 1)$ and hence $\mathbb{E}_{\theta}[e_j(Y, \theta)] \leq 1$. The statistic $e_j(Y, \theta)$ is the likelihood ratio of $\mathcal{N}(c, 1)$ against $\mathcal{N}(0, 1)$ evaluated at data $t_j(Y, \theta)$ —thus tuned to detect alternatives where option j ’s quality is c standard errors above option k .

An e-variable against θ can thus be constructed by treating $e_j(Y, \theta)$ as building blocks—aggregating them over choices of j and choices of c . Such a *merger* of e-variables is a typical recipe in the e-values literature (Vovk and Wang, 2021). In this case, an attractive way of combining them is the following: For a set of e-variables $E_1(Y, \theta), \dots, E_J(Y, \theta)$, we can construct the *normalized maximum likelihood* (Grünwald et al., 2024): $E^*(Y, \theta) = \max_j E_j(Y, \theta) / \mathbb{E}_{\theta}[\max_j E_j(Y, \theta)]$, which is an e-variable by construction. We can thus merge the e-variables $e_j(Y, \theta)$ by applying this transform twice: first over $j \neq k$, and a second time over choices of $c \in \{0.5, 1, 2\}$. This process produces e-variables $E(Y, \theta)$, indexed by θ .

We use these e-variables to derive certified decisions for regret in Table 1 (see Appendix C.2 for how they enable efficient computation). We also show P-certified decisions derived from max- $|t|$ confidence sets. In every setting that we consider, the certified decisions agree with the empirical winner. The E-certificates are numerically

lower than the 95% P-certificates, sometimes substantially, although the two types of certificate provide different statistical guarantees.

5. Conclusion

This paper considers combining statistical inference with statistical decisions. In two leading modes of frequentist inference, statistical inference translates to *certified decisions*—decisions paired with certificates of their performance. We show that it is without loss to base certified decisions on statistical inference procedures, and thus that inferential and decision goals are, at the very least, partially aligned. We also show that these certified decisions provide ambiguity-averse downstream decision-makers with useful risk guarantees and simple adoption decision rules—which may be implemented and understood without statistical sophistication on the part of the decision-maker.

References

- ANDREWS, I., CHEN, J. and TECCHIO, O. (2025). The purpose of an estimator is what it does: Misspecification, estimands, and over-identification. *arXiv preprint arXiv:2508.13076*. [6](#)
- , KITAGAWA, T. and MCCLOSKEY, A. (2024). Inference on winners. *The Quarterly Journal of Economics*, **139** (1), 305–358. [3](#), [4](#), [5](#), [7](#), [8](#), [11](#), [18](#), [19](#), [20](#), [21](#), [22](#), [23](#), [24](#)
- BANERJEE, A. V., CHASSANG, S., MONTERO, S. and SNOWBERG, E. (2020). A theory of experimenters: Robustness, randomization, and balance. *American Economic Review*, **110** (4), 1206–1230. [15](#)
- BATES, S., JORDAN, M. I., SKLAR, M. and SOLOFF, J. A. (2024). Incentive-theoretic bayesian inference for collaborative science. *arXiv preprint arXiv:2307.03748*. [2](#)
- BEN-MICHAEL, E., GREINER, D. J., IMAI, K. and JIANG, Z. (2021). Safe policy learning through extrapolation: Application to pre-trial risk assessment. *arXiv preprint arXiv:2109.11679*. [3](#)
- BENJAMINI, Y., HECHTLINGER, Y. and STARK, P. B. (2019). Confidence intervals for selected parameters. *arXiv preprint arXiv:1906.00505*. [3](#), [7](#)
- BERGMAN, P., CHETTY, R., DELUCA, S., HENDREN, N., KATZ, L. F., and PALMER, C. (2024). Creating moves to opportunity: Experimental evidence on

- barriers to neighborhood choice. *American Economic Review*, **114** (5), 1281–1337. [5](#), [8](#), [17](#)
- CHEN, J. (2026). Empirical bayes when estimation precision predicts parameters. *Econometrica*, **94** (2), 305–340. [20](#)
- CHENG, N., SPECTOR, A. and JANSON, L. (2025). Chiseling: Powerful and valid subgroup selection via interactive machine learning. *arXiv preprint arXiv:2509.19490*. [3](#)
- CHERNOZHUKOV, V., HONG, H. and TAMER, E. (2007). Estimation and confidence regions for parameter sets in econometric models 1. *Econometrica*, **75** (5), 1243–1284. [10](#)
- , LEE, S., ROSEN, A. M. and SUN, L. (2026). Policy learning with confidence. *arXiv preprint arXiv:2502.10653v3*. [3](#), [4](#), [5](#), [11](#), [25](#), [27](#)
- CHETTY, R., FRIEDMAN, J. N., HENDREN, N., JONES, M. R. and PORTER, S. R. (2026). The opportunity atlas: Mapping the childhood roots of social mobility. *American Economic Review*, **116** (1), 1–51. [17](#)
- CHIANG, W.-L., ZHENG, L., SHENG, Y., ANGELOPOULOS, A. N., LI, T., LI, D., ZHANG, H., ZHU, B., JORDAN, M., GONZALEZ, J. E. *et al.* (2024). Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*. [22](#)
- CHRISTENSEN, T., MOON, H. R. and SCHORFHEIDE, F. (2026). Optimal decision rules when payoffs are partially identified. *Review of Economic Studies*, p. rdag017. [10](#)
- EPANOMERITAKIS, A. and VIVIANO, D. (2026). When is statistical evidence strong enough? using hypothesis tests to value data collection. *arXiv preprint arXiv:2609.09544*. [2](#)
- EROL, S. and VOHRA, R. (2026). Coverage guarantees as ambiguity sets: Coherent decision making under statistical uncertainty. [2](#)
- GALEOTTI, A., GOLUB, B., GOYAL, S., TALAMAS, E. and TAMUZ, O. (2024). Robust market interventions. *arXiv preprint arXiv:2411.03026*. [13](#)
- GILBOA, I. and SCHMEIDLER, D. (1989). Maxmin expected utility with non-unique prior. *Journal of mathematical economics*, **18** (2), 141–153. [15](#)
- GOLDFARB, D. and IYENGAR, G. (2003). Robust portfolio selection problems. *Mathematics of operations research*, **28** (1), 1–38. [2](#), [7](#)
- GRÜNWARD, P., HAO, Y. and BALSUBRAMANI, A. (2024). Growth-optimal e-variables and an extension to the multivariate csisz\’ar-sanov-chernoff theorem.

- arXiv preprint arXiv:2412.17554*. 30
- GRÜNWARD, P. D. (2023). The e-posterior. *Philosophical Transactions of the Royal Society A*, **381** (2247), 20220146. 5, 28, 29, 41, 42
- HAMILTON, I., TAWN, N. and FIRTH, D. (2023). The many routes to the ubiquitous bradley-terry model. *arXiv preprint arXiv:2312.13619*. 22
- HENRY, M. and ONATSKI, A. (2012). Set coverage and robust policy. *Economics Letters*, **115** (2), 256–257. 2, 7, 10
- IMBENS, G. W. and MANSKI, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, **72** (6), 1845–1857. 10
- KITAGAWA, T. and TETENOV, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, **86** (2), 591–616. 3, 8
- KIYANI, S., PAPPAS, G., ROTH, A. and HASSANI, H. (2025). Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. *arXiv preprint arXiv:2502.02561*. 4, 6, 10
- KUDO, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, **50** (3/4), 403–418. 25
- LEHMANN, E. L. and ROMANO, J. P. (2024). *Testing statistical hypotheses*. Springer. 12
- LIM, C. and PARK, Y. (2026). Dynamically consistent statistical decisions. *arXiv preprint arXiv:2607.10519*. 7, 10
- MANSKI, C. F. (2021). Econometrics for decision making: Building foundations sketched by haavelmo and wald. *Econometrica*, **89** (6), 2827–2853. 1, 2, 7
- MOGSTAD, M., ROMANO, J. P., SHAIKH, A. M. and WILHELM, D. (2024). Inference for ranks with applications to mobility across neighbourhoods and academic achievement across countries. *Review of Economic Studies*, **91** (1), 476–518. 5, 18, 20, 21
- NEYMAN, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **236** (767), 333–380. 15
- (1977). Frequentist probability and frequentist statistics. *Synthese*, **36** (1), 97–131. 2
- PONOMAREV, K. and SEMENOVA, V. (2024). On the lower confidence band for the optimal welfare. *arXiv preprint arXiv:2410.07443*. 3
- RAMDAS, A. and WANG, R. (2024). Hypothesis testing with e-values. *arXiv preprint arXiv:2410.23614*. 5, 27

- SCHEFFÉ, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, **40** (1-2), 87–110. [25](#)
- TETENOV, A. (2016). An economic theory of statistical testing. [2](#)
- VIVIANO, D., WUTHRICH, K. and NIEHAUS, P. (2025). A model of multiple hypothesis testing. *arXiv preprint arXiv:2104.13367*. [2](#)
- VOVK, V. and WANG, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, **49** (3), 1736–1754. [30](#)
- WALD, A. (1949). Statistical decision functions. *The Annals of Mathematical Statistics*, pp. 165–205. [2](#)
- YAMIN, J. C. (2026). When and how to pilot: Design rules for two-wave experiments. *arXiv preprint arXiv:2607.16982*. [3](#)
- ZRNIC, T. and FITHIAN, W. (2024). Locally simultaneous inference. *The Annals of Statistics*, **52** (3), 1227–1253. [3](#), [5](#), [7](#), [24](#)
- and — (2025). A flexible defense against the winner’s curse. *The Annals of Statistics*, **53** (6), 2516–2535. [3](#), [5](#), [7](#), [22](#), [24](#)

Appendix A. Proofs

Proposition 2. For $u \in [0, 1]$ and any $(\theta, P) \in \mathcal{M}$, any adoption decision Q with $\mathbb{E}[Q \mid Y] \leq u \mathbb{1}(R(Y) \leq C)$ has bounded risk

$$\mathbb{E}_P[QL(\delta(Y), \theta) + (1 - Q)C] \leq C + u\alpha(1 - C).$$

Proof. Fix $(\theta, P) \in \mathcal{M}$, write $A = \{L(\delta(Y), \theta) \leq R(Y)\}$, and let $\bar{q}(Y) = \mathbb{E}[Q \mid Y]$. Since $\bar{q}(Y) \leq u \mathbb{1}\{R(Y) \leq C\}$,

$$\begin{aligned} & \mathbb{E}_P[QL(\delta(Y), \theta) + (1 - Q)C] - C \\ &= \mathbb{E}_P[\bar{q}(Y)(L(\delta(Y), \theta) - C)] \\ &\leq \underbrace{\mathbb{E}_P[\mathbb{1}(A)\bar{q}(Y)(L(\delta(Y), \theta) - C)]}_{\leq 0} + \mathbb{E}_P[\mathbb{1}(A^c)\bar{q}(Y)(1 - C)] \\ &\leq u(1 - C)P(A^c) \leq u\alpha(1 - C). \end{aligned}$$

The second-to-last line uses that the first expectation is nonpositive: whenever $\bar{q}(Y) > 0$, we have $R(Y) \leq C$, and on A this implies $L(\delta(Y), \theta) \leq C$. $\square$

Proposition 3. Assume:

- (1) $C \in (0, 1)$ and $\mathcal{A}$ contains at least two elements.
- (2) The class of decision rules $\mathcal{Q} = \{q : \mathcal{A} \times [0, 1] \rightarrow [0, 1] : q(a, r) = q(r)\}$ consists of rules that do not depend on the reported action $\delta(Y)$.

(3) The prior π in (9) is optimistic in the sense that

$$m(r) := \mathbb{E}[\ell(\delta(Y)) \mid R(Y) = r] \text{ is weakly increasing and } m(C) < C$$

where the expectation is over the joint distribution of $\ell(\cdot), \delta(Y), R(Y)$ implied by $(\ell, G) \sim \pi$ and $(\delta(Y), R(Y)) \sim G$. Suppose $m(r_B) = C$ at some $r_B > C$.

Then there exists a weakly decreasing function $t^* : (0, 1) \rightarrow [0, \frac{r_B - C}{1 - C}]$ with the following property: for every $\omega \in (0, 1)$, there is an optimal adoption rule q_ω^* for (9) that is either identically zero or has the form

$$q_\omega^*(r) = \begin{cases} 1, & \text{if } r \leq C \\ \min \{1, \frac{1-C}{r-C} t^*(\omega)\} & \text{if } C < r < r_B \\ 0 & \text{if } r \geq r_B \end{cases}$$

Proof. Let μ be the marginal distribution of $R(Y)$ under the prior $(\ell, G) \sim \pi$ and $\delta(Y), R(Y) \sim G$. By iterated expectations, the subjective excess risk of a rule $q(r)$ is

$$\mathbb{E}_{(\ell, G) \sim \pi} [R_{\text{adopt}}(q; \ell, C, G)] - C = \int_{[0, 1]} q(r)(m(r) - C) \mu(dr).$$

We first calculate worst-case risk. For a fixed q , define

$$u(q) := \sup_{r \in [0, 1]} q(r), \quad b(q) := \sup_{r \in [C, 1]} q(r)(r - C).$$

We claim that

$$\sup_{(\ell, G) \in \mathcal{M}_{\text{DM}}^{(P, \alpha)}} R_{\text{adopt}}(q; \ell, C, G) = C + \alpha(1 - C)u(q) + (1 - \alpha)b(q). \quad (14)$$

Take $(\ell, G) \in \mathcal{M}_{\text{DM}}^{(P, \alpha)}$. On the event $\{\ell(\delta) > R\}$, the excess loss from adoption is at most $u(q)(1 - C)$. On the event $\{\ell(\delta) \leq R\}$, the excess loss is at most $b(q)$, since the excess loss is nonpositive when $R < C$ and is bounded by $b(q)$ when $R \geq C$. Since $b(q) \leq u(q)(1 - C)$ and $G(\ell(\delta) > R) \leq \alpha$,

$$R_{\text{adopt}}(q; \ell, C, G) - C \leq \alpha u(q)(1 - C) + (1 - \alpha)b(q).$$

This bound is sharp. Choose distinct $a^-, a^+ \in \mathcal{A}$. For $r^- \in [0, 1]$ and $r^+ \in [C, 1]$, set $\ell(a^-) = 1$, $\ell(a^+) = r^+$, and let G put probability α on (a^-, r^-) and probability $1 - \alpha$ on (a^+, r^+) . This choice of (ℓ, G) belongs to $\mathcal{M}_{\text{DM}}^{(P, \alpha)}$, and its excess risk is

$$\alpha q(r^-)(1 - C) + (1 - \alpha)q(r^+)(r^+ - C).$$

Taking sequences r^- and r^+ approaching the two suprema for $u(q)$ and $b(q)$ therefore gives (14).

Let $J_\omega(q)$ denote the criterion in (9). Combining the preceding displays yields

$$J_\omega(q) - C = (1 - \omega) \int q(r)(m(r) - C) \mu(dr) + \omega \{ \alpha(1 - C)u(q) + (1 - \alpha)b(q) \}. \quad (15)$$

If $u(q) > 0$, write $q = u(q)\tilde{q}$, so $u(\tilde{q}) = 1$. Both terms on the right side of (15) are homogeneous of degree one in q , and hence

$$J_\omega(q) - C = u(q)(J_\omega(\tilde{q}) - C).$$

For any $\tilde{q}$, if $J_\omega(\tilde{q}) < C$, then it is optimal to choose $u(q) = 1$. Otherwise it is optimal to choose $u(q) = 0$ so that $q = 0$ identically. Consequently, an optimal rule can be chosen either to be identically zero or to satisfy $u(q) = 1$. It remains to characterize the latter rules.

When $u(q) = 1$, set $t = b(q)/(1 - C)$. Monotonicity of m , together with $m(C) < C = m(r_B)$, implies

$$m(r) \leq C \text{ for } r \leq r_B, \quad m(r) \geq C \text{ for } r \geq r_B.$$

Set $t_B = (r_B - C)/(1 - C)$, and for $t \in [0, t_B]$ define

$$\bar{q}_t(r) := \begin{cases} 1, & r \leq C, \\ \min \left\{ 1, \frac{t(1 - C)}{r - C} \right\}, & C < r < r_B, \\ 0, & r \geq r_B. \end{cases}$$

Fix q with $u(q) = 1$ and let $s := \min\{t, t_B\}$. We first compare q and $\bar{q}_s$ pointwise and claim, for every $r \in [0, 1]$, that

$$(\bar{q}_s(r) - q(r))(m(r) - C) \leq 0. \quad (16)$$

For $r \leq C$, $\bar{q}_s(r) = 1 \geq q(r)$ and $m(r) - C < 0$. For $C < r < r_B$, the definition of $b(q)$ gives

$$q(r) \leq \min \left\{ 1, \frac{b(q)}{r - C} \right\} = \min \left\{ 1, \frac{t(1 - C)}{r - C} \right\}.$$

If $t \leq t_B$, the right side is $\bar{q}_s(r)$. If $t > t_B$, then $s = t_B$ and

$$\frac{s(1 - C)}{r - C} = \frac{r_B - C}{r - C} > 1,$$

so again $\bar{q}_s(r) = 1 \geq q(r)$. Finally, for $r \geq r_B$, $\bar{q}_s(r) = 0 \leq q(r)$ and $m(r) - C \geq 0$. This proves (16). Integrating shows that $\bar{q}_s$ has no greater subjective risk than q :

$$\int_{[0,1]} \bar{q}_s(r)(m(r) - C) \mu(dr) \leq \int_{[0,1]} q(r)(m(r) - C) \mu(dr).$$

We next compare worst-case risks. For every $x \in [0, t_B]$, the definition of $\bar{q}_x$ and the assumption $r_B > C$ imply

$$\begin{aligned} u(\bar{q}_x) &= 1, \\ b(\bar{q}_x) &= \sup_{r \in [C, 1]} \bar{q}_x(r)(r - C) = \sup_{C \leq r < r_B} \min\{r - C, x(1 - C)\} = x(1 - C). \end{aligned}$$

In particular,

$$u(\bar{q}_s) = 1 = u(q), \quad b(\bar{q}_s) = s(1 - C) = \min\{b(q), r_B - C\} \leq b(q).$$

Hence (14) gives

$$\begin{aligned} &\sup_{(\ell, G) \in \mathcal{M}_{\text{DM}}^{(P, \alpha)}} R_{\text{adopt}}(\bar{q}_s; \ell, C, G) - \sup_{(\ell, G) \in \mathcal{M}_{\text{DM}}^{(P, \alpha)}} R_{\text{adopt}}(q; \ell, C, G) \\ &= \alpha(1 - C)(u(\bar{q}_s) - u(q)) + (1 - \alpha)(b(\bar{q}_s) - b(q)) \leq 0. \end{aligned}$$

Thus, for every q with $u(q) = 1$, we have constructed $\bar{q}_s$ of the desired form, for some $s \in [0, t_B]$, such that $u(\bar{q}_s) = 1$ and $J_\omega(\bar{q}_s) \leq J_\omega(q)$. It therefore suffices to minimize $J_\omega(\bar{q}_t)$ over $t \in [0, t_B]$ and show that the optimal t is weakly decreasing in ω .

Define the subjective gain from adoption by

$$H(t) := \int_{[0,1]} \bar{q}_t(r)(C - m(r)) \mu(dr).$$

We may check that the function H is continuous, nondecreasing, and concave. Consequently,

$$J_\omega(\bar{q}_t) - C = \omega(1 - C)\{\alpha + (1 - \alpha)t\} - (1 - \omega)H(t). \quad (17)$$

Continuity and compactness ensure that $J_\omega(\bar{q}_t) - C$ has a minimizer. Minimizing $J_\omega(\bar{q}_t) - C$ is equivalent to maximizing

$$H(t) - \lambda_\omega t, \quad \lambda_\omega := \frac{\omega}{1 - \omega}(1 - C)(1 - \alpha),$$

over $t \in [0, t_B]$. If $0 < \omega_1 < \omega_2 < 1$, and t_j is a maximizer for $H(t) - \lambda_{\omega_j} t$, we have

$$(\lambda_{\omega_2} - \lambda_{\omega_1})(t_2 - t_1) \leq 0.$$

Because λ_ω is strictly increasing, $t_2 \leq t_1$. Thus $t^*(\omega)$ can be selected to be weakly decreasing in ω . This proves the result. $\square$

Theorem 1. Assume that $\inf_{a \in \mathcal{A}} L(a, \theta) > -\infty$ for every $\theta \in \Theta$. For any P -certified decision $(\tilde{\delta}, \tilde{R})$ and any $\epsilon > 0$, there exists some $(\delta, R) \in \mathcal{C}_{P, \alpha, \epsilon}$ that weakly dominates $(\tilde{\delta}, \tilde{R})$.

Proof. Consider

$$\hat{\Theta}(Y) = \left\{ \theta \in \Theta : L(\tilde{\delta}(Y), \theta) \leq \tilde{R}(Y) \right\}.$$

Since $(\tilde{\delta}, \tilde{R})$ is P -certified, $\hat{\Theta}$ is a confidence set:

$$P(\theta \in \hat{\Theta}(Y)) = P(L(\tilde{\delta}(Y), \theta) \leq \tilde{R}(Y)) \geq 1 - \alpha$$

for all $(\theta, P) \in \mathcal{M}$.

Consider $(\delta, R) \in \mathcal{C}_{P, \alpha, \epsilon}$ such that

$$R(Y) \equiv \sup_{\theta \in \hat{\Theta}(Y)} L(\delta(Y), \theta) \leq \left(\inf_{a \in \mathcal{A}} \sup_{\theta \in \hat{\Theta}(Y)} L(a, \theta) + \epsilon \right) \wedge \sup_{\theta \in \hat{\Theta}(Y)} L(\tilde{\delta}(Y), \theta).$$

Such a choice exists:

(1) If $\hat{\Theta}(Y)$ is empty, then we set $R(Y) = -\infty$ and $\delta(Y) = \tilde{\delta}(Y)$. The above inequality holds vacuously.

(2) If $\hat{\Theta}(Y)$ is not empty, then $\inf_{a \in \mathcal{A}} \sup_{\theta \in \hat{\Theta}(Y)} L(a, \theta) > -\infty$ by assumption. If $\tilde{\delta}(Y)$ is already an ϵ -minimizer,

$$\sup_{\theta \in \hat{\Theta}(Y)} L(\tilde{\delta}(Y), \theta) \leq \inf_{a \in \mathcal{A}} \sup_{\theta \in \hat{\Theta}(Y)} L(a, \theta) + \epsilon,$$

then we may choose $\delta(Y) = \tilde{\delta}(Y)$. Otherwise, if $\tilde{\delta}(Y)$ is not an ϵ -minimizer, then we can choose some action with worst-case risk bounded by $\inf_{a \in \mathcal{A}} \sup_{\theta \in \hat{\Theta}(Y)} L(a, \theta) + \epsilon$.

By construction,

$$R(Y) \leq \sup_{\theta \in \hat{\Theta}(Y)} L(\tilde{\delta}(Y), \theta) \leq \tilde{R}(Y).$$

Since $R(Y)$ and $\tilde{R}(Y)$ are ordered almost surely, they are also ordered in the sense of stochastic dominance. This concludes the proof. $\square$

Theorem 4. Suppose $\Theta \subseteq \mathbb{R}$ is compact, $L(a, \theta)$ is weakly decreasing and right-continuous in θ for each a , and $R(\theta) = \inf_{a \in \mathcal{A}} L(a, \theta)$ is achieved by some $a \in \mathcal{A}$ for every $\theta \in \Theta$. Let $\hat{\theta}(Y)$ be a $1 - \alpha$ uniformly most accurate confidence lower bound, assumed to exist. Let (δ, R) be a certified decision that as-if optimizes against $[\hat{\theta}(Y), \infty) \cap \Theta$:

$$R(Y) \equiv L(\delta(Y), \hat{\theta}(Y)) = \inf_{a \in \mathcal{A}} L(a, \hat{\theta}(Y)).$$

Then for any other $1 - \alpha$ P -certified $(\tilde{\delta}, \tilde{R})$ and any $r > R(\theta)$,

$$P(R(Y) \geq r) \leq P(\tilde{R}(Y) \geq r).$$

Proof. Consider

$$\tilde{\theta}(Y) = \min \left\{ \theta \in \Theta : L(\tilde{\delta}(Y), \theta) \leq \tilde{R}(Y) \right\}$$

On the event that this set is empty, set $\tilde{\theta}(Y) = \max \Theta$. Otherwise, the min is attained by right-continuity and monotonicity. Note that $P(\tilde{\theta}(Y) \leq \theta) \geq P\{L(\tilde{\delta}(Y), \theta) \leq \tilde{R}(Y)\} \geq 1 - \alpha$, and thus $\tilde{\theta}(Y)$ is a $1 - \alpha$ confidence lower bound for θ .

Again let $R(\theta) = \inf_{a \in \mathcal{A}} L(a, \theta)$, and note that $R(Y) = R(\hat{\theta}(Y))$. If $\tilde{\theta}(Y)$ does not minimize over an empty set, then $R(\tilde{\theta}(Y)) \leq L(\tilde{\delta}(Y), \tilde{\theta}(Y)) \leq \tilde{R}(Y)$. If $\tilde{\theta}(Y)$ does minimize over an empty set, then $R(\tilde{\theta}(Y)) = R(\max \Theta) \leq R(\theta) < r$, so $\{R(\tilde{\theta}(Y)) \geq r\}$ can occur only if $\tilde{\theta}(Y)$ minimizes over a non-empty set. Decreasingness of $L(a, \theta)$ in θ implies that $R(\theta)$ is decreasing in θ .

Since $\hat{\theta}$ is a uniformly most accurate confidence lower bound, $P\{\hat{\theta}(Y) \leq \theta - t\} \leq P\{\tilde{\theta}(Y) \leq \theta - t\}$ for all $t > 0$. It follows that $P\{\hat{\theta}(Y) \wedge \theta \leq t\} \leq P\{\tilde{\theta}(Y) \wedge \theta \leq t\}$ for all $t \in \Theta$, and thus that $\hat{\theta}(Y) \wedge \theta$ first order stochastically dominates $\tilde{\theta}(Y) \wedge \theta$. Since first order stochastic dominance is preserved by monotone transformations, it follows that $R(\tilde{\theta}(Y) \wedge \theta) = R(\tilde{\theta}(Y)) \vee R(\theta)$ first order stochastically dominates $R(\hat{\theta}(Y) \wedge \theta) = R(\hat{\theta}(Y)) \vee R(\theta)$, and thus that for all $r > R(\theta)$,

$$P\{R(Y) \geq r\} = P\{R(\hat{\theta}(Y)) \geq r\} \leq P\{R(\tilde{\theta}(Y)) \geq r\} \leq P\{\tilde{R}(Y) \geq r\}.$$

This completes the proof. $\square$

Proposition 4. For any E -certified decision $(\tilde{\delta}, \tilde{R})$ and any $\epsilon > 0$, there exists some collection of e -variables $E(Y, \theta)$ for which the corresponding E -certified decision $(\delta, R) \in \mathcal{C}_{E, \epsilon}$ chosen via (13) weakly dominates $(\tilde{\delta}, \tilde{R})$.

Proof. Define

$$E(Y, \theta) = \frac{L(\tilde{\delta}(Y), \theta)}{\tilde{R}(Y)}$$

with the convention that $0/0 = 0$. Because $(\tilde{\delta}, \tilde{R})$ is E -certified, $\mathbb{E}_P[E(Y, \theta)] \leq 1$ for all $(\theta, P) \in \mathcal{M}$. Observe that for all θ ,

$$\frac{L(\tilde{\delta}(Y), \theta)}{E(Y, \theta)} \leq \tilde{R}(Y) \implies \sup_{\theta \in \Theta} \frac{L(\tilde{\delta}(Y), \theta)}{E(Y, \theta)} \leq \tilde{R}(Y) \implies \inf_{a \in \mathcal{A}} \sup_{\theta \in \Theta} \frac{L(a, \theta)}{E(Y, \theta)} \leq \tilde{R}(Y)$$

The inequality is due to cases where $L = 0$ but $\tilde{R} > 0$.

Note that we can pick $(\delta, R) \in \mathcal{C}_{E, \epsilon}$ such that

$$R(Y) := \sup_{\theta \in \Theta} \frac{L(\delta(Y), \theta)}{E(Y, \theta)} \leq \min \left\{ \inf_{a \in \mathcal{A}} \sup_{\theta \in \Theta} \frac{L(a, \theta)}{E(Y, \theta)} + \epsilon, \tilde{R}(Y) \right\}.$$

This is because, if $\tilde{R}(Y)$ is binding on the right-hand side, we can pick $\delta(Y) = \tilde{\delta}(Y)$. If $\inf_{a \in \mathcal{A}} \sup_{\theta \in \Theta} \frac{L(a, \theta)}{E(Y, \theta)} + \epsilon$ is binding, then we can simply choose any approximate optimum $\delta(Y)$.

By definition, $R(Y) \leq \tilde{R}(Y)$. This completes the proof. $\square$

Proposition 5. *Given an E -certified decision rule (δ, R) , for any adoption decision $Q \leq \mathbb{1}(R(Y) \leq C)$ a.s.,*

$$\mathbb{E}_P [L(\delta_Q(Y), \theta)] \leq \mathbb{E}_P \left[\max \left(\frac{L(\delta(Y), \theta)}{R(Y)}, 1 \right) C \right] \leq 2C \text{ for all } (\theta, P) \in \mathcal{M}$$

Proof. Note that, P -almost surely,

$$L(\delta_Q(Y), \theta) = \frac{L(\delta_Q(Y), \theta)}{QR(Y) + (1 - Q)C} (QR(Y) + (1 - Q)C) \leq \max \left(\frac{L(\delta(Y), \theta)}{R(Y)}, 1 \right) C.$$

The equality holds since $QR(Y) + (1 - Q)C > 0$ when $L > 0$ by [Lemma 5](#). The inequality follows from $Q \leq \mathbb{1}(R(Y) \leq C)$ and thus $QR + (1 - Q)C \leq C$.

Now,

$$\mathbb{E}_P \max \left(\frac{L(\delta(Y), \theta)}{R(Y)}, 1 \right) \leq 1 + \mathbb{E}_P [L/R] \leq 2.$$

This completes the proof. $\square$

Lemma 5. *Let $(\delta(\cdot), R(\cdot))$ be an E -certified decision. For all $(\theta, P) \in \mathcal{M}$,*

$$P(R(Y) = 0, L(\delta(Y), \theta) > 0) = 0.$$

Proof. Note that if not, then $\mathbb{E}[L/R] = \infty$ since $L > 0, R = 0$ with positive probability. This contradicts E -certification. $\square$

Appendix B. Modifying E-posteriors to optimize post-adoption bound

Let $\mathcal{C}_{E,1+\gamma}$ denote the following class of decision rules: For an e -variable $E(\theta, Y)$, define $E_\gamma(\theta, Y) = \gamma E(\theta, Y) + 1$,²⁵ and consider

$$\delta(Y) \in \arg \min_{a \in \mathcal{A}} \sup_{\theta \in \Theta} \frac{L(a, \theta)}{E_\gamma(\theta, Y)} \quad R(Y) = \min_{a \in \mathcal{A}} \sup_{\theta \in \Theta} \frac{L(a, \theta)}{E_\gamma(\theta, Y)}. \quad (18)$$

In this section, let us assume $\mathcal{A}$ is finite for simplicity, and thus the minimum always exists. Since $E_\gamma/(1 + \gamma)$ is an e -variable, such a decision rule is E-certified with multiple $1 + \gamma$.

We show that adoption rules with $Q \leq \mathbb{1}(R(Y) \leq C)$ entail risk at most $(1 + \gamma)C$ post adoption, in contrast to [Proposition 5](#). This guarantee holds because

$$\mathbb{E}_P \left[\max \left(\frac{L(\delta(Y), \theta)}{R(Y)}, 1 \right) \right] \leq 1 + \gamma.$$

Moreover, for all decisions (δ, R) satisfying the above, decision rules in $\mathcal{C}_{E,1+\gamma}$ are essentially complete with respect to weak dominance.

Proposition 6. *Given $(\delta, R) \in \mathcal{C}_{E,1+\gamma}$, we have that for any adoption decision $Q \leq \mathbb{1}(R(Y) \leq C)$,*

$$\mathbb{E}_P [L(\delta_Q(Y), \theta)] \leq \mathbb{E}_P \left[\max \left(\frac{L(\delta(Y), \theta)}{R(Y)}, 1 \right) C \right] \leq (1 + \gamma)C \text{ for all } (\theta, P) \in \mathcal{M}$$

Moreover, for any pair $(\tilde{\delta}, \tilde{R})$ such that $\mathbb{E}_P \left[\max \left(\frac{L(\tilde{\delta}(Y), \theta)}{\tilde{R}(Y)}, 1 \right) \right] \leq 1 + \gamma$, there exists some $(\delta, R) \in \mathcal{C}_{E,1+\gamma}$ that weakly dominates it.

Proof. Observe that

$$L(\delta_Q(Y), \theta) = \frac{L(\delta_Q(Y), \theta)}{QR(Y) + (1 - Q)C} (QR(Y) + (1 - Q)C) \leq \max \left(\frac{L(\delta(Y), \theta)}{R(Y)}, 1 \right) C.$$

Since

$$\mathbb{E}_P \left[\max \left(\frac{L(\delta(Y), \theta)}{R(Y)}, 1 \right) \right] \leq \mathbb{E}_P \left[\max \left(\frac{L(\delta(Y), \theta)}{L(\delta(Y), \theta)/E_\gamma(Y, \theta)}, 1 \right) \right] \leq \mathbb{E}_P [E_\gamma(Y, \theta)] \leq 1 + \gamma,$$

this proves the first claim.

²⁵We can also view $E_\gamma = (1 + \gamma) \left(\frac{\gamma}{1 + \gamma} E(\theta, Y) + \frac{1}{1 + \gamma} \right) = (1 + \gamma) S_\theta^{[\gamma/(1 + \gamma)]}(Y)$, where $S_\theta^{[c]}(Y) = cE(\theta, Y) + (1 - c)$ is proposed by [Grünwald \(2023\)](#) as a modification to e -variables that produces upper-bounded E-posteriors. Our subsequent result gives some guidance on how different choices of γ impact the resulting guarantee.

For the second claim, let

$$E_\gamma(Y, \theta) = \max \left(\frac{L(\tilde{\delta}(Y), \theta)}{\tilde{R}(Y)}, 1 \right).$$

Since its expectation is bounded by $1 + \gamma$, $(E_\gamma(Y, \theta) - 1)/\gamma$ is an e-variable. Thus there is some $(\delta, R) \in \mathcal{C}_{E, 1+\gamma}$ that corresponds to $E_\gamma(Y, \theta)$. Note that

$$R(Y) = \min_{a \in \mathcal{A}} \sup_{\theta \in \Theta} \frac{L(a, \theta)}{E_\gamma(Y, \theta)} \leq \sup_{\theta \in \Theta} \min \left(\tilde{R}(Y) \cdot L(\tilde{\delta}(Y), \theta) / L(\tilde{\delta}(Y), \theta), L(\tilde{\delta}(Y), \theta) \right) \leq \tilde{R}(Y).$$

This completes the proof. $\square$

Appendix C. Additional details of empirical applications

C.1. Variable-width confidence sets for the CMTO application. We could consider a variable-width version of $\hat{\Theta}_{\text{FW}}$ ²⁶

$$\hat{\Theta}_{\text{VW}} := \left\{ \vartheta \in \mathbb{R}^n : \begin{array}{l} (q - p)' \vartheta \geq (q - p)' Y - \lambda_\alpha \sum_i |q_i - p_i| \sigma_i \\ \text{for every } S \text{ with } |S| = m, p \in \text{ext } W(S), q \in \text{ext } V(S) \end{array} \right\}$$

where the critical value λ_α is chosen such that

$$\mathbb{P}_{\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)} \left\{ \max_{\substack{S \subseteq \{1, \dots, n\} \\ |S| = m}} \max_{\substack{p \in \text{ext } W(S) \\ q \in \text{ext } V(S), q \neq p}} \frac{(q - p)' \epsilon}{\sum_i |q_i - p_i| \sigma_i} \leq \lambda_\alpha \right\} \geq 1 - \alpha.$$

Here, using $\sum_i |q_i - p_i| \sigma_i$ as a proxy for the sampling noise in $(q - p)' Y$ rather than $\sqrt{\sum_i (q_i - p_i)^2 \sigma_i^2}$ is primarily motivated by computational tractability. Qualitatively, as-if optimization under $\hat{\Theta}_{\text{VW}}$ chooses tracts by penalizing those with high σ_i . We compute similar certificates in [Figure 5](#).

While the certificates are similar to those in [Figure 1](#), $\hat{\Theta}_{\text{VW}}$ and $\hat{\Theta}_{\text{FW}}$ make different recommendations when as-if optimized against. We can also use the two confidence sets to certify *each other's* recommendations in [Table 2](#). We see that the resulting certificates are different, indicating that the selected tracts have some discrepancy between the two confidence sets.

C.2. Details of as-if optimization with e-variables. In particular, with the choice of e-variables in [Section 4.1](#), the as-if optimization following [Grünwald \(2023\)](#)

²⁶Here, $\text{ext } V$ denotes the extreme points of a set V . For $V_\rho(S)$, the extreme points are those movement patterns supported on S with share exactly $1/(1 \vee \lceil \rho m \rceil)$ at exactly $1 \vee \lceil \rho m \rceil$ locations. For a fixed λ_α , removing ext in $\hat{\Theta}_{\text{VW}}$ results in the same as-if decisions as $\hat{\Theta}_{\text{VW}}$. But removing ext would demand a higher critical value, and thus would result in worse certificates than $\hat{\Theta}_{\text{VW}}$.

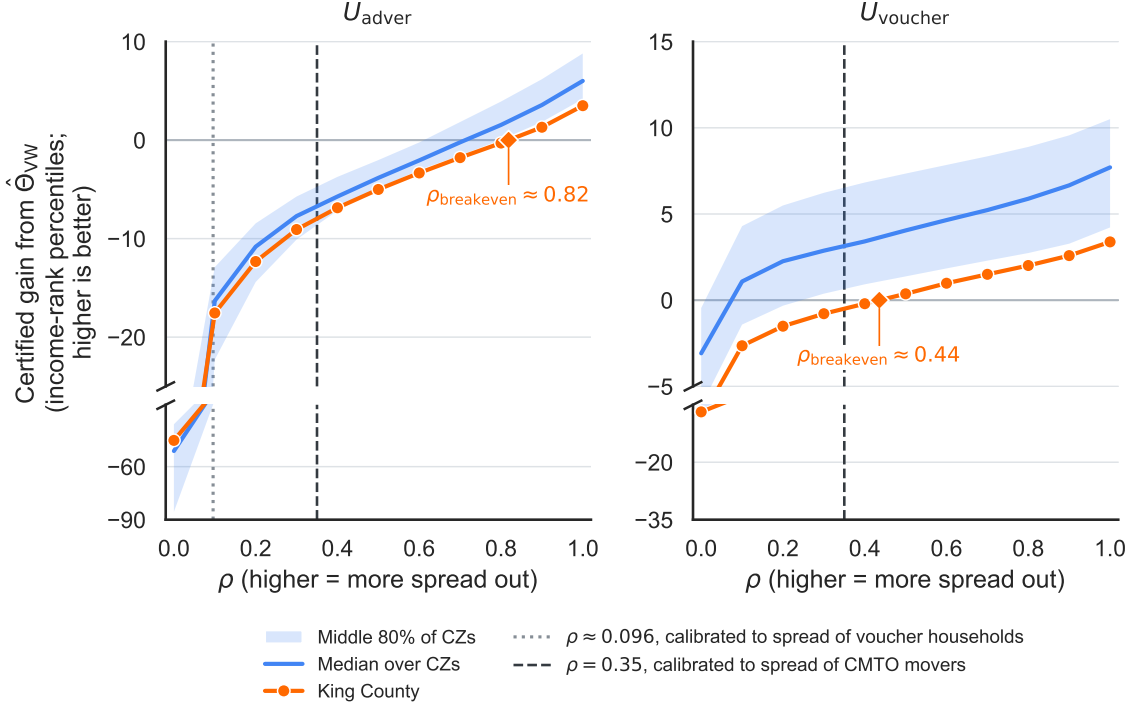


FIGURE 5. Analogue of Figure 1 using $\hat{\Theta}_{vw}$

simplifies considerably due to the observation that

$$\sup_{\theta} \frac{L(a, \theta)}{E(Y, \theta)} = \max_{k \neq a} \sqrt{\sigma_a^2 + \sigma_k^2} \sup_{t \geq \max_{j \neq k} \frac{Y_j - Y_k}{\sqrt{\sigma_j^2 + \sigma_k^2}}} \frac{t - \frac{Y_a - Y_k}{\sqrt{\sigma_a^2 + \sigma_k^2}}}{G_k(t)}$$

for a known non-decreasing function $G_k(\cdot)$.

This follows from two observations. First, if k is the best arm under θ , then $E(Y, \theta) = G_k(\max_{j \neq k} t_j(Y, \theta))$ for some non-decreasing $G_k(\cdot)$. Second, for a fixed k , let $t(\theta, k) = \max_{j \neq k} t_j(Y, \theta)$. We have that

$$\Theta = \bigcup_{k \in [n]} \bigcup_{t \geq \max_{j \neq k} \frac{Y_j - Y_k}{\sqrt{\sigma_j^2 + \sigma_k^2}}} \{\theta : t(\theta, k) = t, k(\theta) = k\}.$$

Maximizing $L(a, \theta)/E(Y, \theta)$ over the inner set of θ , applying the first observation, yields the closed form $G_k(t)^{-1} \sqrt{\sigma_a^2 + \sigma_k^2} \left(t - \frac{Y_a - Y_k}{\sqrt{\sigma_a^2 + \sigma_k^2}} \right)$.

TABLE 2. Certificates under $(\hat{\Theta}_{\text{FW}}, \hat{\Theta}_{\text{VW}})$ for recommended decisions using either confidence set

Selected set	FW certificate	VW certificate
<i>Median of 50 CZs, adversarial origins, $\rho = 0.1$</i>		
FW-selected	-13.8	-16.7
VW-selected	-14.4	-16.3
<i>King County, adversarial origins, $\rho = 0.1$</i>		
FW-selected	-14.9	-18.4
VW-selected	-15.9	-17.5
<i>Median of 50 CZs, voucher origins, $\rho = 0.1$</i>		
FW-selected	-1.6	-1.7
VW-selected	-3.0	1.1
<i>King County, voucher origins, $\rho = 0.1$</i>		
FW-selected	-4.5	-5.4
VW-selected	-5.9	-2.6
<i>Median of 50 CZs, adversarial origins, $\rho = 0.5$</i>		
FW-selected	-3.0	-3.9
VW-selected	-3.1	-3.8
<i>King County, adversarial origins, $\rho = 0.5$</i>		
FW-selected	-4.6	-5.1
VW-selected	-4.8	-5.0
<i>Median of 50 CZs, voucher origins, $\rho = 0.5$</i>		
FW-selected	4.5	3.1
VW-selected	3.8	4.0
<i>King County, voucher origins, $\rho = 0.5$</i>		
FW-selected	0.4	-0.6
VW-selected	-0.5	0.4