Appendix:
Counterfactually Comparing Abstaining Classifiers

Yo Joong Choe[*]
Data Science Institute
University of Chicago
`yjchoe@uchicago.edu`

Aditya Gangrade
Department of EECS
University of Michigan
`aditg@umich.edu`

Aaditya Ramdas
Dept. of Statistics and Data Science
Machine Learning Department
Carnegie Mellon University
`aramdas@cmu.edu`

A Further discussion

A.1 Additional motivating examples for the counterfactual score

Here, we include three additional examples that motivate the counterfactual score. These illustrate cases in which either (a) the missing predictions are utilized in a failure mode (Examples A.1 and A.2) or (b) the missing predictions are relevant to the evaluator's future uses (Examples A.2 and A.3).

Example A.1 (Inattentive driver in a self-driving car)**.** Consider an ML classifier in a semi-autonomous vehicle system that makes a prediction (the weather, time of day, etc.) given the available sensory inputs. The prediction is used by the sequential decision making agent. In principle, when facing a high-uncertainty input, the classifier can abstain from a prediction and alert the driver to take back control. Yet, in reality, we would still greatly prefer a system that can make a safe decision in case the driver is inattentive[2] at the time and cannot take back control. In such a case, we require evaluating what a system would have done in situations where it decided to abstain.

Example A.2 (Comparing ML radiologist assistants)**.** Suppose that a hospital is evaluating third-party radiology application programming interfaces (API) that can assist with its diagnosis system. Each API will either give a prediction or abstain from making one; if it abstains, then a human radiologist will examine the input (Raghu et al., 2019). The hospital is wary that there are inputs for which the professional would also abstain or have cognitive biases against (Busby et al., 2018; Madras et al., 2018). Thus, it would need to occasionally rely on the classifier's predictions even on examples that it chose to abstain. If these "hidden predictions" are not readily available from the third-party providers (e.g., require extra costs), how can the hospital evaluate and compare their services?

Example A.3 (Evaluating an abstaining classifier's internal biases)**.** Suppose that an independent agency is auditing an ML-based recidivism prediction system[3] that has been deployed for a certain amount of time. Given the high stakes of misclassification, the system is equipped with a learned abstention mechanism and decides to occasionally abstain from making a prediction, in which case the rejected cases are examined by human judges. The auditing agency is interested in checking whether the ML classifier possesses internal biases against certain demographic groups, and in particular, it wants to estimate the classifier's accuracy on each demographic group *had it not abstained on any input*. While the agency has access to the system's past predictions and abstentions, it does not have access to the underlying predictive model or its abstention mechanism. In other words, the agency requires a black-box evaluation method that estimates the counterfactual score of this system.

---

[*] This work was submitted while this author was at Carnegie Mellon University.

[2] Driver inattention is a serious issue for semi-autonomous vehicles: studies have shown that the lack of active involvement correlates with both driver fatigue and tardy reactions to take-over requests (Vogelpohl et al., 2019).

[3] Algorithmic approaches to recidivism prediction, such as COMPAS, are both popular and controversial.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

A.2 An equivalent formulation in the potential outcomes framework

There are other equivalent ways to formulate our setup (Section 2.2) using variants of the potential outcomes framework. First, we can define a (potentially observed) prediction $f(X;R)$, which equals $f(X)$ if $R=0$ and $*$ if $R=1$, where the symbol $*$ indicates an abstention (the same notation is used in Rubin (1976)'s missing data framework). The score S is then $\mathsf{s}(f(X),Y)$ if $R=0$ and $*$ otherwise.

Alternatively, we can explicitly invoke Rubin (1974)'s potential outcomes framework to write $S(0) \leftarrow \mathsf{s}(f(X),Y)$ and $S(1) \leftarrow *$, where $S(r)$ refers to the score of the abstaining classifier when $R=r$ for each $r \in \{0,1\}$. We do not use this notation in our main paper because $S(1)$ is not meaningful in our case.

A.3 Comparison with Condessa et al.'s score

To better understand the counterfactual score $\psi = \mathbb{E}[S]$, we can contrast it with Condessa et al. (2017)'s notion of the 'classification quality score' θ. Assuming that $S \in [0,1]$, their *classification quality score* θ can be defined as follows:

$$\theta := \mathbb{E}\left[S \mid R=0\right]\mathbb{P}(R=0) + \mathbb{E}\left[1-S \mid R=1\right]\mathbb{P}(R=1). \tag{A.1}$$

In contrast, note that the counterfactual score (2.1) is decomposed into

$$\psi = \mathbb{E}\left[S \mid R=0\right]\mathbb{P}(R=0) + \mathbb{E}\left[S \mid R=1\right]\mathbb{P}(R=1). \tag{A.2}$$

Thus, our target quantity ψ is large if the classifier is good on all inputs (abstentions or not), while θ is large if the classifier is good on points it predicts on but poor on points it abstains on.

A concrete example To further elucidate the difference in what each score measures, we include a hypothetical example. Consider comparing two abstaining classifiers A and B based on their accuracy score over $n=100$ data points, whose inputs (X_1,X_2) are sampled uniformly on $[-1,1]\times[-1,1]$.

Suppose that the base classifier for A achieves a $S^{\mathsf{A}}=1.0$ accuracy when $X_1<0$ (the "left half") but only a $S^{\mathsf{A}}=0.8$ when $X_1 \geq 0$ (the "right half"). So, it decides to abstain at an 80% rate on the right half ($\pi^{\mathsf{A}}(x_1,x_2)=0.8$ if $x_1<0$), while it does not abstain at all on the left half ($\pi^{\mathsf{A}}(x_1,x_2)=0$ if $x_1 \geq 0$). Assuming for the sake of simplicity that $n/2=50$ points are placed in each half, the classifier makes 50 (out of 50) correct predictions on the left half, while on the right half, it makes 8 (out of 10) correct predictions and 40 abstentions, for which it would have been correct 80% of the time. This classifier's overall selective accuracy is thus $(50+8)/(50+10)=58/60\approx 0.97$, while its coverage is $(50+10)/100=0.6$.

Plugging in the counts and probabilities to (A.2), we can calculate the empirical[4] counterfactual score of classifier A:

$$\hat{\psi}^{\mathsf{A}} = \frac{50+8}{50+10}\cdot\frac{1.0+0.2}{2} + \frac{0+32}{0+40}\cdot\frac{0+0.8}{2} = \frac{58}{60}\cdot 0.6 + \frac{32}{40}\cdot 0.4 = 0.9. \tag{A.3}$$

(A simpler calculation would be to use (2.1) directly, but we use the equivalent decomposition (A.2) here to contrast with (A.1).) The empirical classification quality score of classifier A is

$$\hat{\theta}^{\mathsf{A}} = \frac{58}{60}\cdot 0.6 + \left(1-\frac{32}{40}\right)\cdot 0.4 = 0.66. \tag{A.4}$$

Comparing the two scores, $\hat{\psi}^{\mathsf{A}}$ is much larger than $\hat{\theta}^{\mathsf{A}}$ because the classifier chose to abstain on 40% of the data for which it would have gotten a 0.8 accuracy.

Next, suppose that classifier B is the same as classifier A, except that it achieves a meager 0.6 accuracy on the right half. It also uses the same abstention mechanism as A. Then, analogous calculations show that classifier B's counterfactual score $\hat{\psi}^{\mathsf{B}}$ would be $56/60\cdot 0.6 + 24/40\cdot 0.4 = 0.8$, lower than $\hat{\psi}^{\mathsf{A}}$, whereas the classification quality score $\hat{\theta}^{\mathsf{B}}$ would be $56/60\cdot 0.6 + (1-24/40)\cdot 0.4 = 0.72$, higher than $\hat{\theta}^{\mathsf{A}}$. Thus, comparisons based on the two scores would lead to opposite conclusions. Note

---

[4] This is only 'empirical' up to the sampling of n data points; we do not need to estimate the counterfactuals in this hypothetical example because they are already known.

that the classification quality score rewards B for hiding its low-accuracy predictions, even though A uses the same abstention mechanism and has a better accuracy overall on the right half.

The choice between the two scores should ultimately be determined by the use case, although we focus on the counterfactual score in the main paper, motivated by the various cases we described earlier (Example 1.1 and the additional examples in Appendix A.1). In the example above, if the evaluator needs to later access the classifier's hidden predictions on the right half, then they can use the counterfactual score and choose A, which has a higher accuracy in its hidden predictions than B.

Estimation As mentioned in Section 1, Condessa et al. (2017) focus on the "white-box" setup where the hidden predictions are known to the evaluator, and it is not obvious how to estimate their score in the black-box setup. However, much like the counterfactual score ψ, the challenge of estimating θ is driven entirely by the $\mathbb{E}\left[S|R=1\right]$ term, as the remaining terms are directly observed. As the decompositions (A.1) and (A.2) show, estimates of ψ (from the DR CI in Section 3) also yield estimates of θ, since $\theta + \psi$ is an observable quantity that can be straightforwardly estimated. Subtracting an estimate of ψ from the sum gives an estimate of θ.

A.4 The plug-in and inverse propensity weighting estimators

The uniqueness of efficient influence functions tells us that the DR estimator outperforms two intuitive yet suboptimal estimators in an asymptotic and locally minimax sense. The first is the *plug-in estimator*, which is derived directly from the identified target $\psi = \mathbb{E}[\mu_0(X)]$ in Proposition 2.4:

$$\hat{\psi}_{\mathsf{pi}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_0(X_i), \tag{A.5}$$

where $\hat{\mu}_0$ is any estimate of the regression function $\mu_0(x) = \mathbb{E}\left[S \mid R=0, X=x\right]$. The quality of this simple estimator directly depends on the estimation quality of $\hat{\mu}_0$ for μ_0, and in a nonparametric setting, the estimator can suffer from the statistical curse of dimensionality. Another point of concern is that it makes no use of the missingness patterns.

The second is *inverse probability weighting (IPW)* estimator (Horvitz and Thompson, 1952; Rosenbaum, 1995):

$$\hat{\psi}_{\mathsf{ipw}} = \frac{1}{n}\sum_{i=1}^{n}\frac{(1-R_i)}{1-\hat{\pi}(X_i)}S_i, \tag{A.6}$$

where $\hat{\pi}$ is an estimate of the abstention mechanism $\pi(x) = \mathbb{P}\left(R=1 \mid X=x\right)$. If $\hat{\pi}$ consistently estimates π, the IPW estimator is unbiased; yet, it has the opposite problem to the plug-in estimator as it does not model the conditional score μ_0 at all.

A.5 Positivity and policy

Our identification results in Section 2.2 impose a requirement of positivity (Assumption 2.3) on the abstaining classifier (f, π), i.e., a demand that for some $\epsilon > 0$, the essential supremum of $\pi(x)$ is smaller than $1-\epsilon$. This requirement is necessary: intuitively, if no feedback about the behaviour of f is available in a region, it is impossible (without further strong assumptions about the global structure of f) to determine the behaviour of the score in this region. Operationally, this is seen quite directly in the validity of the confidence intervals inferred from data (Figure App.5). Of course, the parameter ϵ also plays a quantitative role: the higher the ϵ, the better the validity and widths of our CIs. In other words, our ability to identify decays gracefully with ϵ, with complete inability if $\pi(x) = 1$ in a region of large mass.

While necessary, this positivity requirement is at odds with the practical deployment of client-facing abstaining classifiers. Indeed, there are two major reasons to implement an abstaining mechanism in such scenarios. In a positive sense, abstentions signal that the use of the underlying classifier f is inappropriate in a particular domain. However, in a negative sense, abstentions can also be employed in order to artificially limit a vendor's liability when their predictions (and the actions driven by the same) are incorrect. A pertinent example is the recent investigation of the Tesla autopilot by the NHTSA (2022) which found that in 16 incidents, the autopilot would deactivate and hand-off control to the driver at the very last seconds before a crash, thus artificially inflating the safety metrics of the system.

Part of the impetus behind studying a metric such as the counterfactual score is precisely to identify such behaviours before unsafe incidents bring them to light. Nevertheless, if vendors can stymie this investigation simply by ensuring that abstention is accompanied by a very high $\pi(x)$, then the method is not particularly useful.

This technical impasse begs for a policy-level treatment: through regulatory action, the executive may ensure that vendors supply evaluators (whether government agencies or independent reviewers) with abstaining classifiers that reveal the counterfactual decision of f at least an ϵ-fraction of the times when the decision is to abstain, where ϵ is set by mutual agreement of the stakeholders. Note that it is not enough to just supply evaluators with the predictions of f (although this would solve our particular problem formulation), since it is important to understand its behaviour in the context of when the abstaining classifier actually tends to reject points (i.e., it is equally important for evaluators and users to understand $\mathbb{E}\left[S \mid R=1\right]$, which of course is estimable under our setup).

A.6 Extension to learning-to-defer settings

In the learning-to-defer setting involving an expert (Madras et al., 2018), the counterfactual score would refer to the expected score of the overall system had the classifier not deferred at all. The counterfactual score is thus an evaluation metric primarily for the classifier, and it is independent of the expert's predictions, even when the classifier is adaptive to the expert's tendencies.

On the other hand, in the case where the goal is to assess the *joint* performance of the algorithm and the expert, then it may be useful to estimate a variation of the classification quality score (A.1) defined in Appendix A.3. If we denote the expert's score as E, then equation (A.1) can further be generalized to

$$\theta^E := \mathbb{E}[S \mid R=0]\mathbb{P}(R=0) + \mathbb{E}[E-S \mid R=1]\mathbb{P}(R=1). \tag{A.7}$$

For each rejection ($R=1$), the score would assess the system by the difference in the quality of expert prediction and the model prediction ($E-S$); in the black-box evaluation case, the model prediction score in the case of deferral (S given $R=1$) is a counterfactual.

Finally, the estimation approach from Appendix A.3 still applies to θ^E. If the expert is an oracle ($E=1$), then θ^E coincides with the classification quality score θ (A.1). Even if the expert's predictions are random, $\mathbb{E}[E \mid R=1]$ is an observable quantity and θ^E can be re-written as $\theta + \mathbb{E}[E-1 \mid R=1]\mathbb{P}(R=1)$, so θ^E can be estimated within our counterfactual framework.

B Proofs

B.1 Proof of Proposition 2.2

Since (X,Y) is independent of the training data $\mathcal{D}_{\text{train}}$ for (f,π), and because ξ is an independent source of randomness, we can treat the functions f and π as fixed. Then, by definition, $S = \mathsf{s}(f(X),Y)$ is a deterministic function of (X,Y) and $R = \mathsf{r}(\pi(X),\xi)$ is a deterministic function of X and ξ. This means that the condition $S \perp\!\!\!\perp R \mid X$ is equivalent to saying that $Y \perp\!\!\!\perp \xi \mid X$. Given that ξ is independent of (X,Y), the latter condition follows.

B.2 Proof of Proposition 2.4

Positivity (Assumption 2.3) ensures that the conditional expectation $\mu_0(X) = \mathbb{E}\left[S \mid R=0, X\right]$ is well-defined. Then,

$$\mathbb{E}\left[\mu_0(X)\right] = \mathbb{E}\left[\mathbb{E}\left[S \mid R=0, X\right]\right] \overset{\text{(MAR)}}{=} \mathbb{E}\left[\mathbb{E}\left[S \mid X\right]\right] = \mathbb{E}\left[S\right] = \psi, \tag{B.1}$$

where the second inequality follows from the MAR condition (Assumption 2.1), i.e., $S \perp\!\!\!\perp R \mid X$.

B.3 Proof Sketch of Theorem 3.1

We follow the relevant notations and derivations from Kennedy (2022). Denote $\mathbb{P}\left\{f\right\} = \mathbb{E}_{\mathbb{P}}\left[f(Z)\right]$ and $\mathbb{P}_n\left\{f\right\} = n^{-1}\sum_{i=1}^{n} f(Z_i)$ where $Z_i \overset{iid}{\sim} \mathbb{P}$. We use the *centered* influence function for $\psi(\mathbb{P}) =$

$\mathbb{E}_{\mathbb{P}}\left[\mu_0(X)\right]$ (upon identification), defined as follows:

$$\mathsf{IF}_{\mathbb{P}}(x,r,s) := \left[\frac{1-r}{1-\pi(x)}\left(s-\mu_0(x)\right)+\mu_0(x)\right]-\psi(\mathbb{P}). \tag{B.2}$$

Here, $\mathsf{IF}_{\mathbb{P}}$ depends on $\mathbb{P}$, which determines π and μ_0. Analogously, we let $\hat{\mathbb{P}}$ denote the distribution of abstentions and score outcomes involving estimators $\hat{\pi}$ and $\hat{\mu}_0$ (in place of π and μ_0), and let $\mathsf{IF}_{\hat{\mathbb{P}}}$ and $\psi(\hat{\mathbb{P}})$ denote the corresponding influence function and target functional, respectively, defined using $\hat{\pi}$ and $\hat{\mu}_0$. Also, note that an uncentered version is shown in the main text for ease of explanation; the resulting variance does not change due to this centering. Using these definitions, we proceed with the proof in two steps.

Step 1: Showing that IF (B.2) is the efficient influence function for ψ To show that IF is indeed the unique efficient influence function for ψ, we show that $\mathbb{P}\{\mathsf{IF}_{\mathbb{P}}\}=0$ and that its bias term is second-order. The uniqueness and asymptotic efficiency of this EIF in a nonparametric setting, in general, is well-known (e.g., van der Vaart (2002)). First, observe that

$$\begin{aligned}\mathbb{P}\left\{\mathsf{IF}_{\mathbb{P}}\right\} &= \mathbb{E}_{\mathbb{P}}\left[\frac{1-R}{1-\pi(X)}\left(S-\mu_0(X)\right)+\mu_0(X)\right]-\psi(\mathbb{P}) &\text{(B.3)}\\ &= \mathbb{E}_{\mathbb{P}}\left[\frac{\mathbb{E}\left[(1-R)(S-\mu_0(X))\mid X\right]}{1-\pi(X)}\right] &\text{(B.4)}\\ &\overset{\text{(a)}}{=} 0, &\text{(B.5)}\end{aligned}$$

where (a) follows from the fact that

$$\mathbb{E}\left[(1-R)S\mid X\right]=\pi(X)\cdot 0+(1-\pi(X))\mathbb{E}\left[S\mid R=0,X\right]=(1-\pi(X))\mu_0(X). \tag{B.6}$$

Furthermore, for any distributions $\hat{\mathbb{P}}$ and $\mathbb{P}$, the bias term is given by

$$\begin{aligned}R_2(\hat{\mathbb{P}},\mathbb{P}) &= \psi(\hat{\mathbb{P}})-\psi(\mathbb{P})+\mathbb{P}\left\{\mathsf{IF}_{\hat{\mathbb{P}}}\right\} &\text{(B.7)}\\ &= \psi(\hat{\mathbb{P}})-\psi(\mathbb{P})+\mathbb{E}_{\mathbb{P}}\left[\frac{1-R}{1-\hat{\pi}(X)}\left(S-\hat{\mu}_0(X)\right)+\hat{\mu}_0(X)\right]-\psi(\hat{\mathbb{P}}) &\text{(B.8)}\\ &= \mathbb{E}_{\mathbb{P}}\left[\frac{1-R}{1-\hat{\pi}(X)}\left(S-\hat{\mu}_0(X)\right)+\hat{\mu}_0(X)-\mu_0(X)\right] &\text{(B.9)}\\ &\overset{\text{(IE,a)}}{=} \mathbb{E}_{\mathbb{P}}\left[\frac{1-\pi(X)}{1-\hat{\pi}(X)}\left(\mu_0(X)-\hat{\mu}_0(X)\right)-\left(\mu_0(X)-\hat{\mu}_0(X)\right)\right] &\text{(B.10)}\\ &= \mathbb{E}_{\mathbb{P}}\left[\frac{\left(\hat{\pi}(X)-\pi(X)\right)\left(\mu_0(X)-\hat{\mu}_0(X)\right)}{1-\hat{\pi}(X)}\right] &\text{(B.11)}\\ &\leq \frac{1}{\epsilon}\cdot\left\|\hat{\pi}-\pi\right\|_{L_2(\mathbb{P})}\left\|\hat{\mu}_0-\mu_0\right\|_{L_2(\mathbb{P})}. &\text{(B.12)}\end{aligned}$$

This is a second-order product term in the difference of $\hat{\mathbb{P}}$ and $\mathbb{P}$, showing that IF is an influence function for $\mathbb{P}$.

Step 2: Showing the asymptotic normality of $\sqrt{n}(\hat{\psi}_{\mathsf{dr}}-\psi)$ To derive the explicit form of the limiting distribution, denote $\hat{\mathsf{IF}}=\mathsf{IF}_{\hat{\mathbb{P}}}$, and observe that the DR estimator is a “one-step” bias-corrected estimator (Bickel, 1975), given by $\hat{\psi}_{\mathsf{dr}}=\mathbb{P}_n\{\hat{\mathsf{IF}}\}+\psi(\hat{\mathbb{P}})$. Then, we have the following three-term decomposition:

$$\begin{aligned}\hat{\psi}_{\mathsf{dr}}-\psi &= \mathbb{P}_n\left\{\hat{\mathsf{IF}}\right\}+\psi(\hat{\mathbb{P}})-\psi(\mathbb{P}) &\text{(B.13)}\\ &= (\mathbb{P}_n-\mathbb{P})\left\{\hat{\mathsf{IF}}\right\}+R_2(\hat{\mathbb{P}},\mathbb{P}) &\text{(B.14)}\\ &= (\mathbb{P}_n-\mathbb{P})\left\{\mathsf{IF}\right\}+(\mathbb{P}_n-\mathbb{P})\left\{\hat{\mathsf{IF}}-\mathsf{IF}\right\}+R_2(\hat{\mathbb{P}},\mathbb{P}). &\text{(B.15)}\end{aligned}$$

The first term, which is a sample average term, has the desired limiting distribution by the central limit theorem:

$$\sqrt{n} \cdot (\mathbb{P}_n - \mathbb{P})\{\mathsf{IF}\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\mathsf{IF}(Z_i) - \mathbb{E}_{\mathbb{P}}[\mathsf{IF}(Z)]] \rightsquigarrow \mathcal{N}(0, \mathsf{Var}_{\mathbb{P}}[\mathsf{IF}]). \tag{B.16}$$

Then, by Slutsky's theorem, it suffices to show that the other two terms are of order $o_{\mathbb{P}}(1/\sqrt{n})$. The third term, $R_2(\hat{\mathbb{P}}, \mathbb{P})$, is precisely the second-order bias term we derived in (B.12), and it is $o_{\mathbb{P}}(1/\sqrt{n})$ by the DR assumption (3.2).

The second term, called the empirical process term, can be shown to be of order $o_{\mathbb{P}}(1/\sqrt{n})$ when using cross-fitting to estimate $\hat{\mathbb{P}}$. Specifically, the sample splitting procedure guarantees that $\hat{\mathbb{P}} \perp\!\!\!\perp \mathbb{P}_n$ (where $\mathbb{P}_n$ now refers to the held-out fold in each step of cross-fitting), which is enough to show that

$$(\mathbb{P}_n - \mathbb{P})\left\{\hat{\mathsf{IF}} - \mathsf{IF}\right\} = O_{\mathbb{P}}\left(\frac{\|\hat{\mathsf{IF}} - \mathsf{IF}\|_{L^2(\mathbb{P})}}{\sqrt{n}}\right). \tag{B.17}$$

Since $\|\hat{\mathsf{IF}} - \mathsf{IF}\|_{L^2(\mathbb{P})} = o_{\mathbb{P}}(1)$ by assumption, the term itself is of order $o_{\mathbb{P}}(1/\sqrt{n})$ as desired. The loss of sample efficiency due to a single sample splitting can be recovered by the cross-fitting procedure. See, e.g., Lemma 1 and Proposition 1 of Kennedy (2022) for details.

B.4 Proof of Theorem 3.2

Given that $\mathsf{IF}_{\mathbb{P}}^{\mathsf{AB}} = \mathsf{IF}_{\mathbb{P}}^{\mathsf{A}} - \mathsf{IF}_{\mathbb{P}}^{\mathsf{B}}$, it is immediate that it is an influence function for $\Delta^{\mathsf{AB}} = \psi^{\mathsf{A}} - \psi^{\mathsf{B}}$ because $\mathbb{P}\{\mathsf{IF}_{\mathbb{P}}^{\mathsf{AB}}\} = \mathbb{P}\{\mathsf{IF}_{\mathbb{P}}^{\mathsf{A}}\} - \mathbb{P}\{\mathsf{IF}_{\mathbb{P}}^{\mathsf{B}}\} = 0$ and

$$R_2(\hat{\mathbb{P}}, \mathbb{P}) \le \frac{1}{\epsilon} \cdot \left(\|\hat{\pi}_{\mathsf{A}} - \pi_{\mathsf{A}}\|_{L_2(\mathbb{P})} \|\hat{\mu}_{0,\mathsf{A}} - \mu_{0,\mathsf{A}}\|_{L_2(\mathbb{P})} + \|\hat{\pi}_{\mathsf{B}} - \pi_{\mathsf{B}}\|_{L_2(\mathbb{P})} \|\hat{\mu}_{0,\mathsf{B}} - \mu_{0,\mathsf{B}}\|_{L_2(\mathbb{P})} \right). \tag{B.18}$$

The limiting distribution can also be derived analogously, where the upper bound in (B.18) reveals the additive form of the DR assumption (3.4).

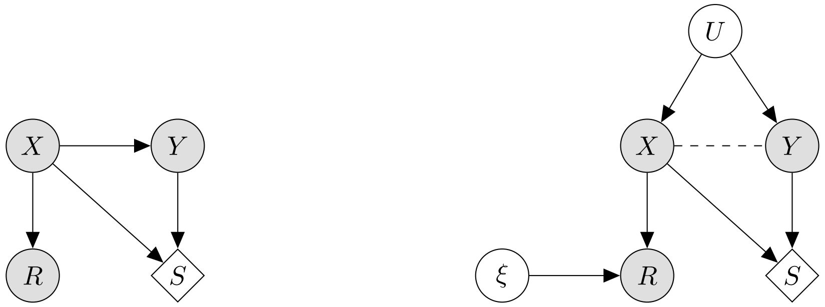
C Illustration of the MAR condition via causal graphs

Intuitively, the MAR condition is satisfied as long as the evaluation label is unknown to either classifier, simply because the classifier cannot access the actual score $S = \mathsf{s}(f(X), Y)$, which is a function of the true label Y, in making its abstention decision. This already implies $P(R = 1|S, X) = P(R = 1|X)$. We can further elucidate how the causal relationships between the random variables in our setup, and highlight how the MAR condition is generally satisfied, via graphical representations of the evaluation setup. The comparison case is an analogous extension to two abstaining classifiers.

Assuming that the abstaining classifier (f, π) does not depend on the evaluation output label Y, we (the evaluator) can treat both functions as fixed given the input X. We can then illustrate the MAR condition via two causal graphs. First, suppose $X \to Y$ (for the sake of simplification). Then, we have the relationships $X \to Y$, $X \to R$ (Bernoulli with probability $\pi(X)$), and $(X, Y) \to S$ (deterministic via f and s). In the resulting graph, shown in Figure App.1a, the variables S and R are d-separated (Pearl, 2000) given X, i.e., $S \perp\!\!\!\perp R \mid X$. Note that S is partially observed and thus drawn as a diamond node, but it does not affect the conditional independence relationship. An alternative representation is possible via missingness graphs (Mohan et al., 2013), which would give us the same conclusion.

Next, we can remove the assumption on the relationship $X \to Y$, and allow any possible relationship between X and Y: $X \to Y$ (causal), $Y \to X$ (anticausal), or $U \to (X, Y)$, where U is an unobserved confounder to the prediction task. This is depicted as a dashed line between X and Y, along with a possible presence of U, in Figure App.1b. We can further allow the abstaining classifier to utilize some internal randomness or bias ξ, which is independent of the randomness in evaluation data, for its decision to abstain R. In the resulting graph, shown in Figure App.1b, none of the generalizations change the fact that S and R are d-separated given X, i.e., the MAR condition is satisfied.

Finally, as mentioned in the main text, the MAR condition can be violated when the evaluation data is not independent of the training data. For example, if the true label Y is used by the abstaining



(a) Simple DAG representation of our setup, assuming $X \to Y$.

(b) A more general graph that allows arbitrary relationships between X and Y as well as the classifier's internal randomness/bias (ξ).

Figure App.1: Two graphical representations of the random variables involved in our evaluation framework from Section 2, assuming that the true label Y is independent of the abstaining classifier (f, π). Shaded variables are observed by the evaluator; the score $S = \mathsf{s}(f(X), Y)$ is *partially* observed by the evaluator (depicted as a diamond node). In plot App.1a, assuming $X \to Y$, the simple DAG illustrates that S and R are d-separated given X. In plot App.1b, we further allow arbitrary relationships between X and Y, including $X \to Y$, $Y \to X$, and $U \to (X, Y)$ for some unobserved confounder U. The classifier's decision to abstain, R, is also allowed to additionally depend on some internal randomness and bias ξ that is independent of the evaluation data. Accounting for these generalizations, S and R are still d-separated given X, irrespective of the causal direction (if any) between X and Y.

classifier during its training to inform its abstention decision, then this would correspond to a graph in which there is an additional edge from Y to R, as the abstention function π now depends on Y. Then, S and R are no longer d-separated because there is a now connecting path via Y (common cause).

D Confidence sequences for anytime-valid counterfactual score estimation

The nonparametric efficiency result of Theorem 3.1 yields an optimal inference procedure (either a hypothesis test or a confidence interval) for evaluating and comparing abstaining classifiers at a fixed sample size. Here, we go one step further and utilize a *confidence sequence (CS)* (Darling and Robbins, 1967; Howard et al., 2021), which is a sequence of confidence intervals whose validity holds uniformly over all sample sizes. This *time-uniform* property allows the evaluator to continuously monitor the result as more data is collected over time. The time-uniform property also implies *anytime-validity* (Johari et al., 2022; Grünwald et al., 2023), which allows the evaluator to run the experiment without pre-specifying the size of the evaluation set and compute the CIs as more data is collected. This implies that anytime-valid methods avoid the issue of inflated miscoverage rates coming from "data peeking." See Ramdas et al. (2023) for an introduction.

Formally, for any $\alpha \in (0, 1)$, a $(1 - \alpha)$-level (non-asymptotic) CS $(C_t)_{t \geq 1}$ for a parameter $\theta \in \mathbb{R}$ is a sequence of confidence intervals (CI) such that

$$\mathbb{P}\left(\forall t \geq 1 : \theta \in C_t\right) \geq 1 - \alpha. \tag{D.1}$$

Importantly, a CS contrasts with a fixed-time CI, whose guarantee no longer remains valid at stopping times: a CI only satisfies $\mathbb{P}\left(\theta \in C_t\right) \geq 1 - \alpha$ for a fixed sample size t.

Here, we describe how we can perform the proposed counterfactual comparison of abstaining classifiers using a variant of a CS that is asymptotic and readily applicable to causal estimands (Waudby-Smith et al., 2021). An (two-sided) $(1 - \alpha)$-*asymptotic CS (AsympCS)* $(\tilde{C}_t)_{t \geq 1}$ for a parameter $\theta \in \mathbb{R}$ is a sequence of intervals, $\tilde{C}_t = (\hat{\theta}_t \pm \tilde{B}_t)$, for which there exists a non-asymptotic CS $(C_t)_{t \geq 1}$ for θ of the form $C_t = (\hat{\theta}_t \pm B_t)$ that satisfies

$$B_t / \tilde{B}_t \xrightarrow{\text{a.s.}} 1. \tag{D.2}$$

The AsympCS has an *approximation rate* of r_t if $\tilde{B}_t - B_t = O(r_t)$ almost surely.

Intuitively, an AsympCS is an arbitrarily precise approximation of a non-asymptotic CS. Because no known non-asymptotic CS exists for counterfactual quantities such as the ATE, AsympCS has been derived as an (only) viable alternative. Waudby-Smith et al. (2021) further leverage the (previously described) nonparametric efficiency theory and doubly robust estimation to derive an AsympCS for the ATE in randomized experiments and observational studies; we apply their theory to estimating the counterfactual scores and their differences. The resulting AsympCS is asymptotically time-uniform and anytime-valid, and its width scales similarly, up to logarithmic factors, to a fixed-time CI derived directly from Theorem 3.1.

Now we describe our main theorem for anytime-valid and counterfactual evaluation of an abstaining classifier. We consider evaluating the classifier on an i.i.d. test set that is continuously collected over time; let n be the (data-dependent) sample size with which inference is performed. As before, the comparison problem reduces to evaluating each abstaining classifier and taking their difference. We suppose that the nuisance functions $\hat{\pi}$ and $\hat{\mu}_0$ are learned via cross-fitting, and these are used to compute the EIF estimate (3.1). Now we can formally state an asymptotic CS for $\psi = \mathbb{E}[S]$ (2.1) that is anytime-valid and doubly robust. In the below, the $o(\cdot)$ notation refers to almost sure convergence.

Theorem D.1 (Anytime-valid DR estimation of the counterfactual score)**.** *Suppose that $\hat{\mu}_0$ and $\hat{\pi}$ consistently estimates μ_0 and π in $L_2(\mathbb{P})$, respectively, at a product rate of $o(\sqrt{\log\log n/n})$:*

$$\|\hat{\mu}_0 - \mu_0\|_{L_2(\mathbb{P})} \|\hat{\pi} - \pi\|_{L_2(\mathbb{P})} = o(\sqrt{\log\log n/n}). \tag{D.3}$$

Also, suppose that $\|\hat{\mathsf{IF}} - \mathsf{IF}\|_{L_2(\mathbb{P})} = o(1)$ and that IF has at least four finite moments.

Then, under Assumption 2.1 and 2.3, for any choice of $\rho > 0$,

$$\hat{\psi}_{\mathsf{dr}} \pm \sqrt{\hat{\mathsf{Var}}_n\left(\hat{\mathsf{IF}}\right)} \cdot \sqrt{\frac{2n\rho^2+1}{n^2\rho^2}\log\left(\frac{\sqrt{n\rho^2+1}}{\alpha}\right)} \tag{D.4}$$

forms a $(1-\alpha)$-AsympCS for ψ with an approximation rate of $\sqrt{\log\log n/n}$.

This result is an adaptation of Theorems 2.2 and 3.2 in Waudby-Smith et al. (2021) to our setup. The assumptions on $\hat{\pi}$ and $\hat{\mu}_0$ are analogous to the double robustness assumptions (3.2) in Theorem 3.1, as they require the same product rate up to logarithmic factors. Here, ρ is a free parameter that can be chosen to optimize the CS width (see Appendix C.3 of Waudby-Smith et al. (2021) for details).

Compared to the fixed-size CI of (3.1), whose width shrinks at a $O(1/\sqrt{n})$ rate, the width of the AsympCS in (D.4) shrinks at a $O(\sqrt{\log n/n})$ rate. This means that, in terms of the CI width, the extra cost of ensuring anytime-validity is logarithmic in n. In practice, the AsympCS may be wider than the CI from Theorem 3.1; nevertheless, the AsympCS may be preferred in scenarios where the evaluation/comparison is performed on continuously collected data. Another potential benefit of the AsympCS is the extension to settings with sequential and time-varying evaluation tasks (e.g., involving time-series forecasters that abstain). We leave the formalization of the time-varying setup as future work.

Finally, to apply Theorem D.1 to a comparison setting, we can construct two $(1-\alpha/2)$-AsympCSs, $C_n^{\mathsf{A}} = (L_n^{\mathsf{A}}, U_n^{\mathsf{A}})$ and $C_n^{\mathsf{B}} = (L_n^{\mathsf{B}}, U_n^{\mathsf{B}})$ for ψ^{A} and B respectively, and then combine them into one $(1-\alpha)$-AsympCS for $\Delta^{\mathsf{AB}} = \psi^{\mathsf{A}} - \psi^{\mathsf{B}}$ via $C_n^{\mathsf{AB}} = (L_n^{\mathsf{A}} - U_n^{\mathsf{B}}, U_n^{\mathsf{A}} - L_n^{\mathsf{B}})$.

E Additional experiments and details

E.1 Details on the simulated data and abstaining classifiers

The evaluation set is generated as follows: $(X_{0i}, X_{1i}) \sim \mathsf{Unif}[0,1]$, $E_i \sim \mathsf{Ber}(0.15)$, and $Y_i = \mathbf{1}(X_{0i} + X_{1i} \geq 1)$ if $E_i = 0$ and $Y_i = \mathbf{1}(X_{0i} + X_{1i} < 1)$ otherwise (label noise). Classifier A uses a logistic regression model with the optimal linear decision boundary, i.e., $f^{\mathsf{A}}(x_0, x_1) = \sigma(x_0 + x_1 - 1)$, where $\sigma(u) = 1/(1+\exp(-u))$, achieving an accuracy of 0.85 by design. Classifier B, on the other hand, has a (suboptimal) curved boundary: $f^{\mathsf{B}}(x_0, x_1) = 0 \vee (\frac{1}{2}(x_0^2 + x_1^2) + \frac{1}{10}) \wedge 1$. Classifier A is thus "oracle" logistic regression model with the same decision boundary, achieving an empirical score of 0.86 before abstentions; classifier B is a biased model that achieves an empirical score of 0.74 before abstentions.

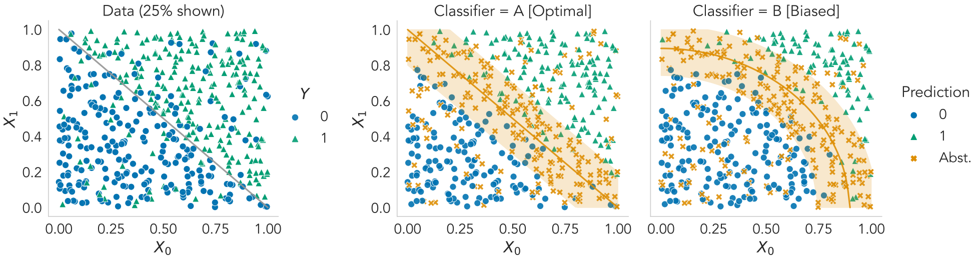


Figure App.2: A simulated example where we compare two hypothetical abstaining classifiers. The left plot shows a binary classification dataset (25% shown) in which the true decision boundary is linear. The two plots on the right show both the predictions (blue circles for 0; green triangles for 1) and the abstentions (orange x's) of two classifiers: A, which has the optimal linear boundary, and B, which has the biased nonlinear boundary. Both classifiers abstain w.p. $1-\epsilon$ in the shaded (orange) region near the decision boundary and w.p. ϵ outside the region. For both classifiers, ϵ is set to 0.2 (positivity is satisfied). Because the abstention mechanism of either classifier is determined by the input, it is not uniformly spread out across the input domain (MAR). As a result, the difference in *selective scores*, i.e., $\mathbb{E}[S^{\mathsf{A}} \mid R^{\mathsf{A}}=0] - \mathbb{E}[S^{\mathsf{B}} \mid R^{\mathsf{B}}=0] \approx 0.044$, is substantially smaller than the difference in the *counterfactual scores*, i.e., $\Delta^{\mathsf{AB}} = \mathbb{E}[S^{\mathsf{A}} - S^{\mathsf{B}}] \approx 0.116$. Our 95% DR CI for Δ^{AB} the yields $(0.077, 0.145)$, using $n = 2,000$.

For both classifiers, $\epsilon = 0.2$ determines the coefficient for positivity, and they are designed to abstain more frequently near their decision boundaries. For classifier A, $\pi^{\mathsf{A}}(x) = 1-\epsilon$ if the distance from x to its boundary is less than δ, and $\pi^{\mathsf{A}}(x) = \epsilon$ otherwise; for classifier B, we use 0.8δ as the threshold, resulting in less abstentions than A. In some sense, this is a setting where ϵ-positivity is "minimally" satisfied because the abstention rate is always either ϵ or $1-\epsilon$, and not in between, in all regions of the input space. If, say, the abstention rate was 0.5 in most parts but ϵ in a small region, the positivity level would still be ϵ but the estimation would in general be easier. Thus, this example can be viewed as a more challenging case than a standard causal inference setup with small regions of ϵ-positivity.

Figure App.2 shows both the predictions (blue circles: 0, green triangles: 1) and the abstention decisions (orange x's: predictions) for each classifier. Each classifier has a high chance of abstaining near its boundary (shaded orange region) and a low chance otherwise, meaning that abstentions are *not* spread out uniformly (MAR but not MCAR). In particular, classifier B hides many of its misclassifications as abstentions, leading to its high selective score ($\mathsf{Sel}^{\mathsf{B}} = 0.81$) relative to its counterfactual score ($\psi^{\mathsf{B}} = 0.74$).

The nuisance functions $\hat{\pi}$ and $\hat{\mu}_0$ for each classifier A and B are learned via 2-fold cross-fitting. In each case, we cap extreme propensity predictions by $\hat{\pi}^{\mathsf{A}}$ and $\hat{\pi}^{\mathsf{B}}$ are capped at $1-\epsilon$.

On a 128-core CPU machine, using parallel processing, the entire compute time it took to produce Table 2 was approximately 5 minutes.

E.2 Power analysis

To examine the efficiency of the DR estimator, we now analyze the power of the statistical test for $H_0 : \Delta^{\mathsf{AB}} = 0$ vs. $H_1 : \Delta^{\mathsf{AB}} \neq 0$ by inverting the DR CI. For different values of the sample size and the underlying performance gap, we compute the rejection rate of the statistical test across 1,000 runs. As before, the classifier A represents the oracle classifier that has the optimal decision boundary, which is linear, but the classifier B now uses a linear decision boundary that is shifted from the optimal one by a fixed amount, thereby shifting Δ^{AB} away from zero. As such, B performs increasingly worse as Δ^{AB} increases.

To increasingly vary the counterfactual score difference between two classifiers, we set A as the same classifier as in Section E.1 and set B to use the (optimal) linear decision boundary of A shifted diagonally by a fixed amount μ. Specifically, $f^{\mathsf{B}}(x_0, x_1) = \sigma(x_0 + x_1 - (1+\mu))$. An example with $\mu = 0.2$ is shown in Figure App.3. While Δ^{AB} is not strictly a linear function of μ, it is gradually increasing as μ increases, as shown in Table App.1. Aside from this difference, both classifiers use

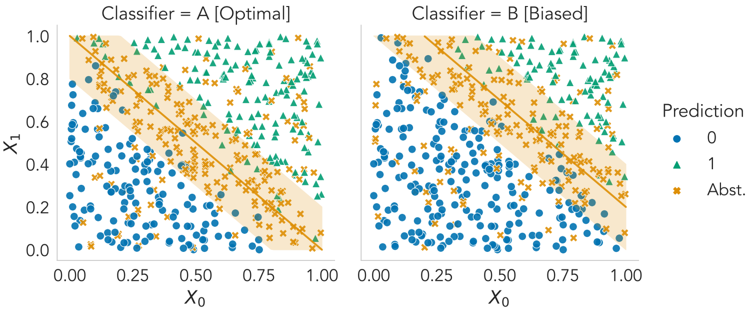


Figure App.3: A simulated example for the power experiment in which $\Delta^{\mathsf{AB}} = 0.123$. The evaluation data is the same as the one in Figure App.2. For B, the decision boundary of A is shifted diagonally upwards by $\mu = 0.2$; in the power experiment, we experiment with various values of μ (and thus Δ^{AB}).

Table App.1: The relationship between Δ^{AB} and μ, the distance between the linear decision boundaries of A and B, in the power experiment of Section E.2.

Δ^{AB}	0.0	0.045	0.069	0.088	0.123	0.152	0.180	0.181	0.219	0.248	0.271
μ	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50

the same abstention mechanism as classifier A from the previous experiment, and the data generating process is also identical to the previous experiment.

Figure App.4 plots the rejection rates of the level-α statistical test, for $\alpha = 0.05$, against different values of Δ^{AB} (0 to 0.27) for various sample sizes ($n = 400, 800, 1600, 3200$). Here, we plot the miscoverage rate as a function of the resulting values of Δ^{AB} directly. We use the super learner to learn the nuisance functions. Overall, we see that as n or Δ^{AB} increases, the power of the statistical test quickly approaches 1, implying that the test can consistently detect a gap in counterfactual scores if either the sample size or the difference gets large.

On a 128-core CPU machine, using parallel processing, the entire compute time it took to produce Figure App.4 was approximately 88 minutes.

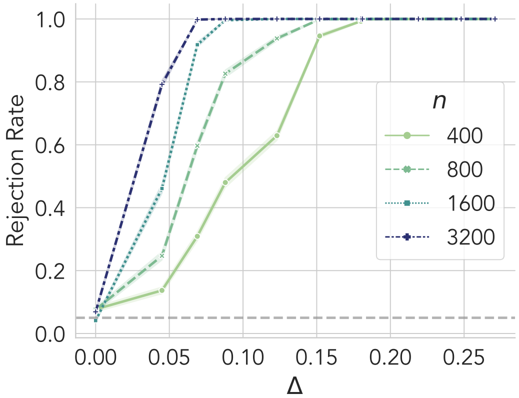


Figure App.4: Power of the statistical test for $H_0 : \Delta^{\mathsf{AB}} = 0$ derived by our 95% DR CIs, plotted for different values of n (sample size) and Δ^{AB}, which varies based on the distance between the (linear) decision boundaries of A and B. Mean rejection rates of H_0 over 1,000 simulations are shown, with 1 standard error as shaded error bars. As either n or Δ^{AB} grows large, the power approaches 1.

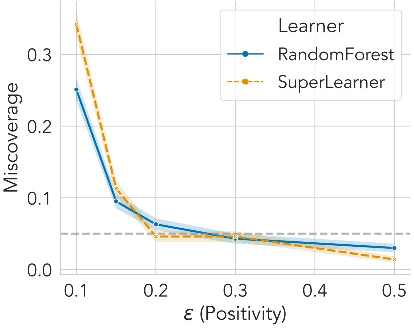


Figure App.5: Miscoverage rates of 95% doubly robust CIs by varying the level of ϵ (positivity), plotted for different nuisance function learners. Each point is the mean over 1,000 repeated simulations; shaded error bars represent 1 standard error.

E.3 Details on the CIFAR-100 experiment

The abstaining classifiers compared in the experiments are variants of the VGG-16 CNN model with batch normalization (Simonyan and Zisserman, 2015). Specifically, the feature representation layers are obtained from a model[5] trained on the training set of the CIFAR-100 dataset and are fixed during evaluation. Using half ($n = 5,000$) of the validation set, we train a L2-regularized softmax output layer and its softmax response (SR) for the abstention mechanism. The comparison is done on the other half ($n = 5,000$) of the validation set. This version of the VGG-16 features and the softmax layer is used for all scenarios, with different abstention mechanisms described in the main text, except for the last comparison, where we compare this softmax layer with VGG-16's original 3-layer output model (2 hidden layers of size 512).

The nuisance functions, $\hat{\pi}$ and $\hat{\mu}_0$ for each classifier in each scenario, also utilize the pre-trained representations of the VGG-16 layer, but their output layers (both L2-regularized linear models) are trained separately via cross-fitting.

The pre-trained VGG-16 features on the CIFAR-100 validation set were first obtained using a single NVIDIA A100 GPU, taking approximately 20 seconds. On a 128-core CPU machine, using parallel processing, the rest of the computation to produce Table 3 took less than 10 seconds (note that there are no repeated runs in this experiment).

E.4 Sensitivity to different positivity levels

Here, we examine how the DR estimator is affected by the level of positivity, i.e., ϵ in (2.3). As discussed in the main paper, positivity violations make it infeasible to properly identify and estimate causal estimands. In practice, we expect the DR estimator to remain valid up until ϵ becomes smaller than a certain (small) number. To validate this, we use the same setting from our first experiment (Section 4.1; Appendix E.1) but vary the level of positivity from $\epsilon = 0.5$ (MCAR) to $\epsilon = 0.1$ (positivity near-violation).

Figure App.5 plots the miscoverage rate of the DR estimator, averaged over 1,000 repeated simulations, using the three nuisance learner choices we used in Section 4.1. The result confirms that the DR estimator, when using either the random forest or the super learner, retains validity as long as $\epsilon \geq 0.2$, in this particular case; as ϵ shrinks to below 0.2, the miscoverage rates start to go above the significance level. This confirms that there is a (problem-dependent) level of positivity we must expect for the DR estimator to work; otherwise, we do not expect the counterfactual target to be a meaningfully identifiable quantity in the first place.

On a 128-core CPU machine, using parallel processing, the entire compute time it took to produce Figure App.5 was approximately 12 minutes.

---

[5] `https://github.com/chenyaofo/pytorch-cifar-models`

References

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *Journal of the American Statistical Association*, 70(350):428–434.

Busby, L. P., Courtier, J. L., and Glastonbury, C. M. (2018). Bias in radiology: the how and why of misses and misinterpretations. *Radiographics*, 38(1):236–247.

Condessa, F., Bioucas-Dias, J., and Kovačević, J. (2017). Performance measures for classification systems with rejection. *Pattern Recognition*, 63:437–450.

Darling, D. A. and Robbins, H. (1967). Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences of the United States of America*, 58(1):66.

Grünwald, P., de Heide, R., and Koolen, W. (2023). Safe testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) (to appear).*

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055 – 1080.

Johari, R., Koomen, P., Pekelis, L., and Walsh, D. (2022). Always valid inference: Continuous monitoring of A/B tests. *Operations Research*, 70(3):1806–1821.

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: A review. *arXiv preprint arXiv:2203.06469*.

Madras, D., Pitassi, T., and Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 31.

Mohan, K., Pearl, J., and Tian, J. (2013). Graphical models for inference with missing data. *Advances in Neural Information Processing Systems*, 26.

NHTSA (2022). INOA-EA22002-3184. Technical report, National Highway Traffic Safety Administration, U.S. Department of Transportation. Report on opening of an engineering analysis regarding Tesla, Inc. products by the Office of Defects Investigation. Accessed on 26th Jan, 2023.

Pearl, J. (2000). Models, reasoning and inference. *Cambridge University Press*, 19(2).

Raghu, M., Blumer, K., Corrado, G., Kleinberg, J., Obermeyer, Z., and Mullainathan, S. (2019). The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science (to appear).*

Rosenbaum, P. R. (1995). *Observational studies*. Springer.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*.

van der Vaart, A. W. (2002). Semiparametric statistics. In *Lectures on Probability Theory and Statistics (Saint-Flour, 1999)*, pages 331–457. Springer.

Vogelpohl, T., Kühn, M., Hummel, T., and Vollrath, M. (2019). Asleep at the automated wheel—sleepiness and fatigue during highly automated driving. *Accident Analysis & Prevention*, 126:70–84.

Waudby-Smith, I., Arbour, D., Sinha, R., Kennedy, E. H., and Ramdas, A. (2021). Time-uniform central limit theory, asymptotic confidence sequences, and anytime-valid causal inference. *arXiv preprint arXiv:2103.06476*.