Quantitative Economics 0 (2025), 1–29 1759-7331/20250001

Comprehensive OOS evaluation of predictive algorithms with statistical decision theory

Jeff Dominitz
Department of Economics and Ken Kennedy Institute, Rice University

Charles F. Manski
Department of Economics and Institute for Policy Research, Northwestern University

We argue that comprehensive out-of-sample (OOS) evaluation using statistical decision theory (SDT) should replace the current practice of K-fold and Common Task Framework validation in machine learning (ML) research on prediction. SDT provides a formal frequentist framework for performing comprehensive OOS evaluation across all possible (1) training samples, (2) populations that may generate training data, and (3) populations of prediction interest. Regarding feature (3), we emphasize that SDT requires the practitioner to directly confront the possibility that the future may not look like the past and to account for a possible need to extrapolate from one population to another when building a predictive algorithm. For specificity, we consider treatment choice using conditional predictions with alternative restrictions on the state space of possible populations that may generate training data. We discuss application of SDT to the problem of predicting patient illness to inform clinical decision making. SDT is simple in abstraction, but it is often computationally demanding to implement. We call on ML researchers, econometricians, and statisticians to expand the domain within which implementation of SDT is tractable.

Keywords. Machine learning, conditional prediction, Wald statistical decision theory.

JEL classification. C44.

1. Introduction

As hopes and fears related to artificial intelligence (AI) continue to grow, it is important to recognize that the impact of any AI system will always be tied to the quality of the predictive algorithms on which the AI is based. Throughout the 20th century, evaluation of predictions made with sample data was primarily the domain of statisticians and econometricians, who use frequentist or Bayesian statistical theory to propose and evaluate probabilistic prediction methods. In the 21st century, prediction is increasingly per-

Jeff Dominitz: jeff.dominitz@rice.edu
Charles F. Manski: cfmanski@northwestern.edu
We have benefitted from the opportunity to present this work in seminars at Northwestern University and the University of Chicago, and at conferences at Brown University and Cornell University. We have further benefited from the comments of reviewers on earlier drafts.

 https://doi.org/10.3982/QE2728

formed by researchers in computer science, who view prediction methods as computational algorithms and who rarely use statistical theory to assess the algorithms. Instead, these researchers commonly perform *out-of-sample* (OOS) tests of predictive accuracy.

In an influential early article, Breiman (2001a) argued for this approach, writing (p. 201): "Predictive accuracy on test sets is the criterion for how good the model is." He asserted that OOS evaluation is close to theory-free, stating (p. 205): "the one assumption made in the theory is that the data is drawn i.i.d. from an unknown multivariate distribution." Although Breiman did not state it explicitly, we conjecture that he had in mind settings where the data are drawn from the distribution of prediction interest, not from just *any* unknown distribution.

Over the past two decades, OOS evaluation has often been applied even without this basic assumption. Using the now familiar term *machine learning* (ML) to describe prevalent prediction practices in computer science, Taddy (2019) remarked on the wide adoption of the approach advocated by Breiman, writing (p. 70): "Machine learning's wholesale adoption of OOS validation as the arbitrator of model quality has freed the ML engineer from the need to *theorize* about model quality."

The term *machine learning* has been used in multiple ways, with occasional lack of clarity. Considered broadly, ML means any computational procedure using data to make predictions. This broad usage encompasses the host of parametric and nonparametric probabilistic prediction methods developed by statisticians and econometricians over the past 150 years, studied through the lens of frequentist or Bayesian statistical theory. The recent usage in computer science is narrower, focusing on certain nonparametric methods (e.g., random forests and deep neural networks), whose predictive properties have been assessed by OOS evaluation. In this paper, we mainly use the term in the narrow sense, as does Taddy.[1]

Taddy described OOS evaluation this way (p. 70):

> "Out-of-sample validation is a basic idea: you choose the best model specification by comparing predictions from models estimated on data that was not used during the model 'training' (fitting). This can be formalized as a cross-validation routine: you split the data into K 'folds,' and then K times fit the model on all data but the K^{th} fold and evaluate its predictive performance (e.g., mean squared error or misclassification rate) on the left-out fold. The model with optimal average OOS performance (e.g., minimum error rate) is then deployed in practice."

Cross-validation (CV) has long been used by statisticians to choose tuning parameters for prediction with nonparametric and semiparametric regression methods.

[1]An apt example of lack of clarity in usage concerns *empirical risk minimization* (*ERM*) as studied by Vapnik and others in the latter part of the 1990s; see Vapnik (1999, 2000). Computer scientists frequently embrace this methodology within the domain of machine learning. However, it firmly is a development of frequentist statistical theory, grounded in limit theorems regarding consistency and rates of convergence. Indeed, ERM is a direct descendant of classical statistical analysis of the method of moments, M-estimation, and minimum distance estimation. All of these methods apply what Goldberger (1968) called the *analogy principle*, using the empirical distribution of a training sample to estimate a population distribution under the assumption of random sampling, relying on a law of large numbers for their basic convergence properties. Computer scientists do not ordinarily refer to these classical statistical methods as machine learning, yet they apply this term to empirical risk minimization.

Statisticians have used frequentist asymptotic statistical theory to evaluate prediction methods with cross-validated tuning parameters—studying their consistency, rates of convergence, and limiting distributions (e.g., Zhang (1993)). The use of CV in OOS evaluation described by Taddy differs. Here, one evaluates predictive performance on alternative subsets of the realized training sample, without considering how the method would perform across repeated draws of the training sample. Thus, K-fold OOS CV does not use frequentist statistical theory. Note, however, that Bates, Hastie, and Tibshirani (2024) have recently offered a frequentist statistical perspective on OOS CV.

Rather than use a K-fold split of the training data to form multiple validation samples, OOS evaluation is sometimes performed in the so-called 'Common Task Framework' (CTF). Here, predictive accuracy is measured in some pre-specified testing data that may differ systematically from the training sample. Donoho (2017) described the CTF as follows:

> "An instance of the CTF has these ingredients: (a) A publicly available training dataset involving, for each observation, a list of (possibly many) feature measurements, and a class label for that observation. (b) A set of enrolled competitors whose common task is to infer a class prediction rule from the training data. (c) A scoring referee, to which competitors can submit their prediction rule. The referee runs the prediction rule against a testing dataset, which is sequestered behind a Chinese wall. The referee objectively and automatically reports the score (prediction accuracy) achieved by the submitted rule. All the competitors share the *common task* of training a prediction rule, which will receive a good score; hence the phase *common task framework*. A famous recent example is the Netflix Challenge, where the common task was to predict Netflix user movie selections."

CTF evaluation is the same as 1-fold CV when the testing dataset behind the wall is a hold-out subsample of the training sample. However, the designated common task often is to predict an outcome in a testing dataset not generated from the training sample. The practice of holding competitions with awards made using the CTF is remote from the concern of our paper, which is decision making by a single agent.[2]

The intuitive appeal of K-fold and CTF OOS evaluation has aided computer scientists in their efforts to market ML predictive algorithms to medical clinicians, judges and police, firm managers, and others who may lack expertise in statistical theory but who feel that they can appraise prediction accuracy heuristically. However, it should be obvious that these types of OOS evaluation cannot yield generalizable lessons. We acknowledge that innovation without theory has sometimes been productive throughout human history, but we are concerned that it sometimes leads society astray.

Caution was expressed early on by Cox (2001), whose published Comment on Breiman (2001a) distinguished between cases where (pp. 216–217): "prediction is localized to situations directly similar to those applying to the [*training*] data" and those

[2] In a recent commentary that enthusiastically portrays the current state of data science, Donoho (2024) asserts that challenge competitions using the CTF have contributed to advances in machine learning. As we interpret him, he conjectures an evolutionary process of "survival of the fittest," whereby competitive challenges stimulate productive innovation. We think this an intriguing conjecture that may warrant study as a topic in the economics of research and development. We are not aware of existing knowledge that corroborates or invalidates the conjecture.

where "prediction is under quite different conditions from the data." On the former, Cox concluded that the algorithmic "empirical black-box approach" may be preferable to explicit modeling of the data generating process, whereas, in the latter case "prediction, always hazardous, without some understanding of underlying process and linking with other sources of information, becomes more and more tentative." In his Rejoinder, Breiman (p. 227) replied to the latter statement by stating, simply, "I agree."

More recently, Taddy (2019) wrote (p. 62): "Machine learning can do fantastic things, but it is basically limited to predicting a future that looks mostly like the past." Poggio and Fraser (2024) caution that the popular ML approach of fitting training data to deep neural networks should be expected to yield good predictions of an outcome y conditional on covariates x only if the true best predictor of interest (often the conditional mean function) is "compositionally sparse;" that is, if it has a structure similar to that of deep neural network functions. We share the important concerns of Cox, Taddy, and Poggio and Fraser.

Efron (2020), in an article contrasting the perspectives of statisticians and computer scientists, wrote (p. S49): "In place of theoretical criteria, various prediction competitions have been used to grade algorithms in the so-called 'Common Task Framework.' ... None of this is a good substitute for a so-far nonexistent theory of optimal prediction." We agree with Efron that the prediction competitions of the CTF are not a satisfactory way to evaluate prediction methods. However, we sharply disagree with Efron's second statement. To reiterate what Manski (2023) observed recently (p. 648):

> "[Efron] was not correct when he stated that a theory of optimal prediction is 'so-far nonexistent'. Wald (1939, 1945, 1950) considered the general problem of using sample data to make decisions. He posed the task as choice of a *statistical decision function*, which maps potentially available data into a choice among the feasible actions. His development of statistical decision theory provides a broad framework for decision making with sample data, yielding optimal decisions when these are well-defined and proposing criteria for reasonable decision making more generally."

Wald recommended ex ante (frequentist) evaluation of statistical decision functions as *procedures* to be applied as a sampling process is engaged repeatedly to draw independent data samples. Statistical decision theory (SDT) measures the mean performance of a prediction method across all possible training samples, when the objective is to predict outcomes in a population that may possibly differ from the one generating the data in the training sample.

We emphasize that SDT performs ex ante evaluation. This is directly at odds with the standard practice of ML researchers to perform ex post evaluation, after observation of a particular training sample and without regard to other training samples that might have been drawn. Ex ante evaluation is the standard practice in frequentist statistical theory, which evaluates estimates by their sampling distributions, using criteria including consistency, unbiasedness, and efficiency.

SDT is remote from the types of OOS evaluation currently practiced by ML researchers, but it is not remote conceptually. Indeed, SDT provides a formal framework for performing what we shall term *comprehensive OOS* evaluation. SDT performs OOS evaluation across all possible (1) training samples, (2) populations that may generate

training data, and (3) populations of prediction interest. We think that feature (3) is particularly important. SDT requires the practitioner to directly confront the possibility that the future may not look like the past and to account for a possible need to extrapolate from one population to another when building a predictive algorithm.

Even in the best-case scenario where the future will credibly look like the past—that is, no extrapolation problem—SDT provides a framework for OOS evaluation that is scientifically more rigorous than the current standard practice of K-fold and CTF validation. When the training sample is known to be drawn i.i.d. from the distribution of prediction interest, one must recognize that this sample is just one draw among all possible samples from this population distribution. With comprehensive OOS evaluation, the performance of the predictive algorithm is assessed across all samples on which it could be trained, rather than just the one sample on which it actually is trained. Thus, SDT embraces the paradigm of frequentist statistical theory.

ML researchers often motivate their versions of OOS evaluation by stating that it protects against drawing misleading conclusions from prediction performance on the training sample, which may be unrealistically high due to so-called "overfitting" the data. Concern with overfitting does not arise in the Wald framework because it evaluates performance across all feasible samples, not a particular training sample. Whereas some may believe that the so-called "big data" revolution renders attention to sampling variation unnecessary, the desire to build ML models with high-dimensional covariates gives rise to the same finite sample concerns that SDT was developed to address. Even if the total sample size is orders of magnitude larger than Wald may have envisioned in the 1940s, statistical imprecision is an important consideration when conditioning on covariates whose distribution has sufficiently large support.

In this paper, we argue that comprehensive OOS evaluation performed using SDT should replace the current practice of K-fold and CTF validation in ML research. The argument is simple to make in abstraction. As we explain in Section 2, the Wald framework is general, transparent, and intellectually attractive. In principle, it enables comparison of essentially all predictive algorithms, the only caveat being satisfaction of weak mathematical regularity conditions. It enables comparison of alternative sampling processes generating training data. It uses no asymptotic approximations when interpreting the training data. Further, the population of interest in prediction may differ from that generating the training data.

The argument is more challenging to make in practice. Application of SDT is easy in some important settings, but it is often computationally demanding. Computation commonly requires numerical methods to find approximate solutions. Moreover, SDT requires the practitioner to specify a decision criterion. Among the criteria that have been studied, we find minimax regret particularly appealing. A statistical decision function (SDF) with small maximum regret is uniformly near-optimal across all possible populations generating training data and populations of prediction interest. In the terminology of SDT, a possible pair of such populations is a *state of nature*. The set of all possible populations is the *state space*. Maximum regret is computed across the state space.

To go beyond abstraction, Sections 3 and 4 consider treatment choice using a prediction of an outcome y conditional on a specified covariate value x. To focus on the

problem of statistical imprecision, we suppose that the population of prediction interest is the same as that generating the training data. The central issue is specification of cross-covariate restrictions on conditional outcome distributions, which enable outcome data associated with covariate values $x' \neq x$ to be informative about $P(y|x)$.

Statistical and econometric theory has long studied settings in which cross-covariate restrictions are imposed by assuming all conditional distributions, or their means, lie in a specified finite-dimensional set (parametric regression) or a specified space of suitable smooth functions (nonparametric regression). The ML literature using deep neural networks as predictor functions has recently favored assumptions that conditional probability distributions or means are compositionally sparse, which Poggio and Fraser (2024) described as (p. 1): "a key principle underlying successful learning architectures." Yet other cross-covariate restrictions on conditional distributions are specified in the bounded-variation assumptions studied by Manski (2023). We do not take a stand favoring any one class of restrictions on the state space over another—the choice should be application specific. Our theme is that SDT may in principle be used to perform comprehensive OOS evaluation, however a researcher chooses to restrict the state space.

To provide an example, we discuss the important problem of predicting patient illness to inform clinical decision making. We summarize the history of probabilistic prediction by biostatisticians and computer scientists. We then focus on the problem of choice between surveillance and aggressive treatment of a patient whose illness status is not observed but can be predicted probabilistically.

Past econometric research applying statistical decision theory

Readers wanting to learn about the early development of statistical decision theory may want to begin with the highly informative monographs of Ferguson (1967) and Berger (1985). Econometric research applying SDT began in the mid-20th century (e.g., Sawa and Hiromatsu (1973)), but it has become common more recently. Much of the focus has been on minimax-regret treatment choice with data from classical randomized trials or quasiexperiments, where the distribution of treatment response is point-identified and the sole concern is statistical imprecision. Contributions include Manski (2004, 2005, 2019), Hirano and Porter (2009, 2020), Stoye (2009), Manski and Tetenov (2016, 2019, 2021), Kitagawa and Tetenov (2018), and Mbakop and Tabord-Meehan (2021).[3]

Attention is increasingly being given to treatment choice with data from trials with imperfect internal validity or from observational studies, where identification is the dominant concern with large sample size and joint attention to identification and statistical imprecision is essential when sample size is small. Contributions include Athey and Wager (2021), Manski (2000, 2007, 2021, 2023), Montiel Olea, Qiu, and Stoye (2024), Stoye (2012), and Yata (2023). Dominitz and Manski (2017, 2022, 2025) studied prediction under square loss with nonrandomly missing data.

[3]With the exception of Hirano and Porter (2009, 2020), these studies apply Wald's finite-sample theory. Hirano and Porter use the local asymptotic theory of Hajek and LeCam to circumvent the computational complexity often encountered when implementing the Wald finite-sample theory. At present, little is known about the adequacy of the the Hajek–LeCam theory to approximate the finite-sample settings that occur in applied research.

2. The statistical decision theory perspective on prediction

When Wald formally considered the general problem of using sample data to make decisions, this included the use of data to make predictions. We present the general structure of SDT in Section 2.1. Section 2.2 explains how SDT views prediction—in particular, SDFs that use predictions to make decisions. Section 2.3 juxtaposes statistical decision theory and research on robust decisions. This exposition draws on Manski (2021, 2023, 2024).

2.1 *Statistical decision theory in abstraction*

Wald utilized the standard decision-theoretic framing of the choice problem of a decision maker (DM) who must choose an action yielding loss that depends on an unknown state of nature. The DM specifies a state space listing the states considered possible. The DM wants to minimize loss, but must choose without knowing the true state.

To begin, Section 2.1.1 presents this problem in a setting where the DM does not observe sample data. Section 2.1.2 explains how Wald generalized this setting by supposing that the DM observes sample data that may be informative about the true state. Section 2.1.3 discusses computation.

Wald's first major contribution was to recast the hypothesis testing problem previously posed by Neyman and Pearson (1928, 1933) as a decision problem in which one must choose between two feasible actions. ML researchers now refer to this type of decision as a binary *classification problem*. We summarize Wald's formalization of the classification problem in the Appendix.

2.1.1 *Decisions without sample data* Consider a DM who faces a predetermined choice set C and believes that the true state s^* lies in state space S. The state space may be finite-dimensional (parametric) or larger (nonparametric). Objective function $L(\cdot,\cdot): C \times S \to R^1$ maps actions and states into loss. The DM ideally would minimize $L(\cdot, s^*)$ over C, but the DM does not know s^*.

It is generally accepted that choice should respect dominance. Action $c \in C$ is weakly dominated if there exists a $d \in C$ such that $L(d,s) \leq L(c,s)$ for all $s \in S$ and $L(d,s) < L(c,s)$ for some $s \in S$. To choose among undominated actions, decision theorists have proposed various ways of using $L(\cdot,\cdot)$ to form functions of actions alone, which can be optimized.[4]

Wald (1945) studied choice when the DM places a subjective probability distribution π on the state space, averages state-dependent loss with respect to π, and minimizes subjective average loss $\int L(c,s)\,d\pi$ over C. The criterion solves

$$\min_{c \in C} \int L(c,s)\,d\pi. \tag{1}$$

[4]In principle, one should only consider undominated actions, but it often is difficult to determine which actions are undominated. Hence, in practice it is common to optimize over the full set of feasible actions. We define decision criteria accordingly. We use max and min notation, without concern for the subtleties that sometimes make it necessary to use sup and inf operations.

Wald (1945, 1950) considered choice when the DM does not place a subjective distribution on the state space. In this setting, he studied minimax choice, which selects an action that works uniformly well over all of S in the sense of minimizing the maximum loss attainable across S. The minimax criterion is

$$\min_{c\in C}\max_{s\in S} L(c,s). \tag{2}$$

Savage (1951), in a book review of Wald (1950), suggested a different formalization of the idea of selecting an action that works uniformly well over all of S. This formalization, which has become known as the minimax-regret (MMR) criterion, solves the problem

$$\min_{c\in C}\max_{s\in S}\Big[L(c,s)-\min_{d\in C} L(d,s)\Big]. \tag{3}$$

Here, $L(c,s)-\min_{d\in C} L(d,s)$ is the *regret* of action c in state s—that is, the increment to loss in state s arising from making choice c rather than the optimal choice in that state. The true state being unknown, one evaluates c by its maximum regret over all states and selects an action that minimizes maximum regret. The maximum regret of an action measures its maximum distance from optimality across states.[5]

2.1.2 *Statistical decision problems* Statistical decision problems add to the above structure by supposing that the DM observes data generated by some sampling distribution. Knowledge of the sampling distribution is generally incomplete. To express this, one extends state space S to list the feasible sampling distributions, denoted $(Q_s, s\in S)$. Let Ψ_s denote the sample space in state s; Ψ_s is the set of samples that may be drawn under distribution Q_s. The literature typically assumes that the sample space does not vary with s and is known. We assume this and denote the sample space as Ψ. Then a statistical decision function, $c(\cdot):\Psi\to C$, maps the sample data into a chosen action. Henceforth, $\psi\in\Psi$ is a possible realization of the sample data; that is, a possible training sample.

Wald's concept of a statistical decision function embraces all mappings [data → action]. SDF $c(\cdot)$ is a deterministic function after realization of the sample data, but it is a random function ex ante. Hence, the loss achieved by $c(\cdot)$ is a random variable ex ante. Wald's theory evaluates the performance of $c(\cdot)$ in state s by $Q_s\{L[c(\psi),s]\}$, the ex ante distribution of loss that it yields across realizations ψ of the sampling process.

It remains to ask how DMs might compare the loss distributions yielded by different SDFs. DMs want to minimize loss, so it seems self-evident that they should prefer SDF d(·) to c(·) in state s if $Q_s\{L[d(\psi),s]\}$ is stochastically dominated by $Q_s\{L[c(\psi),s]\}$. It is

[5]Criteria (1)–(3) have become the most prominent studied in decision theory, but they are not alone. Hurwicz (1951) suggested minimization of a weighted average of worst and best possible outcomes. Rather than assert a complete subjective distribution on the state space or none, a DM might assert a partial subjective distribution, placing lower and upper probabilities on states. One then might minimize maximum subjective average loss or minimize maximum subjective average regret. These criteria combine elements of averaging across states and concern with uniform performance across states. It appears that this idea was first suggested by Hurwicz (1951). The idea was later taken up by statistical decision theorists. See Berger (1985) and Walley (1991).

less obvious how they should compare SDFs whose loss distributions do not stochastically dominate one another.

Wald proposed measurement of the performance of c(·) in state s by its expected loss across samples; that is, $E_s\{L[c(\psi), s]\} \equiv \int L[c(\psi), s]\, dQ_s$. He used the term *risk* to denote the mean performance of an SDF across samples. An alternative that has drawn only slight attention measures performance by quantile loss (Manski and Tetenov (2023)).

Not knowing the true state, a DM evaluates $c(\cdot)$ by the expected loss vector $(E_s\{L[c(\psi), s]\}, s \in S)$. Using the term *inadmissible* to denote weak dominance when evaluating performance by risk, Wald recommended elimination of inadmissible SDFs from consideration. As in decisions without sample data, there is no clearly best way to choose among admissible SDFs. SDT has mainly studied the same criteria as has decision theory without sample data. Let Γ be a specified set of SDFs, each mapping $\Psi \to C$. The statistical versions of criteria (1), (2), and (3) are

$$\min_{c(\cdot)\in\Gamma} \int E_s\big\{L\big[c(\psi), s\big]\big\}\, d\pi, \tag{4}$$

$$\min_{c(\cdot)\in\Gamma} \max_{s\in S} E_s\big\{L\big[c(\psi), s\big]\big\}, \tag{5}$$

$$\min_{c(\cdot)\in\Gamma} \max_{s\in S}\Big(E_s\big\{L\big[c(\psi), s\big]\big\} - \min_{d\in C} L(d, s)\Big). \tag{6}$$

Each of criteria (4)–(6) has been deemed reasonable by some decision theorists, but each has also drawn criticism. Minimization of subjective average loss (4) (aka Bayes risk) may be appealing if one has a credible basis to place a subjective probability distribution on the state space, but not otherwise. Concern with specification of priors motivated Wald to study the minimax criterion (5). However, minimax was criticized by Savage (1951) as "ultrapessimistic." Manski (2004, 2021) argued for the MMR criterion (6), observing that an SDF with small maximum regret is uniformly near-optimal across all states.

Whichever criteria (4)–(6) one uses, SDT performs a comprehensive OOS evaluation of an SDF. In each state of nature s, the expected loss $E_s\{L[c(\psi), s]\}$ of SDF $c(\cdot)$ measures its performance across all possible training samples. The state-dependent vector $\{E_s\{L[c(\psi), s]\}, s \in S\}$ of expected loss measures performance across all possible populations that may have generated the training sample and all possible populations of decision interest. Direct measurement of performance across all possible populations replaces any need for test samples to be drawn.

Recall that, when arguing for his narrow form of OOS evaluation, Breiman (2001a) stated that "the one assumption made in the theory is that the data is drawn i.i.d. from an unknown multivariate distribution." This assumption is unnecessary in SDT. Any sampling distribution can generate the data.

2.1.3 *Computation* Subject to regularity conditions ensuring that the relevant expectations and extrema exist, problems (4)–(6) offer criteria for decision making with sample data that are broadly applicable in principle. The primary challenge is computational. Problems (4)–(6) have tractable analytical solutions only in certain cases. Computation commonly requires numerical methods to find approximate solutions.

Expected loss $E_s\{L[c(\psi), s]\}$ typically does not have an explicit form, but it can be well approximated by Monte Carlo integration. One draws repeated values of ψ from distribution Q_s, computes the average value of $L[c(\psi), s]$ across the values drawn, and uses this to estimate $E_s\{L[c(\psi), s]\}$. Monte Carlo integration can also be used in criterion (4) to approximate the subjective average of expected loss.

The main computational challenges are determination of the extrema across actions in problem (6), across states in problems (5)–(6), and across SDFs in problems (4)–(6). Solution of $\min_{d\in C} L(d, s)$ in (6) is often straightforward but sometimes difficult. Finding extrema over S must cope with the fact that the state space commonly is uncountable. In applications where the quantity to be optimized varies smoothly over S, a simple approach is to compute the extremum over a suitable finite grid of states.

The most difficult computational challenge usually is to optimize over the feasible SDFs. No generally applicable approach is available. Hence, applications of SDT necessarily proceed case-by-case. It may not be tractable to find the best feasible SDF, but one often can evaluate the performance of relatively simple SDFs that researchers use in practice.

2.2 *Prediction-based SDFs*

We have observed that Wald's concept of an SDF embraces all mappings [data → action]. This very broad domain includes SDFs that make predictions and use the predictions to make decisions. These have the form [data → prediction → action], first making a prediction and then using the prediction to make a decision. There seems to be no accepted term for such SDFs, so we call them *prediction-based* SDFs.

To formalize prediction-based SDFs, we first need to formalize the concept of prediction and embed it in a decision problem. To do this, we bring to bear the venerable idea of probabilistic conditional prediction. This idea postulates a population characterized by a joint probability distribution $P(y, x)$, where (y, x) takes values in some set $Y \times X$. A member is drawn at random from the subpopulation with a specified value of x. Then the conditional distribution $P(y|x)$ provides the ideal probabilistic prediction of y conditional on x.

We embed prediction in a decision problem by considering settings in which a value of x is specified and the state space contains a set of feasible conditional distributions $P(y|x)$. Then $P^*(y|x)$ is the unknown distribution of prediction interest. The DM wants to minimize $L[\cdot, P^*(y|x)]$ over C, but does not know $P^*(y|x)$. One faces a statistical decision problem if one observes data ψ drawn from some sampling process.

In a setting of this type, an SDF is prediction-based if the DM uses the data ψ to form an estimate of $P^*(y|x)$ and acts as if the estimate is accurate. Let $f(\cdot) : \Psi \to S$ be the predictor function used to estimate $P^*(y|x)$. The chosen action with this SDF is $d(\psi) = \arg\min_{c\in C} L[c, f(\psi)]$.

Rather than estimate the entire conditional distribution $P^*(y|x)$, researchers often use sample data to form a point prediction of outcome y. Let $p(\cdot) : \Psi \to Y$ denote a predictor function that generates a point prediction. The chosen action with an SDF of this type acts as if $P^*(y|x)$ is degenerate, with all its mass at the outcome value $p(\psi)$. Thus, the SDF is $d(\psi) = \arg\min_{c\in C} L[c, p(\psi)]$.

Important special cases occur when the choice set C coincides with the set Y of potential outcomes. Let y be real-valued and let loss be mean square error (MSE) when using c to predict y. Then $C = Y = R$ and

$$L\big[c, P(y|x)\big] = E\big[(y-c)^2|x\big] = \int (y-c)^2\, dP(y|x). \tag{7}$$

Or y may be binary, and loss may be the misclassification rate (MCR) when using c to predict y. Then $C = Y = \{0, 1\}$ and

$$L\big[c, P(y|x)\big] = P(y \neq c|x) = \int 1[y \neq c]\, dP(y|x). \tag{8}$$

Observe that $1[y \neq c] = (y-c)^2$ when $C = \{0, 1\}$, but not otherwise. See the Appendix for further discussion.

From the perspective of SDT, the performance of a prediction-based SDF should be evaluated in the same manner as any SDF. Thus, one should first determine if the SDF is undominated. If so, one may evaluate it by the subjective expected loss it yields, the maximum loss it yields, by its maximum regret, or by some other reasonable criterion.

2.3 *Robust decisions*

To conclude this introduction to statistical decision theory, we explain its relationship to research on *robust decisions*. The field of robust statistics began with study of prediction when data contamination makes the probability distribution generating the training data lie in a neighborhood of the distribution of prediction interest (e.g., Huber (1981), Hampel, Elvezio, Rousseeuw, and Stahel (1986), Horowitz and Manski (1995)). Subsequently, engineers, economists, and computer scientists have studied mathematically similar but distinct problems of robust decision making in which a researcher specifies a probabilistic model space and assumes that the true probability distribution generating outcomes lies in a neighborhood of the model space (e.g., Zhou, Doyle, and Glover (1995), Hansen and Sargent (2008)). Blanchet, Li, Lin, and Zhang (2024) refers to this as *distributional shift*.

Study of robust decisions has proceeded in a different manner than statistical decision theory. Rather than initially specify a state space that lists all possible states of nature, the researcher initially poses a model space. The model specifies a single state or a relatively small set of states, often a finite-dimensional family. Having posed the model space, a researcher may be concerned that it does not contain the true state; that is, the model may not be correct. To recognize this possibility, the researcher enlarges the model space locally, using some metric to generate a neighborhood thereof. He then acts as if the locally enlarged model space is correct. Watson and Holmes write (p. 465): "We then consider formal methods for decision making under model misspecification by quantifying stability of optimal actions to perturbations to the model within a neighbourhood of [the] model space."

Although research on robust decisions differs procedurally from statistical decision theory, one can subsume the former within the latter if one considers the locally enlarged model space to be the state space. It is unclear how often this perspective characterizes what researchers have in mind. Articles often do not state explicitly whether the constructed neighborhood of the model space encompasses all states that authors deem sufficiently feasible to warrant consideration. The models specified in robust decision analyses often make strong assumptions and generated neighborhoods often relax these assumptions only modestly.

3. Conditional prediction with cross-covariate restrictions on the state space

In principle, SDT may be used to perform comprehensive OOS evaluation of any of the huge variety of probabilistic conditional prediction methods that have been developed from the late 1800s onward by statisticians, econometricians, and ML researchers. In practice, SDT has rarely been used to evaluate conditional prediction methods by researchers in these fields and applied practitioners. Prediction has mainly been studied under the assumption that the training data are a random sample drawn from the distribution $P(y, x)$ of prediction interest, or at least that $P(y|x)$ generates the training realizations of y conditional on x. The literature on Bayesian prediction predominately uses the *conditional Bayes* paradigm. This centers attention on the posterior predictive distribution, computed after training data are observed, not on ex ante Bayes risk as in SDT. There has been little minimax or MMR evaluation of conditional prediction methods; isolated cases of MMR evaluation of linear regression are Sawa and Hiromatsu (1973) and Chamberlain (2000). Research on nonparametric regression has mainly studied asymptotic properties of estimates when the covariate distribution has positive density in a neighborhood of a value of interest and the conditional expectation $E(y|x)$ varies smoothly with x in a local sense, such as differentiability. Donoho, Johnstone, Kerkyacharian, and Picard (1995) reviews many findings. An isolated instance of MMR evaluation of kernel nonparametric regression is Manski (2023), which bounds the global rather than local variation of $E(y|x)$ with x. See Section 3.2 below for discussion.

Whereas statisticians and econometricians have used statistical theory to study prediction methods, ML researchers have largely performed K-fold and CTF OOS validation. Absent statistical theory, ML researchers assert that K-fold and CTF validation exercises show that certain favored methods commonly "work" in practice. Yet the reasons why these methods "work" and the circumstances in which they "work" continue to be controversial. This is clear from a widely cited report of Bommasani, Hudson, Adeli et al. (2021), who concluded (p. 1): "Despite the impending widespread deployment of foundation models, we currently lack a clear understanding of how they work, when they fail, and what they are even capable of."

We are aware of only a few frequentist statistical studies of ML methods. Schmidt-Hieber (2020) studied the rates of convergence of deep neural network predictors when conditional means are compositionally sparse. Bartlett, Montanari, and Rakhlin (2021) used asymptotic statistical theory to study the predictive accuracy of deep learning

methods that involved "benign overfitting." Bates, Hastie, and Tibshirani (2024) provided a frequentist analysis of measures of prediction error when certain prediction models are evaluated using CV methods. These recent frequentist studies are constructive, but they focus on evaluating predictive accuracy per se. They do not study the performance of predictions in decision making, which is our concern.

3.1 *The fundamental importance of the cross-covariate structure of the state space*

A focus of controversy has been the asserted capacity of certain prediction methods advocated by ML researchers to successfully predict outcomes when covariate vectors have high dimension relative to sample size. Reacting to the longstanding concern that prediction becomes increasingly difficult as the dimension of the covariate vector increases (aka the curse of dimensionality), Breiman (2001a) expressed considerable optimism when he wrote (p. 209): "For decades, the first step in prediction methodology was to avoid the curse. . . . Recent work has shown that dimensionality can be a blessing." He continued (p. 209):

> "Reducing dimensionality reduces the amount of information available for prediction. The more predictor variables, the more information. There is also information in various combinations of the predictor variables. Let's try going in the opposite direction: Instead of reducing dimensionality, increase it by adding many functions of the predictor variables."

Logically, Breiman's optimism cannot be completely realistic. The foundation of the curse of dimensionality is that, as the support of the covariate distribution grows, the probability $P(x)$ of drawing any specified value of x into a random training sample of fixed size N necessarily decreases. Hence, the expected number of observations of y associated with any specified value of x necessarily falls. Indeed, it is zero when the covariate distribution is continuous rather than discrete. It follows that, as the support of the covariate distribution grows, prediction of y conditional on the specified value of x with a training sample of size N must increasingly rely on outcome data associated with other covariate values. However, observation of outcomes associated with other covariate values per se conveys no information about $P(y|x)$ at the specified x-value of interest. These data become informative *only* given cross-covariate restrictions on the state space that relate $P(y|x)$ to the conditional outcome distributions $P(y|x')$, $x' \neq x$.

Frequentist statistical theory has long sought to characterize how the statistical imprecision of conditional prediction varies with the maintained cross-covariate restrictions. Analysis is straightforward in parametric modeling of conditional distributions, where the distributions $\{P(y|x), x \in X\}$, or at least the conditional expectations $\{E(y|x), x \in X\}$ are assumed to all be functions of a finite-dimensional parameter vector. Then cross-covariate restrictions are formalized in the specification of the parameter space. Analysis is relatively straightforward in the semiparametric approach to conditional classification called *maximum score* estimation in econometrics (Manski (1975, 1985); Manski and Thompson (1989)) and *empirical risk minimization* in the statistical learning theory of Vapnik (1999, 2000). In both the parametric and semiparametric settings, researchers have mainly studied asymptotic questions of consistency and rates of convergence.

Analysis is more subtle, but still intuitive, in classical research on nonparametric regression, where $\{P(y|x), x \in X\}$ or $\{E(y|x), x \in X\}$ are assumed to vary in a smooth manner across suitably nearby values of x. The more recent body of research assuming some concept of *dimensional sparsity* is yet more subtle in that it does not a priori fix the covariate space over which the conditional distribution or mean may vary. Instead, it constrains the number of elements of the covariate vector across which variation may occur.

The ML literature using deep neural networks as predictor functions replaces dimensional sparsity with the concept of compositional sparsity. PMRM+ (2017) wrote this (p. 503):

> "The main message is that deep networks have the theoretical guarantee, which shallow networks do not have, that they can avoid the curse of dimensionality for an important class of problems, corresponding to compositional functions, i.e., functions of functions. An especially interesting subset of such compositional functions are hierarchically local compositional functions where all the constituent functions are local in the sense of bounded small dimensionality. The deep networks that can approximate them without the curse of dimensionality are of the deep convolutional type."

Poggio and Fraser (2024) defined compositional sparsity as follows (p. 1): "The property that a compositional function have [*sic*] "few" constituent functions, each depending on only a small subset of inputs."

Thus, if the function governing variation in the outcome y with covariates x is compositionally sparse, then the claim is that a deep neural network will approximate it well. It should not be surprising that a predictor function whose structure is similar to the unknown, true function of interest will approximate it well. Recognizing this led Shamir (2020), commenting on Schmidt-Hieber (2020), to assert (p. 1912): "Essentially, we have replaced a 'curse of dimensionality' effect with a 'curse of sparsity'."

A logical follow-up question concerns whether it is realistic in practice to assume that unknown functions governing variation in y with x are compositionally sparse. PMRM+ (2017) asked this question and conjectured a partial answer, writing (p. 517):

> "This line of arguments defines a class of algorithms that is universal and can be optimal for a large set of problems. It does not however explain why problems encountered in practice should match this class of algorithms. Though we and others have argued that the explanation may be in either the physics or the neuroscience of the brain, these arguments... are not (yet) rigorous. Our conjecture is that compositionality is imposed by the wiring of our cortex and is reflected in language. Thus, compositionality of many computations on images may reflect the way we describe and think about them."

As we see it, compositional sparsity is an intriguing concept that warrants further study. In principle, SDT can specify that a state space is composed of compositionally sparse functions and evaluate the predictive performance of deep neural networks that have this structure. However, we expect that this implementation of SDT will be computationally challenging. We also think it important to question the realism of assuming that unknown conditional means and other features of conditional distributions actually are compositionally sparse. PMRM+ (2017) and Poggio and Fraser (2024) express optimism about this. We are less sure.

3.2 *Bounded-variation restrictions on the state space*

In some applications, it is realistic to globally bound the variation across covariate values of a conditional mean $E(y|x)$ or other feature of $P(y|x)$. Focusing on the problem of identification rather than decision making with sample data, Manski and Pepper (2000) initiated study of identification with *monotone instrumental variable* assumptions, which assume that a subvector of x is ordered, and that $E(y|x)$ varies monotonically across these covariate values. More general *bounded-variation* assumptions have been used to study conditional prediction of risk of illness by Manski (2018) and Li, Litvin, and Manski (2023). They have been used to study prediction of treatment response in criminal justice settings by Manski and Pepper (2013, 2018).

In settings where the outcome y is binary, bounded-variation assumptions have the simple form

$$P_s\big(y=1|x'\big)+\lambda_-\big(x,x'\big)\leq P_s(y=1|x)\leq P_s\big(y=1|x'\big)+\lambda_+\big(x,x'\big),\quad \text{all } s\in S. \tag{9}$$

Here, x is a covariate value of prediction interest, x' is a different value, and $\lambda_-(x,x')\leq \lambda_+(x,x')$ are specified real numbers. Bounded-variation assumptions do not assert that $P_s(y=1|x)$ varies locally smoothly with x, as in classical nonparametric regression analysis. Nor do they assume the types of cross-covariate invariance embedded in dimensional or compositional sparsity assumptions. They impose a distinct type of cross-covariate restriction on the state space.

As far as we are aware, the only research studying SDT with bounded-variation assumptions are Stoye (2012), Manski (2023), and Montiel Olea, Qiu, and Stoye (2024). These focused on the maximum regret of SDFs in certain problems of treatment choice. Stoye (2012) obtained an analytical expression for the MMR decision function when $\lambda_-(x,x')=-\lambda_+(x,x')=\kappa$, where κ is a specified positive real number that does not vary with (x,x').

Manski (2023) considered numerical computation of maximum regret for general assumptions of form (9), when loss is an extension of the MCR where the magnitude of the loss incurred with a prediction error varies with x and s. For concreteness, the analysis considered a medical setting in which a clinician's problem is to predict illness and to use the prediction to choose between surveillance and aggressive treatment of a disease. The assumptions made about patient outcomes with each treatment implied that aggressive treatment is optimal if $P^*(y=1|x)$ exceeds a known threshold and surveillance is optimal otherwise. This will be discussed further in Section 4.

It was supposed that, not knowing $P^*(y=1|x)$, the clinician uses a kernel method to predict it and then acts as if the prediction is correct. A kernel estimate is a weighted average of the outcomes across different values of the covariate. Computation of maximum regret using alternative weights enables one to minimize maximum regret within the class of kernel predictors.

4. Probabilistic prediction for clinical decision making

Our discussion of statistical decision theory has been abstract, with mention of potential but not actual applications. We are concerned that the scarcity of applied use of SDT to

evaluate predictive algorithms has been accompanied by a rapid growth in use of OOS validation methods without providing a foundation in statistical theory. Considering the many domains in which OOS validation is becoming common, we are especially troubled by its increasing acceptance in medical research aiming to predict patient health outcomes. Probabilistic prediction of illness and treatment response is used widely to inform clinical decision making. It matters considerably to individual patients and to society at large that researchers use a well-grounded approach to assess the reliability of predictions.

Section 4.1 summarizes the trajectory of prediction methodology in medicine. Section 4.2 focuses on an important medical setting, where a clinician wants to predict illness and to use the prediction to choose between surveillance and aggressive treatment of a disease. Section 4.3 discusses a recent application of OOS validation to study this prediction and decision problem.

4.1 *A brief history of probabilistic prediction in medicine*

Empirical research on medical risk assessment has long used sample data on illness in study populations to estimate conditional probabilities of illness and to make probabilistic predictions of treatment response. Throughout the 20th century, risk assessment conditional on observed patient covariates was performed by biostatisticians who use frequentist statistical theory to propose methodology and assess findings. Certain parametric and semiparametric models became standard. The logit model introduced by Berkson (1944) has been used to predict binary outcomes. The proportional hazards model introduced by Cox (1972) has been used to predict the timing of death and other future events. Linear regression models introduced by Galton (1886) have been used to predict the means of real-valued outcomes. Poisson regression has been used to predict outcomes taking count values (Nelder (1974)). The parameters of these models have been estimated by maximum likelihood or least squares, which have well-understood frequentist statistical properties.

Although the motivation ostensibly is to improve patient care, frequentist biostatistical analysis has commonly viewed prediction as a self-contained inferential problem rather than as a task undertaken specifically to inform treatment choice. Confidence intervals and standard errors have been used to measure statistical imprecision. When treatment choice is studied, Neyman–Pearson hypothesis tests have been used to judge whether treatment B is better than treatment A, the null hypothesis being that B is no better than A, and the alternative being that it is better. These inferential methods are remote from SDT. Confidence intervals and standard errors have no obvious interpretation in the Wald theory. As discussed in the Appendix, Wald (1939) initially proposed SDT as a superior replacement for Neyman–Pearson testing.

In the 21st century, medical risk assessment is increasingly performed by computer scientists, who view prediction methods as computational algorithms rather than through the lens of statistical theory. Whereas biostatisticians have favored the use of estimation of relatively simple parametric and semiparametric probabilistic prediction models, computer scientists fit data with increasingly complex nonparametric algorithms whose frequentist statistical properties are poorly understood.

Early on, computer scientists applied what are now considered classical nonparametric methods, such as kernel and nearest-neighbor regression; see, for example, the *Journal of Machine Learning Research* special issue on kernel methods (Cristianini, Shawe-Taylor, and Williamson (2001)). These are well understood from the perspective of asymptotic statistical theory. They later focused on random forest methods (Ho (1995); Breiman (2001b)), for which some statistical theory exists. However, they have increasingly used methods whose statistical properties are hardly understood at all. We have already noted the weak state of knowledge regarding deep neural networks. Another prominent approach without a foundation in statistical theory, which we have not mentioned thus far, is the Synthetic Minority Over-Sampling Technique (SMOTE) introduced by Chawla, Bowyer, Hall, and Kegelmeyer (2002). Whereas frequentist biostatisticians study how methods perform across repetitions of a sampling process, computer scientists engaged in medical prediction perform K-fold or CTF OOS validation, as championed by Breiman (2001a). See Deo (2015) and Rajkomar, Dean, and Kohane (2019) for two among many review articles written for audiences of clinicians.

4.2 *Prediction for choice between surveillance and aggressive treatment*

Choice between surveillance and aggressive treatment is a frequent clinical decision. The broad problem concerns a clinician caring for patients with observed covariates x. There are two care options for a specified disease, A denoting surveillance and B denoting aggressive treatment. The clinician must choose without knowing a patient's illness status; $y = 1$ if a patient is ill and $y = 0$ if not. Observing x, the clinician can attempt to learn the conditional probability of illness, $p_x \equiv P^*(y = 1|x)$. Medical research often proposes as-if optimization, using sample data to estimate p_x and acting as if the estimate is correct.

We summarize here the simple version of this decision problem studied in Manski (2023). As there, we assume that patient welfare with care option $c \in \{A, B\}$ has the known form $U_x(c, y)$; thus, welfare varies with whether the disease occurs and with the patient covariates x. We assume that aggressive treatment is better if the disease occurs, and surveillance is better otherwise. That is,

$$U_x(B, 1) > U_x(A, 1), \tag{10a}$$

$$U_x(A, 0) > U_x(B, 0). \tag{10b}$$

The specific form of welfare function $U_x(\cdot, \cdot)$ depends on the clinical context, but inequalities (10a)–(10b) are typically realistic.

We assume that the chosen care option does not affect whether the disease occurs; hence, a patient's illness probability is simply p_x rather than a function $p_x(c)$ of the care option. Treatment choice still matters because it affects the severity of illness and the patient's experience of side effects. Aggressive treatment is beneficial to the extent that it lessens the severity of illness, but harmful if it yields side effects that do not occur with surveillance.

4.2.1 *Optimal treatment choice with knowledge of p_x* Before considering decision making with sample data, suppose that the clinician knows p_x and maximizes expected patient welfare conditional on x. Then an optimal decision is

$$\begin{aligned}\text{Choose } A \text{ if } \quad & p_x \cdot U_x(A,1) + (1-p_x)\cdot U_x(A,0) \\ & \geq p_x \cdot U_x(B,1) + (1-p_x)\cdot U_x(B,0), \qquad (11a)\\ \text{Choose } B \text{ if } \quad & p_x \cdot U_x(B,1) + (1-p_x)\cdot U_x(B,0) \\ & \geq p_x \cdot U_x(A,1) + (1-p_x)\cdot U_x(A,0). \qquad (11b)\end{aligned}$$

The decision yields optimal expected patient welfare

$$\begin{aligned}\max\big[& p_x \cdot U_x(A,1) + (1-p_x)\cdot U_x(A,0), \\ & p_x \cdot U_x(B,1) + (1-p_x)\cdot U_x(B,0)\big]. \qquad (12)\end{aligned}$$

The optimal decision is easy to characterize when inequalities (10a)–(10b) hold. Let $p_x^\#$ denote the threshold value of p_x that makes options A and B have the same expected utility. This value is

$$p_x^\# = \frac{U_x(A,0) - U_x(B,0)}{\big[U_x(A,0) - U_x(B,0)\big] + \big[U_x(B,1) - U_x(A,1)\big]}. \qquad (13)$$

Option A is optimal if $p_x \leq p_x^\#$ and B if $p_x \geq p_x^\#$. Thus, optimal treatment choice does not require exact knowledge of p_x. It only requires knowing whether p_x is larger or smaller than $p_x^\#$.

Manski (2023) focused on an instructive special case that occurs when aggressive treatment neutralizes disease, in the sense that $U_x(B,0) = U_x(B,1)$. Let U_{xB} denote welfare with aggressive treatment. Then (10a) and (10b)–(13) reduce to

$$U_x(A,0) > U_{xB} > U_x(A,1). \qquad (14)$$

$$\text{Choose } A \text{ if } \quad p_x \cdot U_x(A,1) + (1-p_x)\cdot U_x(A,0) \geq U_{xB}, \qquad (15a)$$

$$\text{Choose } B \text{ if } \quad U_{xB} \geq p_x \cdot U_x(A,1) + (1-p_x)\cdot U_x(A,0). \qquad (15b)$$

$$\max\big[p_x \cdot U_x(A,1) + (1-p_x)\cdot U_x(A,0),\, U_{xB}\big]. \qquad (16)$$

$$p_x^\# = \frac{U_x(A,0) - U_{xB}}{U_x(A,0) - U_x(A,1)}. \qquad (17)$$

Further simplification occurs when one normalizes the location and scale of welfare by setting $U_x(A,0) = 1$ and $U_x(A,1) = 0$. Then $1 > U_{xB} > 0$ and $p_x^\# = 1 - U_{xB}$. We consider this special case below.

4.2.2 *Maximum regret of as-if optimization* Now suppose that the clinician does not know p_x. The state space lists all feasible values of p_x. In the abstract notation of Section 2, the objective function $w(\cdot,\cdot) : C \times S \to R^1$ has this form: $w(A,s) = p_{sx} \cdot U_x(A,1) + (1-p_{sx})\cdot U_x(A,0)$ and $w(B,s) = U_{xB}$.

Algorithm 1 Algorithm for calculation of maximum regret

1. Choose a grid on S
2. For each s in the grid, calculate expected regret (19)
   a. Compute $|(1 - p_{sx}) - U_{xB}|$
   b. Use Monte Carle integration to approximate $Q_s\{e[p_{sx}, \varphi_x(\psi), U_{xB}] = 1\}$ (see Section 2.1.3)
   c. Multiply $a \times b$
3. Maximize expected regret over S

Suppose that the clinician does not know whether p_x is smaller or larger than $p_x^{\#} = 1 - U_{xB}$. Formally, suppose the clinician knows that $p_{mx} < 1 - U_{xB} < p_{Mx}$, where $p_{mx} \equiv \min_{s\in S} p_{sx}$ and $p_{Mx} \equiv \max_{s\in S} p_{sx}$. Then the clinician cannot maximize expected patient welfare conditional on x.

The clinician can, however, use sample data to estimate p_x and then act as if the estimate is correct. Let Ψ be a sample space and let Q_s be a sampling distribution with realizations $\psi \in \Psi$. Let $\varphi_x(\psi)$ be a point estimate of p_x. The clinician can maximize expected welfare acting as if $p_x = \varphi_x(\psi)$. Then the chosen care option is A if $\varphi_x(\psi) \leq 1 - U_{xB}$ and is B if $\varphi_x(\psi) > 1 - U_{xB}$.

Now consider the regret of treatment choice acting as if $p_x = \varphi_x(\psi)$. Let $e[p_{sx}, \varphi_x(\psi), U_{xB}]$ denote the occurrence of an error in state s when $\varphi_x(\psi)$ is used to choose treatment. That is, $e[p_{sx}, \varphi_x(\psi), U_{xB}] = 1$ when p_{sx} and $\varphi_x(\psi)$ yield different treatments, while $e[p_{sx}, \varphi_x(\psi), U_{xB}] = 0$ when p_{sx} and $\varphi_x(\psi)$ yield the same treatment. Regret using estimate $\varphi_x(\psi)$ is

$$\begin{aligned} R_{sx}\big[\varphi_x(\psi)\big] &= \max(1 - p_{sx}, U_{xB}) - (1 - p_{sx}) \cdot 1\big[1 - \varphi_x(\psi) \geq U_{xB}\big] \\ &\quad - U_{xB} \cdot 1\big[U_{xB} > 1 - \varphi_x(\psi)\big] \\ &= \big|(1 - p_{sx}) - U_{xB}\big| \\ &\quad \cdot 1\big[1 - p_{sx} \geq U_{xB} > 1 - \varphi_x(\psi) \text{ or } 1 - p_{sx} < U_{xB} \leq 1 - \varphi_x(\psi)\big] \\ &= \big|(1 - p_{sx}) - U_{xB}\big| \cdot e\big[p_{sx}, \varphi_x(\psi), U_{xB}\big]. \end{aligned} \tag{18}$$

Expected regret across repeated samples is

$$E_s\big\{R_{sx}\big[\varphi_x(\psi)\big]\big\} = \big|(1 - p_{sx}) - U_{xB}\big| \cdot Q_s\big\{e\big[p_{sx}, \varphi_x(\psi), U_{xB}\big] = 1\big\}. \tag{19}$$

Maximum expected regret across the state space is $\max_{s\in S} E_s\{R_{sx}[\varphi_x(\psi)]\}$.

Manski (2023) applied Algorithm 1 to numerical computation of maximum regret with as-if optimization when bounded-variation assumptions restrict the state space and $\varphi_x(\psi)$ is a kernel estimate of p_x. In the present paper, Section 4.2.3 considers computation when $\varphi_x(\psi)$ is an estimate that uses CV to choose tuning parameters.

4.2.3 *Illustration: Maximum regret of as-if optimization with cross-validated estimation of p_x* We earlier mentioned the recent contribution of Bates, Hastie, and Tibshirani (2024), who performed a frequentist analysis of measures of prediction error when certain prediction models are evaluated using CV methods. Their study focused on the prevalent ML use of CV to evaluate predictive accuracy. They did not address the distinct use of CV to choose tuning parameters in predictive statistical models, such as the bandwidth in kernel estimation or the penalty in LASSO estimation. They wrote (p. 1436):

> "CV is often used to compare predictive models, such as when selecting a model or a good value of a learning algorithm's hyperparameters. . . . While we expect that our proposed estimator would be of use for hyperparameter selection because it yields more accurate confidence intervals for prediction error, we do not pursue this problem further in the present work."

From the perspective of decision making using sample data, we question their conjecture relating hyperparameter selection to accuracy of confidence intervals for prediction error. We noted earlier that confidence intervals have no obvious interpretation in SDT. We recommend evaluation of CV choice of tuning parameters by the maximum regret of the resulting decisions.

To illustrate the striking difference between the type of analysis performed by Bates et al. and what we have in mind, consider the example they present in their opening Section 1.1. They write (p. 1434):

> "we consider a sparse logistic regression model. . . with $n = 90$ observations of $p = 1000$ features, and a coefficient vector $\theta = c \cdot (1, 1, 1, 1, 0, 0, \ldots)^T \in R^p$ with four nonzero entries of equal strength. The feature matrix $X \in R^n \times p$ is comprised of iid standard normal variables, and we chose the signal strength c so that the Bayes misclassification rate is 20%. We estimate the parameters using ℓ_1-penalized logistic regression with a fixed penalty level. In this case, naïve confidence intervals for prediction error are far too small: intervals with desired miscoverage of 10% give 31% miscoverage in our simulation. . . . The intervals need to be made larger by a factor of about 1.6 to obtain coverage at the desired level in this case."

Their reference to the "Bayes misclassification rate" indicates that their concern is to measure the expected frequency of misclassification across the distribution of X when the estimated logit probability is used to predict the binary event, the prediction being 1 if the estimated probability is greater than 1/2 and 0 otherwise. With their assumption that X is i.i.d. standard normal and that the binary outcome is generated by the specified sparse logit model, they state that the oracle Bayes misclassification rate is 20%. Using sample data to estimate the logit model yields a larger misclassification rate across repeated samples, which they seek to characterize using CV.

The nature of their analysis differs in numerous ways from the computation of maximum regret that we described in Section 4.2.2. There we considered treatment choice for patients with a specified value of X, rather than across the distribution of X. In a specified state of nature, expected regret across repeated samples was given in (19), which multiplies the magnitude of a treatment error times the probability of an error. Calculation of maximum regret across the state space recognizes that, in practice, the true distribution of X and of the binary outcome conditional on X are not known, whereas

Bates et al. presume them to be i.i.d. normal and given by a particular sparse logistic model. Finally, we would envision using CV to choose the penalization parameter in the ℓ_1-penalized logistic regression whereas they fix the penalty level.

4.3 *A machine learning model of the probability of a coronary blockage*

It should by now be evident that evaluation of an estimate $\varphi_x(\cdot)$ of p_x by maximum regret differs fundamentally from ML evaluation by ex post prediction accuracy. In the CTF paradigm, one uses a training sample ψ_{tr} to compute $\varphi_x(\psi_{\text{tr}})$ and then uses this estimate to predict illness in a test sample ψ_{test}. A standard practice is to predict $y = 1$ if $\varphi_x(\psi_{\text{tr}})$ exceeds a specified threshold, say γ, and $y = 0$ otherwise. Prediction accuracy is measured by the *positive predictive value*, the fraction of the test sample with $y = 1$ conditional on $\varphi_x(\psi_{\text{tr}}) \geq \gamma$, and the *negative predictive value*, the fraction of the test sample with $y = 0$ conditional on $\varphi_x(\psi_{\text{tr}}) \leq \gamma$. These measures of ex post prediction accuracy do not consider performance across samples or over the state space. Nor do they consider the patient's welfare achieved when an estimate is used to choose a treatment.

An ambitious application of OOS evaluation of ML to choice between surveillance and aggressive treatment occurs in Mullainathan and Obermeyer (2022), who study the decision faced by an emergency room physician when a patient presents with symptoms that may indicate a heart attack. In this case, aggressive treatment begins by ordering an invasive and costly test, such as cardiac catheterization, to determine whether a new blockage in the coronary arteries exists and, if so, the location of the blockage. Mullainathan and Obermeyer specify a blockage-probability threshold model to determine whether or not to order the test for a blockage. They use data on almost 250,000 emergency room visits to estimate an ML model of the blockage probability conditional on over 16,000 patient covariates. Considering the information available to physicians who make the testing decision and the incentives they face, the authors seek to determine whether actual testing decisions deviate from those that would be optimal with their model. They use (p. 679): "actual health outcomes to evaluate whether those deviations represent mistakes or physicians' superior knowledge."

The ML method they employ is described as an "ensemble" model that combines health outcome predictions of four different models. Two algorithms—gradient boosted trees and LASSO—are each trained on two subsets of the data—tested patients and untested patients. The parameters of the four models are each tuned using 5-fold CV in the same 70% training sample. The predictions of the four models are then used as the four covariates in a logit regression that predicts patient outcomes in a separate 5% ensemble training sample The remaining 25% holdout sample is used to identify cases where low-probability patients were tested but the finding was negative (called "overtesting") and high-probability patients were not tested but subsequently have "major adverse cardiac events" (called "undertesting").

The methods used in this study exemplify the current acceptance and growing use of OOS validation methods without a foundation in statistical theory. These validation methods differ sharply from how one would proceed using SDT to study the choice between surveillance and aggressive treatment, as discussed in Section 4.2. Application of

SDT to high-dimensional covariate contexts and complex modeling approaches such as in the Mullainathan–Obermeyer study is not currently computationally tractable. We see this as a challenge to be overcome.

5. Conclusion

Whereas statisticians and econometricians have used frequentist or Bayesian statistical theory to propose and evaluate prediction methods, ML researchers have largely performed K-fold and CTF OOS validation. These types of OOS evaluation cannot yield generalizable lessons.

Considering our critique of OOS evaluation, a reviewer of an earlier draft of this paper commented: "while it is clear that the authors believe modern machine learning lacks a particular style of rigor to their taste, it is also true that the field has produced breakthroughs that would have been unimaginable 15 years ago." The reviewer then provided an impressive list of examples.

We agree that the capacity to perform conditional prediction of events in stable populations has improved remarkably recently, but we do not see a basis to attribute the improvement to OOS evaluation of predictive algorithms. Our perception is that the improvements stem mainly from the combination of (1) vast increases in the availability of data to train predictive models, (2) vast increases in the capacity of computer hardware to perform prediction rapidly, and (3) addition of new approaches to the body of nonparametric regression methods developed in statistics from the 1950s onward, perhaps most notably the methodology of deep neural networks for approximation of functions with compositional sparsity.

It may be that the simplicity of OOS evaluation has contributed to the adoption of ML predictive algorithms, but we do not see how it has directly contributed to the improvements themselves. To the contrary, we are concerned that selective publication of purportedly successful OOS evaluations may yield a misleadingly positive impression of some algorithms and encourage their premature adoption.[6] We see this as particularly problematic in medical risk assessment.

In Sections 1 through 3, we argued abstractly that ML researchers should instead perform comprehensive OOS evaluation using SDT. SDT is remote from the types of OOS evaluation currently practiced, but it is not remote conceptually. It performs OOS evaluation across all possible (1) training samples, (2) populations that may generate training data, and (3) populations of prediction interest.

SDT is simple in abstraction, but it is often computationally demanding to implement. Some progress has been made in tractable implementation of SDT in prediction problems with low-dimensional covariates. We recognize that extending the approach to prediction problems with high-dimensional covariates may be computationally challenging. Yet we should confront the challenge. We are concerned about potentially harmful consequences, especially in clinical decision making but also in many other

[6]Whereas we are concerned with the possible negative effects of selective publication, Donoho (2024) asserted that competitive challenges in the CTF encourage productive innovation. It may be that both conjectures have partial validity.

settings, arising from the use of predictive algorithms evaluated by K-fold or CTF validation rather than by comprehensive OOS evaluation using SDT. We therefore call on ML researchers to join with econometricians and statisticians in expanding the domain within which implementation of SDT is tractable.

Appendix: Binary Classification Problems

Binary classification problems are ones in which the choice set C contains two feasible actions, say A and B. An SDF $c(\cdot)$ partitions Ψ into two regions that separate the data yielding choice of each action. These are $\Psi_{c(\cdot)A} \equiv [\psi \in \Psi : c(\psi) = A]$ and $\Psi_{c(\cdot)B} \equiv [\psi \in \Psi : c(\psi) = B]$. Thus, one chooses A if the data lie in set $\Psi_{c(\cdot)A}$ and B if the data lie in $\Psi_{c(\cdot)B}$.

Choice between two actions can be viewed as hypothesis tests, but the Wald perspective on testing differs from that of Neyman and Pearson (1928, 1933). Wald (1939) proposed evaluation of the performance of an SDF for binary classification by the expected loss that it yields across realizations of the sampling process. The loss distribution in state s is Bernoulli, with mass points $\max[L(a,s), L(b,s)]$ and $\min[L(a,s), L(b,s)]$. These coincide if $L(a,s) = L(b,s)$. When $L(a,s) \neq L(b,s)$, let $R_{c(\cdot)s}$ denote the probability that $c(\cdot)$ yields an error, choosing the inferior action over the superior one. That is,

$$\begin{aligned} R_{c(\cdot)s} &= Q_s\big[c(\psi) = b\big] \quad \text{if } L(a,s) < L(b,s), \\ &= Q_s\big[c(\psi) = a\big] \quad \text{if } L(b,s) < L(a,s). \end{aligned} \tag{A1}$$

These are the probabilities of Type I and Type II errors.

The probabilities that loss equals $\min[L(a,s), L(b,s)]$ and $\max[L(a,s), L(b,s)]$ are $1 - R_{c(\cdot)s}$ and $R_{c(\cdot)s}$. Hence, expected loss is

$$\begin{aligned} E_s\big\{L\big[c(\psi), s\big]\big\} &= R_{c(\cdot)s}\big\{\max\big[L(a,s), L(b,s)\big]\big\} \\ &\quad + [1 - R_{c(\cdot)s}]\big\{\min\big[L(a,s), L(b,s)\big]\big\} \\ &= \min\big[L(a,s), L(b,s)\big] + R_{c(\cdot)s} \cdot \big|L(a,s) - L(b,s)\big|. \end{aligned} \tag{A2}$$

$R_{c(\cdot)s} \cdot |L(a,s) - L(b,s)|$ is the expected regret of $c(\cdot)$. Thus, expected regret, defined in abstraction in (6), has a simple form when choice is binary. It is the product of the error probability and the magnitude of the incremental loss when an error occurs.

It is particularly simple to determine the regret of a point predictor $p(\cdot)$ of a binary outcome y, when measuring loss by mean square error (MSE) or the misclassification rate (MCR). Let y be generated by a Bernoulli distribution $P(y)$, ψ be generated by a sampling distribution $Q(\psi)$, and let the realizations of y and ψ be statistically independent.

When loss is MSE and the point predictor $p(\cdot)$ can take any real value, the risk of $p(\cdot)$ in state s is

$$E\big[y - p(\psi)\big]^2 = V_s(y) + V_s\big[p(\psi)\big] + \big\{E_s(y) - E_s\big[p(\psi)\big]\big\}^2. \tag{A3}$$

Risk in state s is minimized by setting $p(\psi) = E_s(y)$ for all data realizations ψ, yielding the minimum risk value $V_s(y)$. Hence, the regret of $p(\cdot)$ in state s is $V_s[p(\psi)] + \{E_s(y) - E_s[p(\psi)]\}^2$.

When y is binary and the point predictor $p(\cdot)$ is constrained to be binary, $V_s[p(\psi)] = Q_s[p(\psi) = 1] - Q_s[p(\psi) = 1]^2$, $E_s(y) = P_s(y = 1)$, and $E_s[p(\psi)] = Q_s[p(\psi) = 1]$. Hence, regret is

$$\begin{aligned}
&Q_s\big[p(\psi) = 1\big] - Q_s\big[p(\psi) = 1\big]^2 + \big\{P_s(y = 1) - Q_s\big[p(\psi) = 1\big]\big\}^2 \\
&\quad = Q_s\big[p(\psi) = 1\big] - Q_s\big[p(\psi) = 1\big]^2 + P_s(y = 1)^2 \\
&\qquad + Q_s\big[p(\psi) = 1\big]^2 - 2 \cdot P_s(y = 1) \cdot Q_s\big[p(\psi) = 1\big] \\
&\quad = Q_s\big[p(\psi) = 1\big]\big[1 - P_s(y = 1)\big] + P_s(y = 1)\big\{P_s(y = 1) - Q_s\big[p(\psi) = 1\big]\big\} \\
&\quad = Q_s\big[p(\psi) = 1\big]\big[1 - P_s(y = 1)\big] + P_s(y = 1)\big\{1 - Q_s\big[p(\psi) = 1\big]\big\} \\
&\qquad - P_s(y = 1)\big[1 - P_s(y = 1)\big].
\end{aligned} \tag{A4}$$

The first term is the probability of a misclassification of the form $[y = 0,\ p(\psi) = 1]$. The second term is the probability of a misclassification of the form $[y = 1,\ p(\psi) = 0]$. The third term is the outcome variance $V_s(y)$.

When loss is MCR, the risk of $p(\cdot)$ in state s is

$$\begin{aligned}
&P_s Q_s\big[y \neq p(\psi)\big] \\
&\quad = Q_s\big[p(\psi) = 1\big]\big[1 - P_s(y = 1)\big] + P_s(y = 1)\big\{1 - Q_s\big[p(\psi) = 1\big]\big\}.
\end{aligned} \tag{A5}$$

Risk in state is minimized by setting $p(\psi) = 1$ for all ψ if $P_s(y = 1)] > 1/2$ and $p(\psi) = 0$ for all ψ if $P_s(y = 1)] < 1/2$, yielding the minimum risk value $\min[P_s(y = 1),\ 1 - P_s(y = 1)]$. Hence, regret is

$$\begin{aligned}
&Q_s\big[p(\psi) = 1\big]\big[1 - P_s(y = 1)\big] + P_s(y = 1)\big\{1 - Q_s\big[p(\psi) = 1\big]\big\} \\
&\qquad - \min\big[P_s(y = 1),\ 1 - P_s(y = 1)\big].
\end{aligned} \tag{A6}$$

Thus, regret when loss is MSE or MCR differs only in that the former subtracts $P_s(y = 1)[1 - P_s(y = 1)]$ from the probability of misclassification and the latter subtracts $\min[P_s(y = 1),\ 1 - P_s(y = 1)]$. The two differ because the MSE calculation permits the prediction $p(\cdot)$ to take any real value, whereas the MCR calculation constrains the prediction to take the value 0 or 1.

References

Athey, Susan and Stephan Wager (2021), "Policy learning with observational data." *Econometrica*, 89, 133–161. [0006]

Bartlett, Peter, Andrea Montanari, and Alexander Rakhlin (2021), "Deep learning: A statistical viewpoint." *Acta Numerica*, 30, 87–201. [0012]

Bates, Stephen, Trevor Hastie, and Robert Tibshirani (2024), "Cross-validation: What does it estimate and how well does it do it?" *Journal of the American Statistical Association*, 119, 1434–1445. [0003, 0013, 0020]

Berger, James (1985), *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York. [0006, 0008]

Berkson, Joseph (1944), "Application of the logistic function to bio-assay." *Journal of the American Statistical Association*, 39, 357–365. [0016]

Blanchet, Jose, Jiajin Li, Sirui Lin, and Xuhui Zhang (2024), "Distributionally robust optimization and robust statistics." https://arxiv.org/abs/2401.14655. [0011]

Bommasani, Rishi, Drew Hudson, Ehsan Adeli et al. (2021), "On the opportunities and risks of foundation models." arXiv preprint arXiv:2108.07258. [0012]

Breiman, Leo (2001a), "Statistical modeling: The two cultures." *Statistical Science*, 16, 199–231. [0002, 0003, 0009, 0013, 0017]

Breiman, Leo (2001b), "Random forests." *Machine Learning*, 45, 5–32. [0017]

Chamberlain, Gary (2000), "Econometric applications of maxmin expected utility." *Journal of Applied Econometrics*, 15, 625–644. [0012]

Chawla, N., K. Bowyer, L. Hall, and W. Kegelmeyer (2002), "SMOTE: Synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research*, 16, 321–357. [0017]

Cox, David (1972), "Regression model and life-tables." *Journal of the Royal Statistical Society, Series B*, 34, 187–202. [0016]

Cox, David (2001), "Comment on 'Statistical modeling: The two cultures'." *Statistical Science*, 16, 216–218. [0003]

Cristianini, Nello, John Shawe-Taylor, and Robert Williamson (2001), "Introduction to the special issue on Kernel methods." *Journal of Machine Learning Research*, 2, 95–96. [0017]

Deo, Rahul (2015), "Machine learning in medicine." *Circulation*, 132, 1920–1930. [0017]

Dominitz, Jeff and Charles Manski (2017), "More data or better data? A statistical decision problem." *Review of Economic Studies*, 84, 1583–1605. [0006]

Dominitz, Jeff and Charles Manski (2022), "Minimax-regret sample design in anticipation of missing data, with application to panel data." *Journal of Econometrics*, 226, 104–114. [0006]

Dominitz, Jeff and Charles Manski (2025), "Accounting for nonresponse in election polls: Total margin of error." *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2025.2537459. Forthcoming. [0006]

Donoho, David (2017), "50 years of data science." *Journal of Computational and Graphical Statistics*, 26, 745–766. [0003]

Donoho, David (2024), "Data science at the singularity." *Harvard Data Science Review*, 6 (1). https://hdsr.mitpress.mit.edu/pub/g9mau4m0. [0003, 0022]

Donoho, David, Ian Johnstone, Gerard Kerkyacharian, and Dominique Picard (1995), "Wavelet shrinkage: Asymptopia?" *Journal of the Royal Statistics Society, Series B*, 57, 301–369. [0012]

Efron, Bradley (2020), "Estimation, prediction, and attribution." *International Statistical Review*, 88, S28–S59. [0004]

Ferguson, Thomas (1967), *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, San Diego. [0006]

Galton, Francis (1886), "Regression towards mediocrity in hereditary stature." *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263. [0016]

Goldberger, Arthur (1968), *Topics in Regression Analysis*. McMillan, New York. [0002]

Hampel, Frank, Ronchetti Elvezio, Peter Rousseeuw, and Werner Stahel (1986), *Robust Statistics*. Wiley, New York. [0011]

Hansen, Lars and Thomas Sargent (2008), *Robustness*. Princeton University Press, Princeton. [0011]

Hirano, Keisuke and Jack Porter (2009), "Asymptotics for statistical treatment rules." *Econometrica*, 77, 1683–1701. [0006]

Hirano, Keisuke and Jack Porter (2020), "Statistical decision rules in econometrics." In *Handbook of Econometrics*, Vol. 7 (Steven Durlauf, Lars Hansen, James Heckman and Rosa Matzkin, eds.), 283–354, North Holland, Amsterdam. [0006]

Ho, Tin (1995), "Random decision forests." In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, 278–282, Montreal, QC. [0017]

Horowitz, Joel and Charles Manski (1995), "Identification and robustness with contaminated and corrupted data." *Econometrica*, 63, 281–302. [0011]

Huber, Peter (1981), *Robust Statistics*. Wiley, New York. [0011]

Hurwicz, Leo (1951), "Some specification problems and applications to econometric models." *Econometrica*, 19, 343–344. [0008]

Kitagawa, Toru and Aleksey Tetenov (2018), "Who should be treated? Empirical welfare maximization methods for treatment choice." *Econometrica*, 86, 591–616. [0006]

Li, Sheyu, Valentyn Litvin, and Charles Manski (2023), "Partial identification of personalized treatment response with trial-reported analyses of binary subgroups." *Epidemiology*, 34, 319–324. [0015]

Manski, Charles (1975), "Maximum score estimation of the stochastic utility model of choice." *Journal of Econometrics*, 3, 205–228. [0013]

Manski, Charles (1985), "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator." *Journal of Econometrics*, 27, 313–333. [0013]

Manski, Charles (2000), "Identification problems and decisions under ambiguity: Empirical analysis of treatment response and normative analysis of treatment choice." *Journal of Econometrics*, 95, 415–442. [0006]

Manski, Charles (2004), "Statistical treatment rules for heterogeneous populations." *Econometrica*, 72, 221–246. [0006, 0009]

Manski, Charles (2005), *Social Choice With Partial Knowledge of Treatment Response*. Princeton University Press, Princeton. [0006]

Manski, Charles (2007), "Minimax-regret treatment choice with missing outcome data." *Journal of Econometrics*, 139, 105–115. [0006]

Manski, Charles (2018), "Credible ecological inference for medical decisions with personalized risk assessment." *Quantitative Economics*, 9, 541–569. [0015]

Manski, Charles (2019), "Treatment choice with trial data: Statistical decision theory should supplant hypothesis testing." *The American Statistician*, 73, 296–304. [0006]

Manski, Charles (2021), "Econometrics for decision making: Building foundations sketched by haavelmo and Wald." *Econometrica*, 89, 2827–2853. [0006, 0007, 0009]

Manski, Charles (2023), "Probabilistic prediction for binary treatment choice: With focus on personalized medicine." *Journal of Econometrics*, 234, 647–663. [0004, 0006, 0007, 0012, 0015, 0017, 0018, 0019]

Manski, Charles (2024), *Discourse on Social Planning Under Uncertainty*. Cambridge University Press, Cambridge. [0007]

Manski, Charles and John Pepper (2000), "Monotone instrumental variables: With an application to the returns to schooling." *Econometrica*, 68, 997–1010. [0015]

Manski, Charles and John Pepper (2013), "Deterrence and the death penalty: Partial identification analysis using repeated cross sections." *Journal of Quantitative Criminology*, 29, 123–141. [0015]

Manski, Charles and John Pepper (2018), "How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions." *Review of Economics and Statistics*, 100, 232–244. [0015]

Manski, Charles and Aleksey Tetenov (2019), "Trial size for near-optimal choice between surveillance and aggressive treatment: Reconsidering MSLT-II." *The American Statistician*, 73 (S1), 305–311. [0006]

Manski, Charles and Aleksey Tetenov (2021), "Statistical decision properties of imprecise trials assessing COVID-19 drugs." *Value in Health*, 24, 641–647. [0006]

Manski, Charles and Aleksey Tetenov (2023), "Statistical decision theory respecting stochastic dominance." *The Japanese Economic Review*, 74, 447–469. [0009]

Manski, Charles and Alexsey Tetenov (2016), "Sufficient trial size to inform clinical practice." *Proceedings of the National Academy of Sciences*, 113, 10518–10523. [0006]

Manski, Charles and Scott Thompson (1989), "Estimation of best predictors of binary response." *Journal of Econometrics*, 40, 97–123. [0013]

Mbakop, Eric and Max Tabord-Meehan (2021), "Model selection for treatment choice: Penalized welfare maximization." *Econometrica*, 89, 825–848. [0006]

Montiel Olea, Jose, Chen Qiu, and Jörg Stoye (2024), "Decision theory for treatment choice problems with partial identification." https://arxiv.org/abs/2312.17623. [0006, 0015]

Mullainathan, Sendhil and Ziad Obermeyer (2022), "Diagnosing physician error: A machine learning approach to low-value health care." *Quarterly Journal of Economics*, 137, 679–727. [0021]

Nelder, John (1974), "Log linear models for contingency tables: A generalization of classical least squares." *Journal of the Royal Statistics Society, Series C*, 23, 323–329. [0016]

Neyman, Jerzey and Egon Pearson (1928), "On the use and interpretation of certain test criteria for purposes of statistical inference." *Biometrika*, 20A, 175–240. 263–294. [0007, 0023]

Neyman, Jerzey and Egon Pearson (1933), "On the problem of the most efficient tests of statistical hypotheses." *Philosophical Transactions of the Royal Society of London, Ser. A*, 231, 289–337. [0007, 0023]

Poggio, Tomaso and Maia Fraser (2024), "Compositional sparsity of learnable functions." *Bulletin of the American Mathematical Society*, 61, 438–456. [0004, 0006, 0014]

[PMRM+] Poggio, Tomaso, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao (2017), "Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review." *International Journal of Automation and Computing*, 14, 503–519. [0014]

Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane (2019), "Machine learning in medicine." *New England Journal of Medicine*, 380, 1347–1358. [0017]

Savage, Leonard (1951), "The theory of statistical decision." *Journal of the American Statistical Association*, 46, 55–67. [0008, 0009]

Sawa, Takamitsu and Takeshi Hiromatsu (1973), "Minimax regret significance points for a preliminary test in regression analysis." *Econometrica*, 41, 1093–1101. [0006, 0012]

Schmidt-Hieber, Johannes (2020), "Nonparametric regression using deep neural networks with ReLU activation function." *Annals of Statistics*, 48, 1875–1899. [0012, 0014]

Shamir, Ohad (2020), "Discussion of: 'Nonparametric regression using deep neural networks with RELU activation function'." *The Annals of Statistics*, 48, 1911–1915. [0014]

Stoye, Jörg (2009), "Minimax regret treatment choice with finite samples." *Journal of Econometrics*, 151, 70–81. [0006]

Stoye, Jörg (2012), "Minimax regret treatment choice with covariates or with limited validity of experiments." *Journal of Econometrics*, 166, 138–156. [0006, 0015]

Taddy, Matt (2019), "The technological elements of artificial intelligence." In *The Economics of Artificial Intelligence: An Agenda* (Ajai Agrawal, Joshua Gans, and Avi Goldfarb, eds.), University of Chicago Press, Chicago. [0002, 0004]

Vapnik, Vladimir (1999), "An overview of statistical learning theory." *IEEE Transactions on Neural Networks*, 10, 988–999. [0002, 0013]

Vapnik, Vladimir (2000), *The Nature of Statistical Learning Theory*. Springer, New York. [0002, 0013]

Wald, Abraham (1939), "Contribution to the theory of statistical estimation and testing hypotheses." *Annals of Mathematical Statistics*, 10, 299–326. [0004, 0016, 0023]

Wald, Abraham (1945), "Statistical decision functions which minimize the maximum risk." *Annals of Mathematics*, 46, 265–280. [0004, 0007, 0008]

Wald, Abraham (1950), *Statistical Decision Functions*. Wiley, New York. [0004, 0008]

Walley, Peter (1991), *Statistical Reasoning With Imprecise Probabilities*. Chapman and Hall, New York. [0008]

Watson, James and Chris Holmes (2016), "Approximate models and robust decisions." *Statistical Science*, 31, 465–489. [0011]

Yata, Kohei (2023), "Optimal decision rules under partial identification." https://arxiv.org/abs/2111.04926. [0006]

Zhang, Ping (1993), "Model selection via multifold cross validation." *The Annals of Statistics*, 21, 299–313. [0003]

Zhou, Kemin, John Doyle, and Keith Glover (1995), *Robust and Optimal Control*. Prentice Hall, New Jersey. [0011]

Co-editor Stéphane Bonhomme handled this manuscript.

Manuscript received 14 October, 2024; final version accepted 21 July, 2025; available online 22 July, 2025.

All authors assume responsibility for all aspects of the paper.