A Additional related work

Human decision making, interplay with algorithms. Our work contributes to a vast literature on understanding how humans, and particularly human experts, make decisions. We do not attempt to provide a comprehensive summary of this work, but refer the reader to Tversky and Kahneman (1974) and Camerer and Johnson (1991) for general background. Of particular relevance for our setting is work which investigates whether humans make *systematic* mistakes in their decisions, which has been studied in the context of bail decisions (Kleinberg et al. (2017), Rambachan (2022), Lakkaraju et al. (2017) and Arnold et al. (2020)), college admissions (Kuncel et al. (2013), Dawes (1971)) and patient triage and diagnosis (Currie and MacLeod (2017), Mullainathan and Obermeyer (2019)) among others. One common theme in these works is that the decision made by the human expert will often influence the outcome of interest; for example, an emergency room doctor's initial diagnosis will inform the treatment a patient receives, which subsequently affects their health outcomes. Furthermore, it is often the case that even observing the outcome of interest is contingent on the human's decision: for example, in a college admissions setting, we might only observe historical outcomes for *admitted* students, which makes it challenging to draw inferences about *applicants*. This one-sided labeling problem is a form of endogeneity which has been well studied in the context of causal inference, and these works often adopt a causal perspective to address these challenges.

As discussed in Section 6, our instead work assumes that all outcomes are observable and, importantly, that they are not affected by the human predictions. We also do not explicitly grapple with whether the human expert has an objective other than maximizing accuracy under a known metric (e.g., squared error). Though this is often a primary concern in many high-stakes settings – for example, ensuring that bail decisions are not only accurate but also nondiscriminatory – it is outside the scope of our work, and we refer the reader instead to Rambachan (2022) for further discussion.

As discussed in section 1, another closely related theme is directly comparing human performance to that of an algorithm (Cowgill (2018), Dawes et al. (1989), Grove et al. (2000)), and developing learning algorithms which are complementary to human expertise (Madras et al. (2018), Raghu et al. (2019), Mozannar and Sontag (2020), Keswani et al. (2021), Agrawal et al. (2018) and Bastani et al. (2021)). A key design consideration when designing algorithms to complement human expertise involves reasoning about the ways in which humans may *respond* to the introduction of an algorithm, which may be strategic (e.g. Kleinberg and Raghavan (2018), Perdomo et al. (2020), Cen and Shah (2021), Hardt et al. (2015), Liu et al. (2020)) or subject to behavioral biases (Kleinberg et al. (2022)). These behaviors can make it challenging to design algorithms which work with humans to achieve the desired outcomes, as humans may respond to algorithmic recommendations or feedback in unpredictable ways.

Conditional independence testing. We cast our setting as a special case of conditional independence testing, which has been well studied in the statistics community. For background we refer the reader to Dawid (1979). It has long been known that testing conditional independence between three (possibly high-dimensional) random variables is a challenging problem, and the recent result of Shah and Peters (2018) demonstrates that this is in fact impossible in full generality. Nonetheless, there are many methods for testing conditional independence under natural assumptions; perhaps the most popular are the kernel-based methods introduced by Fukumizu et al. (2004) and subsequently developed in Gretton et al. (2007) and Zhang et al. (2011), among others.

Our work instead takes inspiration from the 'knockoffs' framework developed in Candès et al. (2016), Barber et al. (2018) and Barber and Candès (2019), as well as the closely related conditional permutation test of Berrett et al. (2018). These works leverage the elementary observation that, under the null hypothesis that (specialized to our notation) the outcome Y and prediction $\hat{Y}$ are independent conditional on the observed data X, new samples from the distribution of $\hat{Y} \mid X$ should be exchangeable with $\hat{Y}$. Thus, if we know – or can accurately estimate – the distribution of $\hat{Y} \mid X$, it is straightforward to generate fresh samples ('knockoffs') which are statistically indistinguishable from the original data under the null hypothesis $H_0 : Y \perp\!\!\!\perp \hat{Y} \mid X$. Thus, if the observed data appears anomalous with respect to these knockoffs, this may provide us a basis on which to reject H_0.

Our work avoids takes inspiration from this framework, but avoids estimating the distribution of $\hat{Y} \mid X$ by instead leveraging a simple nearest-neighbors style algorithm for generating knockoffs. In that sense, our technique builds upon the nearest-neighbors based estimator of Runge (2017), and is nearly identical to the one-nearest-neighbor procedure proposed in the 'model-powered' conditional independence test of Sen et al. (2017). This algorithm is a subroutine in their more complicated end-to-end procedure, which involves training a model to distinguish between the observed data and knockoffs generated via swapping the 'predictions' (again specializing their general test to our setting) associated with instances which are as close as possible under the ℓ_2 norm. By contrast, we analyze a similar procedure under different smoothness assumptions which allow us to recover p-values that are entirely model free.

B Proof of Theorem 1

We establish the proof of Theorem 1 following the intuition presented in Section 3. Specifically, we first bound the type I error of **ExpertTest** in the idealized case where the data set contains L identical pairs of observations $x = x'$. We then refine this bound to handle the case, which is more likely in practice, that the pairs chosen are merely close together. Our final bound thus includes additional approximation error to account for the 'similarity' of the pairs – if we succeed in finding L pairs which are identical, we get nearly exact type I error control, whereas if we are forced to pair instances which are 'far apart', we incur additional approximation error. We formalize this intuition below.

An idealized bound. We first establish that $\mathbb{P}(\tau_K \leq \alpha) \leq \alpha + \frac{1}{K+1}$ for any $\alpha \in [0, 1]$ when $x = x'$ for every (x, x') pair chosen by **ExpertTest**.

To that end, we observe n data points $(x_i, y_i, \hat{y}_i),\ i \in [n]$. Let $\mathcal{L} = \{i_{2\ell-1}, i_{2\ell} : \ell \in [L]\}$ denote the indices of the pairs chosen by **ExpertTest**, with $(x_{i_{2\ell-1}}, x_{i_{2\ell}})$ for $\ell \in [L]$ denoting the pairs themselves.

By assumption, **ExpertTest** succeeds in finding identical pairs:

$$x_{i_{2\ell-1}} = x_{i_{2\ell}},\ \forall\, \ell \in [L]. \tag{14}$$

Therefore, from the definition (9) it follows that $r((x_{i_{2\ell-1}}, \hat{y}_{i_{2\ell-1}}), (x_{i_{2\ell}}, \hat{y}_{i_{2\ell}})) = 1$ for all $\ell \in [L]$.

As discussed in Section 3, **ExpertTest** will repeatedly generate n fresh data points, denoted by $\tilde{D}$, as follows. For each index $i \in [n] \backslash \mathcal{L}$, i.e. those not corresponding to those selected in L pairs, we select exactly the observed data $(x_i, y_i, \hat{y}_i)$.

For $i \in \mathcal{L}$, we sample a data triplet as follows: for $i \in \{i_{2\ell-1}, i_{2\ell}\}$, we let $(x_{i2\ell-1}, y_{i2\ell-1}), (x_{i2\ell}, y_{i2\ell})$ be the observed values but sample the corresponding $\hat{y}$ values from $\{(\hat{y}_{2\ell-1}, \hat{y}_{2\ell}), (\hat{y}_{2\ell}, \hat{y}_{2\ell-1})\}$ with equal probability. That is, we *swap* the $\hat{y}$ values associated with $(x_{i2\ell-1}, y_{i2\ell-1}), (x_{i2\ell}, y_{i2\ell})$ with probability $\frac{1}{2}$. We argue that this resampling process is implicitly generating a fresh, identically distributed dataset from the underlying distribution $\mathcal{D}$ conditioned on the following event $\mathcal{F}$:

$$\mathcal{F} = \{(x_i, y_i, \hat{y}_i) : i \in [n] \backslash \mathcal{L}\} \cup \{(x_i, y_i) : i \in \mathcal{L}\} \cup \{(\hat{y}_{i_{2\ell-1}}, \hat{y}_{i_{2\ell}}) \vee (\hat{y}_{i_{2\ell}}, \hat{y}_{i_{2\ell-1}}) : \ell \in [L]\}. \tag{15}$$

Why condition on $\mathcal{F}$? As discussed in section 3, a straightforward test would involve simply resampling K fresh datasets from the underlying distribution $\mathcal{D}_X \times \mathcal{D}_{\hat{Y}|X} \times \mathcal{D}_{Y|X}$ and observing that, by definition, these datasets are distributed identically to the observed data D_0 under $H_0 : Y \perp\!\!\!\perp \hat{Y} \mid X$. While this would form the basis for a valid test along the lines of the one described in Section 3, it requires knowledge of the underlying distribution which we are unlikely to have in practice. Thus, we instead condition on nearly everything in the observed data – the values and exact ordering of X and the values and exact ordering of Y, and the values of $\hat{Y}$ up to a specific set of allowed permutations (those induced by swapping 0 or more paired $\hat{y}_{i_{2\ell-1}}, \hat{y}_{i_{2\ell}}$ values). This substantially simplifies the resampling problem, as we only need to reason about the correct 'swap' probability for each such pair. This can be viewed as an alternative factorization of the underlying distribution $\mathcal{D}$ under H_0 – rather than sampling $X \sim \mathcal{D}_X, Y \sim \mathcal{D}_{Y|X}, \hat{Y} \sim \mathcal{D}_{\hat{Y}|X}$, instead sample an event $\mathcal{F} \sim \mathcal{D}_{\mathcal{F}}$ from the induced distribution over events of the form (15), and then sample $\hat{Y} \sim \mathcal{D}_{\hat{Y}|\mathcal{F}}$.

First, we show that conditional on $\mathcal{F}$, the resampled dataset $\tilde{D}$ and the observed dataset D_0 are indeed identically distributed under $H_0 : Y \perp\!\!\!\perp \hat{Y} \mid X$ (that they are also independent, conditional on $\mathcal{F}$, is clear by construction). To see this, observe that for each $\ell \in [L]$:

$$\begin{aligned}
&\mathbb{P}((x_{i_{2\ell-1}}, y_{i_{2\ell-1}}, \hat{y}_{i_{2\ell-1}}), (x_{i_{2\ell}}, y_{i_{2\ell}}, \hat{y}_{i_{2\ell}})) && (16)\\
&\quad = \mathbb{P}(x_{i_{2\ell-1}})\mathbb{P}(y_{i_{2\ell-1}} \mid x_{i_{2\ell-1}})\mathbb{P}(\hat{y}_{i_{2\ell-1}} \mid x_{i_{2\ell-1}})\mathbb{P}(x_{i_{2\ell}})\mathbb{P}(y_{i_{2\ell}} \mid x_{i_{2\ell}})\mathbb{P}(\hat{y}_{i_{2\ell}} \mid x_{i_{2\ell}}) && (17)\\
&\quad = \mathbb{P}(x_{i_{2\ell-1}})\mathbb{P}(y_{i_{2\ell-1}} \mid x_{i_{2\ell-1}})\mathbb{P}(\hat{y}_{i_{2\ell}} \mid x_{i_{2\ell-1}})\mathbb{P}(x_{i_{2\ell}})\mathbb{P}(y_{i_{2\ell}} \mid x_{i_{2\ell}})\mathbb{P}(\hat{y}_{i_{2\ell-1}} \mid x_{i_{2\ell}}) && (18)\\
&\quad = \mathbb{P}((x_{i_{2\ell-1}}, y_{i_{2\ell-1}}, \hat{y}_{i_{2\ell}}), (x_{i_{2\ell}}, y_{i_{2\ell}}, \hat{y}_{i_{2\ell-1}})) && (19)
\end{aligned}$$

In above, (17) follows from H_0 and the assumption that the data are drawn i.i.d., and (18) follows from assumption (14) that $x_{i_{2\ell-1}} = x_{i_{2\ell}}$. By construction, the events in (16) and (19) are the only two possible outcomes after conditioning on $\mathcal{F}$, and this simple argument shows that in fact they are equally likely.

Thus, let $\tilde{D}_1, \ldots, \tilde{D}_K$ be K independent and identically distributed datasets generated by the above procedure. Let $\tilde{D}_0$ be one additional sample from this distribution, which we showed was distributed identically to D_0 under the idealized assumption (14).

As discussed in Section 3, for any real-valued function F that maps each dataset to $\mathbb{R}$, we have

$$\tau_K = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}[F(\tilde{D}_0) \lesssim F(\tilde{D}_k)] \tag{20}$$

where we use definition of $\mathbb{1}[\cdot \lesssim \cdot]$ as in (7).

Because $\tilde{D}_0, \ldots, \tilde{D}_K$ are i.i.d., and thus exchangeable, it follows that $\frac{1}{K}\sum_{k=1}^{K} \mathbb{1}[F(\tilde{D}_0) \lesssim F(\tilde{D}_k)]$ is uniformly distributed $\{0, \frac{1}{K}, \ldots, 1\}$. Therefore, with a little algebra it can be verified that for any $\alpha \in [0, 1]$, τ_K satisfies

$$\mathbb{P}_{\tilde{D}_0,\ldots,\tilde{D}_K|\mathcal{F}}(\tau_K \leq \alpha) \leq \alpha + \frac{1}{K+1}. \tag{21}$$

Because D_0 and $\tilde{D}_0$ are independent and identically distributed under (14), the same holds if we replace $\tilde{D}_0$ with D_0. Thus, **ExpertTest** provides nearly exact type I error control in the case that the idealized assumption (14) holds. This result will serve as a useful building block, as we'll now proceed to relax this assumption and bound the type I error of **ExpertTest** in terms of the total variation distance between $\tilde{D}_0$ and D_0.

Fixing the approximation. $\tilde{D}_1, \ldots, \tilde{D}_K$ are synthetically generated datasets that are independent and identically distributed. The argument above replaced the observed dataset D_0 with a resampled 'idealized' dataset $\tilde{D}_0$, which is also independent and identically distributed with respect to $\tilde{D}_1, \ldots, \tilde{D}_K$, and then used this fact to demonstrate that $\mathbb{P}_{\tilde{D}_0,\ldots,\tilde{D}_K|\mathcal{F}}(\tau \leq \alpha) \leq \alpha + \frac{1}{K+1}$. If the idealized assumption (14) holds, replacing D_0 with $\tilde{D}_0$ is immaterial as we showed the two are identically distributed conditional on $\mathcal{F}$. Of course, this assumption will not hold in general, and this is what we seek to correct next.

Let $\bar{D}_0 \sim \mathcal{D}_{\cdot|\mathcal{F}}$ be a random variable distributed according to the true underlying distribution $\mathcal{D}$, conditional on the event $\mathcal{F}$. The observed data D_0 can be interpreted as one realization of this random variable. One way to quantify the excess type I error incurred by using $\tilde{D}_0$ in place of D_0 is to bound the total variation distance between the joint distributions of $(\tilde{D}_0, \ldots \tilde{D}_K)$ and that of $(\bar{D}_0, \tilde{D}_1, \ldots \tilde{D}_K)$. Specifically, it follows from the definition of total variation distance that:

$$\mathbb{P}_{\bar{D}_0,\ldots,\tilde{D}_K|\mathcal{F}}(\tau_K \leq \alpha) \leq \mathbb{P}_{\tilde{D}_0,\ldots,\tilde{D}_K|\mathcal{F}}(\tau_K \leq \alpha) + \mathsf{TV}(\mathbb{P}_{\bar{D}_0,\ldots,\tilde{D}_K|\mathcal{F}}, \mathbb{P}_{\tilde{D}_0,\ldots,\tilde{D}_K|\mathcal{F}}), \tag{22}$$

where $\mathsf{TV}(\mathbb{P}_{\bar{D}_0,\ldots,\tilde{D}_K|\mathcal{F}}, \mathbb{P}_{\tilde{D}_0,\ldots,\tilde{D}_K|\mathcal{F}})$ denotes the total variation distance between its arguments. Due to the independence of the resampled datasets, this simplifies to:

$$\mathsf{TV}(\mathbb{P}_{\bar{D}_0,\ldots,\tilde{D}_K|\mathcal{F}}, \mathbb{P}_{\tilde{D}_0,\ldots,\tilde{D}_K|\mathcal{F}}) = \mathsf{TV}(\mathbb{P}_{\bar{D}_0|\mathcal{F}}, \mathbb{P}_{\tilde{D}_0|\mathcal{F}}). \tag{23}$$

Therefore, we need only bound the total variation distance between $\mathbb{P}_{\bar{D}_0|\mathcal{F}}$ and $\mathbb{P}_{\tilde{D}_0|\mathcal{F}}$ to conclude the proof.[6]

As defined in (11), the $\varepsilon^*_{n,L}$ provides us with a way of bounding the total variation distance between the distribution of $\bar{D}_0$ and $\tilde{D}_0$. To see this, observe that the distributions of $\tilde{D}_0$ and $\bar{D}_0$, conditioned on $\mathcal{F}$, can be described as follows. To construct $\tilde{D}_0$, we can imagine flipping L fair coins to decide the assignment of $\hat{y}_i$ in each of the $(\hat{y}_{i_{2\ell-2}}\hat{y}_{i_{2\ell}})$ pairs; if it comes up heads, we swap the observed pair $(\hat{y}_{i_{2\ell-2}}\hat{y}_{i_{2\ell}})$ and if it comes up tails we do not. The observed (x_i, y_i) as well as $\hat{y}_i$ for $i \notin \mathcal{L}$ are set in $\tilde{D}_0$ as they are observed in D_0.

$\bar{D}_0$ is constructed similarly, but we instead flip a coin with bias $(1+r((x_{i_{2\ell-1}}, \hat{y}_{i_{2\ell-1}}), (x_{i_{2\ell}}, \hat{y}_{i_{2\ell}})))^{-1}$ to decide the assignment of $(\hat{y}_{i_{2\ell-2}}\hat{y}_{i_{2\ell}})$ – again, heads indicates that we swap the observed ordering, and tails indicates that we do not.

By construction, the distributions of $\bar{D}_0$ and D_0 are identical conditioned on $\mathcal{F}$, as $r((x_{i_{2\ell-1}}, \hat{y}_{i_{2\ell-1}}), (x_{i_{2\ell}}, \hat{y}_{i_{2\ell}}))$ denotes the true relative odds of observing each of the two possible $(x, \hat{y})$ pairings. In contrast, the distribution of $\tilde{D}_0$ is different, as it was sampled using the simplifying assumption (14) – in particular, $\tilde{D}_0$ is generated assuming $r((x_{i_{2\ell-1}}, \hat{y}_{i_{2\ell-1}}), (x_{i_{2\ell}}, \hat{y}_{i_{2\ell}})) = 1$!

The difference between the biases of these coins is bounded above by $\varepsilon^*_{n,L}$. We'll use this observation, along with the following lemma, to complete the proof.

Lemma 3 (Bounding the total variation distance between i.i.d. coin flips) *Let $i \in [L]$ index a sequence of i.i.d. coin flips $u_1 \dots u_L$ each with bias p_i, and $v_1 \dots v_L$ be a sequence of i.i.d. coin flips with bias q_i. Then we can show:*

$$\mathsf{TV}((u_1 \dots u_L), (v_1 \dots v_L)) \leq 1 - (1 - \max_i |p_i - q_i|)^L \tag{24}$$

We defer the proof of lemma 3 to Appendix D. This implies that the total variation distance between $\bar{D}_0$ and $\tilde{D}_0$ is bounded above by $1 - (1 - \varepsilon^*_{n,L})^L$. This, along with (21), (22) and (23) concludes the proof of Theorem 1.

Corollary 3.1 (Weaker type I error bound)

$$\mathbb{P}(\tau \leq \alpha) \leq \alpha + \varepsilon^*_{n,L} L + \frac{1}{K+1} \tag{25}$$

Corollary 3.1 is a weaker bound than the one given in Theorem 1, but is easier to interpret and manipulate. We will make use of this fact in the following section; the proof is an immediate consequence of theorem 1 and provided in Appendix D for completeness.

C Proof of Theorem 2

To establish theorem 2, we will argue that $\varepsilon^*_{n,L}$ goes to 0 at a rate of $O(n^{-\frac{1}{d}})$. This implies that, provided $L = o(n^{\frac{1}{d}})$, the excess type I error established in theorem 1 is $o(1)$ as desired. To do this, we first show that each pair $(x_{i_{2\ell-1}}, x_{i_{2\ell}})$ chosen by **ExpertTest** will be close under the ℓ_2 norm (lemmas 4 and 5 below). We then leverage the smoothness assumption (12) to demonstrate that this further implies that $\varepsilon^*_{n,L}$ concentrates around 0. For clarity we state auxiliary lemmas inline, and defer proofs to Appendix D.

Finding pairs which are close under the ℓ_2 norm.

Let M_L to be the set of matchings of size L on $x_1 ... x_n$; i.e. each element of M_L is a set of L disjoint (x, x') pairs. Let m^*_L be the 'optimal' matching satisfying:

$$m^*_L \in \arg\min_{z \in M_L} \max_{(x,x') \in z} \|x - x'\|_2. \tag{26}$$

[6]This technique is inspired by the proof of type I error control given for the Conditional Permutation Test in Berrett et al. (2018); see Appendix A.2 of their work for details

That is, m_L^* minimizes the maximum distance between any pair of observations in a mutually disjoint pairing of $2L$ observations. Let

$$d_L^* = \max_{(x,x')\in m_L^*} \|x - x'\|_2. \tag{27}$$

That is, the smallest achievable maximum ℓ_2 distance over all matchings of size L. We'll first show that:

Lemma 4 (Existence of an optimal matching) *If $\mathcal{X} = [0,1]^d$ for some $d \geq 1$,*

$$d_{\frac{n}{4}}^* = O\left(n^{-\frac{1}{d}}\right) \tag{28}$$

with probability 1.

That is, there exists a matching of size at least $\frac{n}{4}$ such the maximum pairwise distance in this matching scales like $O(n^{-\frac{1}{d}})$. Lemma 4 demonstrates the existence of a sizable matching in which the maximum pairwise distance indeed tends to 0.[7] We next demonstrate that this approximates the optimal matching, at the cost of a factor of 2 on L.

Lemma 5 (Greedy approximation to the optimal matching)

$$\max_{l\in[L]} ||x_{2l-1} - x_{2l}||_2 \leq d_{2L}^* \tag{29}$$

That is, the maximum distance between any of the L pairs of observations chosen by our algorithm will be no more than the maximum such distance in the optimal matching of size $2L$.

Corollary 5.1 *For $L \leq \frac{n}{8}$, we have:*

$$\max_{l\in[L]} ||x_{2l-1} - x_{2l}||_2 = O\left(n^{-\frac{1}{d}}\right) \tag{30}$$

This follows immediately by invoking lemma 4 to bound the right hand side of lemma 5. Corollary 5.1 demonstrates that as n grows large, the maximum pairwise ℓ_2 distance between L greedily chosen pairs will go to zero at a rate of $O\left(n^{-\frac{1}{d}}\right)$ provided $L \leq \frac{n}{8}$. We now show that the smoothness condition (12) further implies that, under these same conditions, we recover the asymptotic validity guarantee (13).

From approximately optimal pairings to asymptotic validity.

With the previous lemmas in place, the proof of theorem 2 is straightforward. Plugging the smoothness condition (12) into the definition of the odds ratio (9) yields the following:

For all $(x_{2\ell-1}, y_{2\ell-1}), (x_{2\ell}, y_{2\ell})$,

$$r((x_{2\ell-1}, y_{2\ell-1}), (x_{2\ell}, y_{2\ell})) \in \left[\frac{1}{(1 + C||x_{2\ell-1} - x_{2\ell}||_2)^2}, (1 + C||x_{2\ell-1} - x_{2\ell}||_2)^2\right] \tag{31}$$

Where $C > 0$ is the same constant in the definition of the smoothness condition (12). Corollary 5.1 shows that $||x_{2\ell-1} - x_{2\ell}||_2 = O\left(n^{-\frac{1}{d}}\right)$, so (31) immediately implies that $\varepsilon_{n,L}^*$, defined in (11), also goes to zero at a rate of $O\left(n^{-\frac{1}{d}}\right)$. Thus, if we take L to be a constant and $K \to \infty$, the type I error given in (10) can be rewritten as

$$\begin{aligned}
\mathbb{P}\big(\tau_K \leq \alpha\big) &\leq \alpha + (1 - (1 - \varepsilon_{n,L}^*)^L) + \frac{1}{K+1} && (32)\\
&\leq \alpha + \varepsilon_{n,L}^* L + \frac{1}{K+1} && (33)\\
&= \alpha + O\left(n^{-\frac{1}{d}}\right) && (34)
\end{aligned}$$

[7] In principle, we could find this optimal matching by binary searching for d_L^* using the non-bipartiate maximal matching algorithm of Edmonds (1965); for simplicity, our implementation uses a greedy matching strategy instead.

Where (33) follows from corollary 3.1. If we instead allow L to scale like $o(n^{\frac{1}{d}})$ (still taking $K \to \infty$), (33) implies:

$$\mathbb{P}\big(\tau_K \leq \alpha\big) \leq \alpha + o(1) \tag{35}$$

which concludes the proof of theorem 2.

D Proofs of auxiliary lemmas

Proof of Lemma 3.

Recall that one definition of the total variation distance between two distributions P and Q is to consider the set of *couplings* on these distributions. In particular, the total variation distance can be equivalently defined as:

$$\mathsf{TV}(P, Q) = \inf_{(X,Y)\sim C(P,Q)} \mathbb{P}(X \neq Y) \tag{36}$$

Where $C(\cdot, \cdot)$ is the set of couplings on its arguments. Consider then the following straightforward coupling on $X := (u_1 \dots u_L)$ and $Y := (v_1 \dots v_L)$: draw L random numbers independently and uniformly from the interval $[0, 1]$. Denote these by $c_1 \dots c_L$. Let $u_i = \mathbb{1}[c_i \leq p_i]$, and $v_i = \mathbb{1}[c_i \leq q_i]$. It's clear that X and Y are marginally distributed according to $p_1 \dots p_L$ and $q_1 \dots q_L$, respectively. Furthermore, the probability that $u_i \neq v_i$ is $|p_i - q_i|$ by construction. Thus we have:

$$\mathbb{P}(X \neq Y) = 1 - \mathbb{P}(X = Y) = 1 - \Pi_{i\in[L]}(1 - |p_i - q_i|) \leq 1 - (1 - \max_i |p_i - q_i|)^L \tag{37}$$

This concludes the proof.

Proof of Corollary 3.1.

In the preceding proof of lemma 3, observe that we could have instead written:

$$\mathbb{P}(X \neq Y) = \bigcup_{i\in[L]} \{v_i \neq u_i\} \underbrace{\leq}_{\text{union bound}} \sum_{i\in[L]} |p_i - q_i| \leq L \max_{i\in[L]} |p_i - q_i| \tag{38}$$

Specializing this result to the definitions $\bar{D}_0$ and $\tilde{D}_0$ (and, in particular, the definition of $\varepsilon^*_{n,L}$) completes the proof.

Proof of Lemma 4.

Our proof will proceed via a covering argument. In particular, we cover the feature space $[0, 1]^d$ with a set of non-overlapping d-dimensional hypercubes, each of which has edge length $0 < b < 1$, and show that sufficiently many pairs (x, x') must lie in the same 'small' hypercube. To that end, let $C = \{c_1 \dots c_k\}$ be a set of hypercubes of edge length b with the following properties:

$$\forall c \in C, c \subseteq [-b, 1 + b]^d \tag{39}$$

$$\forall c, c' \in C, c \cap c' = \emptyset \tag{40}$$

$$\forall x \in D_0, \exists c \in C \mid x \in c \tag{41}$$

Where D_0 is the observed data. It's clear that such a covering C must exist, for example by arranging $c_1 \dots c_k$ in a regularly spaced grid which cover $[0, 1]^d$ (though note that per condition (39), some of these 'small' hypercubes may extend outside $[0, 1]^d$ if b does not evenly divide 1). Such a covering may be difficult to index as care must be exercised around the boundaries of each small hypercube; however, as we only require the existence of such a covering, we ignore these details. We now state the following elementary facts:

$$|C| \leq \lfloor \frac{(1 + 2b)^d}{b^d} \rfloor \tag{42}$$

$$\forall c \in C, x, x' \in c, ||x - x'||_2 \leq b\sqrt{d} \tag{43}$$

Where (42) follows because the volume of each $c \in C$ is b^d, and the total volume of all such hypercubes cannot exceed the volume of the containing hypercube $[-b, 1+b]^d$, which gives us an upper bound on the size of the cover C. Furthermore, (43) tells us that for any (x, x') which lie in the same 'small' hypercube c, we have $\|x - x'\|_2 \leq b\sqrt{d}$.

Let $n_c := |\{x_i \mid x_i \in c\}|$ denote the number of observations contained in each small hypercube $c \in C$.

Corollary 5.2 *For any $c \in C$, there exist at least $\lfloor \frac{n_c}{2} \rfloor$ disjoint pairs $(x, x') \in c$ such that $||x - x'||_2 \leq b\sqrt{d}$.*

With these preliminaries in place, we'll proceed to prove lemma 4. To do this, we'll first state one additional auxiliary lemma.

Let $N_{a,b} := \frac{a^d}{b^d} \geq \lfloor \frac{a^d}{b^d} \rfloor$, an upper bound on the number of non-overlapping 'small' hypercubes with edge length b which can fit into $[0, a]^d$. We'll show for any $z > 0$, with $b := \frac{z}{\sqrt{d}}, a := 1 + 2b$, we have:

Lemma 6 (Pairwise distance in terms of packing number)

$$n \geq 2N_{a,b} \Rightarrow \exists \; \frac{n}{4} \textit{ pairs satisfying } ||x - x'||_2 \leq z \tag{44}$$

That is, the pairwise distance between the closest set of $\frac{n}{4}$ pairs (half the observed data in total) can be written in terms of the appropriately parameterized covering number. We defer the proof of this lemma to the following section. For now, we simply plug in the definition of $N_{a,b}$ and rearrange to recover:

$$n \geq 2N_{a,b} = 2\frac{\left(1 + 2\frac{z}{\sqrt{d}}\right)^d}{\left(\frac{z}{\sqrt{d}}\right)^d} \Rightarrow \frac{2^{\frac{1}{d}}\sqrt{d}}{n^{\frac{1}{d}} - 2^{1+\frac{1}{d}}} \leq z \tag{45}$$

Recall that z is the maximum distance between any pairs (x, x') contained in the same small hypercube with edge length $\frac{z}{\sqrt{d}}$. The preceding argument holds for all $z > 0$ which satisfy (45), so in particular, it holds for

$$z^* := \frac{2^{\frac{1}{d}}\sqrt{d}}{n^{\frac{1}{d}} - 2^{1+\frac{1}{d}}}. \tag{46}$$

z^* is the maximum pairwise distance corresponding to one possible matching on $\frac{n}{4}$ (x, x') pairs, so this further implies that there exists a matching M of size $\frac{n}{4}$ such that:

$$\max_{(x,x') \in M} ||x - x'||_2 \leq \frac{2^{\frac{1}{d}}\sqrt{d}}{n^{\frac{1}{d}} - 2^{1+\frac{1}{d}}} = O(n^{-\frac{1}{d}})$$

With probability 1. Thus, it follows that the maximum distance between any pair in the optimal matching $d^*_{\frac{n}{4}}$ also satisfies:

$$d^*_{\frac{n}{4}} = O\left(\frac{2^{\frac{1}{d}}\sqrt{d}}{n^{\frac{1}{d}} - 2^{1+\frac{1}{d}}}\right) = O\left(n^{-\frac{1}{d}}\right)$$

With probability 1, as desired. This establishes the existence of a matching of up to $L = \frac{n}{4}$ disjoint pairs $(x, x') \in [0, 1]^d$ such that the maximum distance between any such pair scales like $O\left(n^{-\frac{1}{d}}\right)$.

We also consider the case where instead of $\mathcal{X} := [0, 1]^d$, we instead have $\mathbb{P}\left(X \in [0, 1]^d\right) \geq 1 - \delta$ for some $\delta \in (0, 1)$. For example, this will capture the case where X is a (appropriately re-centered and re-scaled) multivariate Gaussian. In this case, we provide a corresponding high probability version of lemma 4.

Corollary 6.1 *Suppose instead of $\mathcal{X} := [0,1]^d$, we have for some $\delta \in (0,1)$:*

$$\mathbb{P}(X \in [0,1]^d) \geq 1 - \delta \tag{47}$$

Define $m := (1-\delta)^2 n$

We can then show:

$$\mathbb{P}\left(d^*_{\frac{m}{4}} \leq \frac{2^{\frac{1}{d}}\sqrt{d}}{m^{\frac{1}{d}} - 2^{1+\frac{1}{d}}} \right) \geq 1 - e^{-\frac{\delta^2(1-\delta)n}{2}} \tag{48}$$

That is, we can still achieve a constant factor approximation to the optimal matching in Lemma 4 with probability that exponentially approaches 1.

Proof of Corollary 6.1

Define the set of points which falls in $[0,1]^d$ as follows:

$$S_0 := \{X_i \mid X_i \in [0,1]^d\} \tag{49}$$

and

$$n_0 := |S_0| \tag{50}$$

It is clear that in this setting, the proof of lemma 4 holds if we simply replace n with n_0, the realized number of observations which fall in $[0,1]^d$. However, n_0 is now a random quantity which follows a binomial distribution with mean $(1-\delta)n$ (recall that we assume $(x_i, y_i, \hat{y}_i)$ are drawn i.i.d. throughout). Thus, all that remains is to bound n_0 away from 0, which we can do via a simple Chernoff bound:

$$\mathbb{P}(n_0 \leq (1-\delta)^2 n) \leq e^{-\frac{\delta^2(1-\delta)n}{2}} \tag{51}$$

Thus, it follows that

$$\mathbb{P}(n_0 \geq (1-\delta)^2 n) \geq 1 - e^{-\frac{\delta^2(1-\delta)n}{2}} \tag{52}$$

Thus, we have shown $n_0 \geq m$ with the desired probability. It is clear that we only require a lower bound on n_0 to recover the result of Theorem 4, as additional observations which fall in $[0,1]^d$ can only improve the quality of the optimal matching $d^*_{\frac{m}{4}}$.

Proof of Lemma 5

We will show that the procedure in **ExpertTest** which greedily pairs the closest remaining pair of points L times will always be able to choose at least one of the pairs in an optimal matching of size $2L$. Intuitively, this is because each pair (x, x') chosen by **ExpertTest** can only 'rule out' at most two pairs $(x, x''), (x', x''')$ in any optimal matching of size $2L$. Thus, our greedy algorithm for choosing L pairs can perform no worse than an optimal matching of size $2L$, the sense of minimizing the maximum pairwise distance.

Let m^*_{2L} be an optimal matching of size $2L$ in the sense of (26). Then suppose towards contradiction that:

$$\max_{l \in [L]} ||x_{2l-1} - x_{2l}||_2 > d^*_{2L} \tag{53}$$

Where d^*_{2L} is the smallest achievable maximum distance for any matching of size $2L$ as in (27).

Finally, let $l_m := \arg\min_{l \in [L]} ||x_{2l-1} - x_{2l}||_2 > d^*_{2L}$; i.e. the first pair which is chosen by **ExpertTest** that violates (53). Because pairs are chosen greedily to minimize ℓ_2 distance, and m^*_{2L} is a matching of size $2L$ where all pairs are separated by at most d^*_{2L} under the ℓ_2 norm, it must be that *none* of the pairs which make up m^*_{2L} were available to **ExpertTest** at the l_m-th iteration. In particular, at least one element of every (x, x') pair in m^*_{2L} must have been selected on a previous iteration:

$$\forall (x, x') \in m^*_{2L}, x \in \{x_1 \ldots x_{2l_m - 2}\} \vee x' \in \{x_1 \ldots x_{2l_m - 2}\} \tag{54}$$

As m_{2L}^* contains $2L$ disjoint pairs – $4L$ observations total – this implies that $2l_m - 2 \geq 2L \Rightarrow l_m - 1 \geq L \Rightarrow l_m > L$. This is a contradiction, as **ExpertTest** only chooses L pairs, so l_m only ranges in $[1, L]$. This completes the proof.

Corollary 6.2 *Validity in finite samples*

Theorem 2 implies that we can achieve a bound on the excess type one error in finite samples if we knew the constant C in (12)*. In particular, let*

$$m^* := \max_{\ell \in [L]} ||x_{2\ell-1} - x_{2\ell}||_2 \tag{55}$$

$$\epsilon^* := \max_{r \in [(1+Cm^*)^{-2}, (1+Cm^*)^2]} \left| \frac{1}{r+1} - \frac{1}{2} \right| \tag{56}$$

Then (10) *implies that we can always construct a valid (if underpowered) test at exactly the nominal size α by updating our* `REJECT` *threshold to*

$$\alpha - \left(1 - (1-\epsilon^*)^L\right) - \frac{1}{K+1}$$

Proof of lemma 6

let $C := \{c_1 ... c_k\}$ denote any set of k 'small' nonoverlapping hypercubes of edge length b satisfying properties (39), (40) and (41). As discussed in the proof of lemma 4, each element of C is not guaranteed to lie strictly in $[0,1]$. Rather, each $c \in C$ must merely intersect $[0,1]^d$, implying that each element of the cover is instead contained in the slightly larger hypercube $[-b, 1+b]^d$. As in the proof of lemma 4, we'll again let n_c denote the number of observations x_i which lie in some $c \in C$.

By Corollary 5.2, we have that $\lfloor \frac{n_c}{2} \rfloor$ pairs in each $c \in C$ will satisfy $||x - x'||_2 \leq b\sqrt{d} = z$. Thus what's left to show is that:

$$n \geq 2N_{a,b} \Rightarrow \sum_{j \in [k]} \lfloor \frac{n_{c_j}}{2} \rfloor \geq \frac{n}{4}$$

We can see this via the following argument:

$$\sum_{j \in [k]} \lfloor \frac{n_{c_j}}{2} \rfloor \geq \sum_{j \in [k]} \left(\frac{n_{c_j}}{2} - \frac{1}{2} \right) \tag{57}$$

$$= \frac{n}{2} - \frac{k}{2} \tag{58}$$

$$\geq \frac{n}{2} - \frac{N_{a,b}}{2} \tag{59}$$

$$\geq \frac{n}{2} - \frac{n}{4} = \frac{n}{4} \tag{60}$$

Where (59) follows from (42) and the definition of $N_{a,b}$, and (60) follows because $n \geq 2N_{a,b}$ by assumption. This completes the proof.

E Omitted Details from Section 5

E.1 Identifying relevant patient encounters and classifying outcomes

As described in Section 5, we consider a set of 3617 patients who presented with signs or symptoms of acute gastrointestinal bleeding at the emergency department at a large quaternary academic hospital system from January 2014 to December 2018. These patient encounters were identified using a database mapping with a standardized ontology (SNOMED-CT) and verified by manual physician chart review. Criteria for inclusion were the following: any text that identifies acute gastrointestinal bleeding for hematemesis, melena, hematochezia from either patient report or physical exam findings (which were considered equally valid for the purposes of inclusion). Exclusion criteria were the following: patients with other reasons for overt bleeding symptoms (e.g. epistaxis) or missingness in input variables required to calculate the Glasgow-Blatchford Score.

This identified a set of 3627 patients, of which a further 10 were removed from consideration due to unclear emergency department disposition (neither `Admit` nor `Discharge`). As described in Section 5, we record an adverse outcome ($Y = 1$) for admitted patients who required some form of hemostatic intervention (excluding a diagnostic endoscopy or colonoscopy), or patients who are readmitted or die within 30 days. We record an outcome of 0 for all other patients.

The use of readmission as part of the adverse event definition is subject to two important caveats. First, we are only able to observe patients who are readmitted within the *same* hospital system. Thus, although the hospital system we consider is the dominant regional health care network, it is possible that some patients subseqeuently presented elsewhere with signs or symptoms of AGIB; such patients would be incorrectly classified as not having suffered an adverse outcome. Second, we only record an outcome of 1 for patients who are readmitted with signs or symptoms of AGIB, subject to the same inclusion criteria defined above. Patients who are readmitted for other reasons are not recorded as having suffered an adverse outcome.

E.2 The special case of binary outcomes and predictions

In our experiments we define the loss measure $F(D) := \frac{1}{n}\sum_i \mathbb{1}[y_i \neq \hat{y}_i]$, but it's worth remarking that this is merely one choice within a large class of natural loss functions for which **ExpertTest** produces identical results when $Y, \hat{Y}$ are binary. In particular, observe that a swap of $(y_1, \hat{y}_1), (y_2, \hat{y}_2)$ can only change the value of $F(\cdot)$ if $y_1 \neq y_2$ and $\hat{y}_1 \neq \hat{y}_2$ (we'll assume throughout that all observations contribute equally to the loss; i.e. it is invariant to permutations of the indices $i \in [n]$). This implies that there are only 2^2 out of 2^4 possible configurations of $(y_1, \hat{y}_1, y_2, \hat{y}_2)$ where a swap can change the loss at all. Of these, two configurations create a false negative and a false positive in the synthetic data which did not exist in the observed data:

$$\underbrace{(y_1 = 1, \hat{y}_1 = 1, y_2 = 0, \hat{y}_2 = 0)}_{\text{original data}} \underset{\text{swap}}{\rightarrow} \underbrace{(y_1 = 1, \hat{y}_1 = 0, y_2 = 0, \hat{y}_2 = 1)}_{\text{synthetic data}}$$
$$(y_1 = 0, \hat{y}_1 = 0, y_2 = 1, \hat{y}_2 = 1) \underset{\text{swap}}{\rightarrow} (y_1 = 0, \hat{y}_1 = 1, y_2 = 1, \hat{y}_2 = 0)$$

The other two configurations which change the loss are symmetric, in that a swap *removes* both a false negative and false positive that exists in the observed data:

$$(y_1 = 0, \hat{y}_1 = 1, y_2 = 1, \hat{y}_2 = 0) \underset{\text{swap}}{\rightarrow} (y_1 = 0, \hat{y}_1 = 0, y_2 = 1, \hat{y}_2 = 1)$$
$$(y_1 = 1, \hat{y}_1 = 0, y_2 = 0, \hat{y}_2 = 1) \underset{\text{swap}}{\rightarrow} (y_1 = 1, \hat{y}_1 = 1, y_2 = 0, \hat{y}_2 = 0)$$

Thus, for any natural loss function which is strictly increasing in the number of mistakes $\sum_i \mathbb{1}[y_i \neq \hat{y}_i]$, the first two configurations of $(y_1, \hat{y}_1, y_2, \hat{y}_2)$ will induce swaps which strictly increase the loss, while the latter two will induce swaps that strictly decrease the loss. This means that for a given set of L pairs, we can compute the number of swaps which would increase (respectively, decrease) the loss for *any* function in this class of natural losses. In particular, this class includes loss functions which may assign arbitrarily different costs to false negatives and false positives. Thus, in the particular context of assessing physician triage decisions, our results are robust to variation in the way different physicians, patients or other stakeholders might weigh the relative cost of false negatives (failing to hospitalize patients who should have been admitted) and false positives (hospitalizing patients who could have been discharged to outpatient care).

F Numerical Experiments

We first elaborate here on the example 1 presented in the introduction. Consider the following stylized data generating process:

Example: experts can add value despite poor performance.

Let $X, U, \epsilon_1, \epsilon_2$ be independent random variables distributed as follows:

$$X \sim \mathcal{U}([-2,2]), U \sim \mathcal{U}([-1,1]), \epsilon_1 \sim \mathcal{N}(0,1), \epsilon_2 \sim \mathcal{N}(0,1)$$

Where $\mathcal{U}(\cdot)$ and $\mathcal{N}(\cdot,\cdot)$ are the uniform and normal distribution, respectively. Suppose the true data generating process for the outcome of interest Y is

$$Y = X + U + \epsilon_1$$

Suppose a human expert constructs a prediction $\hat{Y}$ which is intended to forecast Y and can be modeled as:

$$\hat{Y} = \text{sign}(X) + \text{sign}(U) + \epsilon_2$$

Where $\text{sign}(X) \coloneqq \mathbb{1}[X > 0] - \mathbb{1}[X < 0]$.

We compare this human prediction to that of an algorithm $\hat{f}(\cdot)$ which can only observe X, and correctly estimates

$$\hat{f}(X) = \mathbb{E}[Y \mid X] = X$$

As described in the introduction, we use this example to demonstrate that **ExpertTest** can detect that the forecast $\hat{Y}$ is incorporating the unobserved U even though the accuracy of $\hat{Y}$ is substantially worse than that of $\hat{f}(X)$. In particular, we consider the *mean squared error* (MSE) of each of these predictors:

$$\text{Algorithm MSE} \coloneqq \frac{1}{n}\sum_i (Y_i - \hat{f}(X_i))^2$$

$$\text{Human MSE} \coloneqq \frac{1}{n}\sum_i (Y_i - \hat{Y}_i)^2$$

We'll show below that the Algorithm MSE is substantially smaller than the Human MSE. However, we may also wonder whether the performance of the human forecast $\hat{Y}$ is somehow artificially constrained by the the relative scale of $\hat{Y}$ and Y, as the sign$(\cdot)$ operation restricts the range of $\hat{Y}$. For example, a forecaster who always outputs $\hat{Y} = \frac{Y}{100}$ is perfectly correlated with the outcome but will incur very large squared error; this is a special case of the more general setting where human forecasts are directionally correct but poorly *calibrated*. To test this hypothesis, we can run ordinary least squares regression (OLS) of Y on $\hat{Y}$ and compute the squared error of this rescaled prediction. It is well known OLS estimates the optimal linear rescaling with respect to squared error, and we further use the *in sample* MSE of this rescaled prediction to provide a lower bound on the achievable loss. In particular, let:

$$(\beta^*, c^*) \coloneqq \min_{\beta, c \in \mathbb{R}} ||Y - \beta\hat{Y} - c||_2^2 \tag{61}$$

$$\text{Rescaled Human MSE} \coloneqq \frac{1}{n}\sum_i (Y_i - \beta^*\hat{Y}_i - c^*)^2 \tag{62}$$

In Table 3 we report the mean squared error (plus/minus two standard deviations) over 100 draws of $n = 1000$ samples from the data generating process described above. As we can see, both the original and rescaled human forecasts substantially underperform $\hat{f}(\cdot)$.

Table 3: Expert vs Algorithm Performance

Algorithm MSE	Human MSE	Rescaled Human MSE
1.33 ± 0.12	2.67 ± 0.24	1.92 ± 0.16

We now assess the power of **ExpertTest** in this setting by repeatedly simulating $n = 1000$ draws of $(X, U, \epsilon_1, \epsilon_2)$ along with the associated outcomes $Y := X + U + \epsilon_1$ and expert predictions $\hat{Y} := \text{sign}(X) + \text{sign}(U) + \epsilon_2$. We sample 100 datasets in this manner, and run **ExpertTest** on each one with $L, K = 100$, and the distance metric $m(x, x') := \sqrt{(x - x')^2}$. The distribution of p-values $\tau_1 ... \tau_{100}$ is plotted in Figure 1.

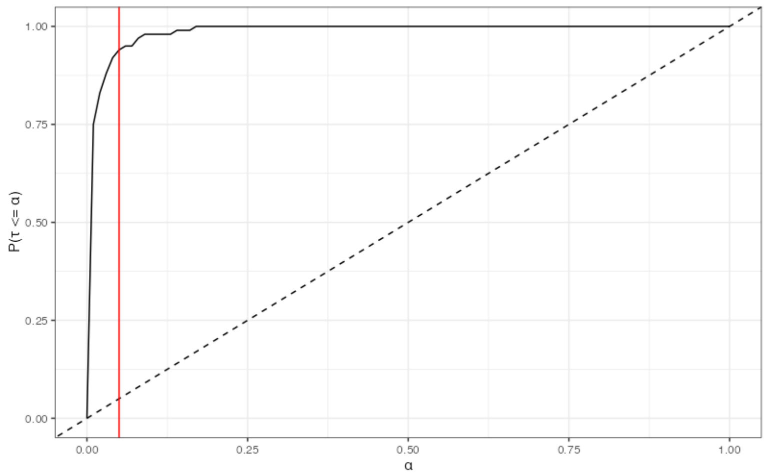

Figure 1: distribution of τ is sharply nonuniform when the expert incorporates unobserved information U in the toy example. The vertical red line indicates a critical threshold of $\alpha = .05$, and the dashed line traces a uniform distribution.

We see that **ExpertTest** produces a highly nonuniform distribution of the p-value τ, and rejects the null hypothesis 94% of the time at a critical value of $\alpha = .05$. To assess whether this power comes at the expense of an inflated type I error, we also run **ExpertTest** with both X and U 'observed'; in particular, suppose the distance measure was instead $m((x, u), (x', u')) = \sqrt{(x - x')^2 + (u - u')^2}$ with everything else defined as above. The distribution of τ in this setting is again plotted in Figure 2.

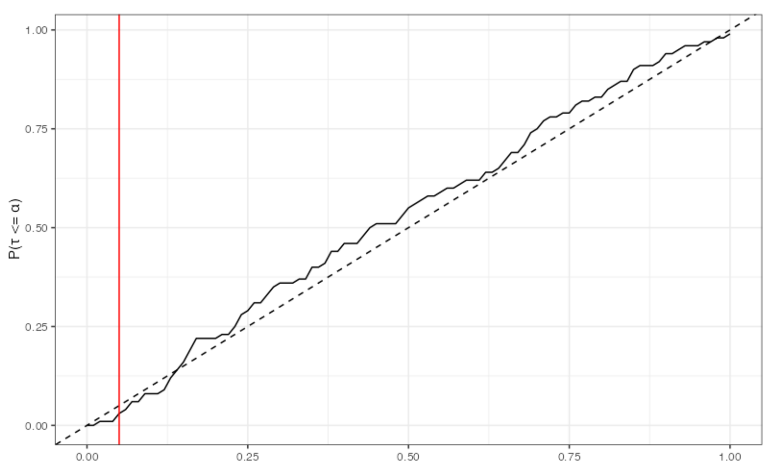

Figure 2: distribution of τ is approximately uniform when the expert does not incorporate unobserved information in the toy example. The vertical red line indicates a critical threshold of $\alpha = .05$, and the dashed line traces a uniform distribution.

When both X and U are observed, and thus the null hypothesis should not be rejected, we instead see that we instead get an approximately uniform distribution of τ with a false discovery rate of only .03 at a critical value of $\alpha = .05$. Thus, the power of **ExpertTest** to detect that the synthetic expert is incorporating some unobserved information U does not come at the expense of inflated type I error, at least in this synthetic example.

Assessing the power of ExpertTest

We now present additional simulations to highlight how the power of **ExpertTest** scales with the number of pairs L and the sample size n in a more general setting. In particular, we consider a simple synthetic dataset $(x_i, y_i, \hat{y}_i), i \in [n] \equiv \{1, \ldots, n\}$ where $x_1 \ldots x_n = [1, 1, 2, 2, \ldots \frac{n}{2}, \frac{n}{2}]'$ and $y_1 \ldots y_n$ is the alternating binary string $[0, 1, 0, 1 \ldots 0, 1]'$ (we consider only even n for simplicity). This guarantees that each of the L pairs chosen are such that $(x_{2\ell-1} = x_{2\ell})$ and $y_{2\ell-1} \neq y_{2\ell}$. Importantly, it's also clear that x is uninformative about the true outcome y – if the expert can perform better than random guessing, it must be by incorporating some unobserved signal U.

We model this unobserved signal by an 'expertise parameter' $\delta \in [0, \frac{1}{2}]$. In particular, for each pair $(y_{2\ell-1}, y_{2\ell})$ for $\ell \in [1 \ldots \frac{n}{2}]$, we sample $(\hat{y}_{2\ell-1}, \hat{y}_{2\ell})$ such that $(\hat{y}_{2\ell-1}, \hat{y}_{2\ell}) = (y_{2\ell-1}, y_{2\ell})$ with probability $\frac{1}{2} + \delta$ and $(y_{2\ell}, y_{2\ell-1})$ otherwise. Intuitively, δ governs the degree to which the expert predictions $\hat{Y}$ incorporate unobserved information – at $\delta = 0$, we model an expert who is randomly guessing, whereas at $\delta = \frac{1}{2}$ the expert predicts the outcome with perfect accuracy.

First, we consider the case of $n \in \{200, 600, 1200\}$ and fix L at $\frac{n}{8}$ as suggested by the proof of Theorem 2. For each of these cases, we examine how the discovery rate scales with the expertise parameter $\delta \in \{0, .05 \ldots .45, .50\}$. In particular, we choose a critical threshold of $\alpha = .05$ and compute how frequently **ExpertTest** rejects H_0 over 100 independent draws of the data for each value of δ. These results are plotted below in Figure 3.

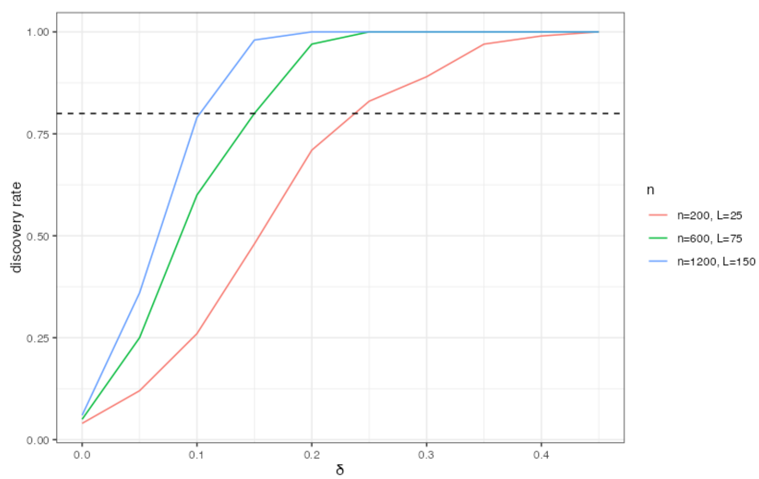

Figure 3: The power of **ExpertTest** as a function of sample size n and expertise parameter δ. The horizontal dashed line corresponds to a power of 80%

Unsurprisingly, the power of **ExpertTest** depends critically on the sample size – at $n = 1200$, **ExpertTest** achieves 80% power in rejecting H_0 when the expert only performs modestly better than random guessing ($\delta \approx .1$). In contrast, at $n = 200$, **ExpertTest** fails to achieve 80% power until $\delta \approx .25$ – corresponding to an expert who provides the correct predictions over 75% of the time even when the observed x is completely uninformative about the true outcome.

Next we examine how the power of **ExpertTest** scales with L. We now fix $n = 600$ and let $\delta = .2$ to model an expert who performs substantially better than random guessing, but is still far from providing perfect accuracy. We then vary $L \in \{20, 40 \ldots 200\}$ and plot the discovery rate (again at a critical value of $\alpha = .05$, over 500 independent draws of the data) for each choice of L. These results are presented below in Figure 4.

As expected, we see that power is monotonically increasing in L, and asymptotically approaching 1. With $\delta = .2$, we see that **ExpertTest** achieves power in the neighborhood of only 50% with $L = 20$ pairs, but sharply improves to approximately 80% power once L increases to 40. Beyond this threshold we see that there are quickly diminishing returns to increasing L.

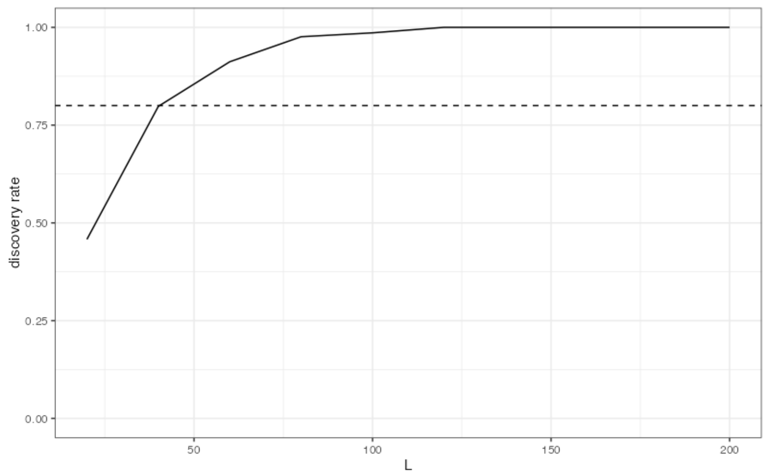

Figure 4: The power of **ExpertTest** as a function of L, with $n = 600, \delta = .2$. The horizontal dashed line corresponds to a power of 80%

Excess type I error of ExpertTest

Recall that, per Theorem 1, **ExpertTest** becomes more likely to incorrectly reject H_0 as L increases relative to n. In particular, larger values of L will force **ExpertTest** to choose (x, x') pairs which are farther apart under any distance metric $m(\cdot, \cdot)$, and thus induce larger values of $\varepsilon^*_{n,L}$ as defined in (11). Furthermore, even for fixed $\varepsilon^*_{n,L} > 0$, the type one error bound given in Theorem 1 degrades with L. We empirically investigate this phenomenon via the following numerical simulation.

First, let $X = (X_1, X_2, X_3) \subset \mathbb{R}^3$ be uniformly distributed over $[0, 10]^3$. Let $Y = X_1 + X_2 + X_3 + \epsilon_1$ and $\hat{Y} = X_1 + X_2 + X_3 + \epsilon_2$, where ϵ_1, ϵ_2 are independent standard normal random variables. In this setting, it's clear that $H_0 : Y \perp\!\!\!\perp \hat{Y} \mid X$ holds.

We repeatedly sample $n = 500$ independent observations from this distribution over $(X, Y, \hat{Y})$ and run **ExpertTest** for each $L \in \{25, 50 \ldots 250\}$. We let $K = 50$ and $m(x, x') := ||x - x'||_2^2$ be the ℓ_2 distance. We let the loss function $F(\cdot)$ be the mean squared error of $\hat{Y}$ with respect to Y. For each scenario we again choose a critical threshold of $\alpha = .05$, and report how frequently **ExpertTest** incorrectly rejects the null hypothesis over 50 independent simulations in Figure 5.

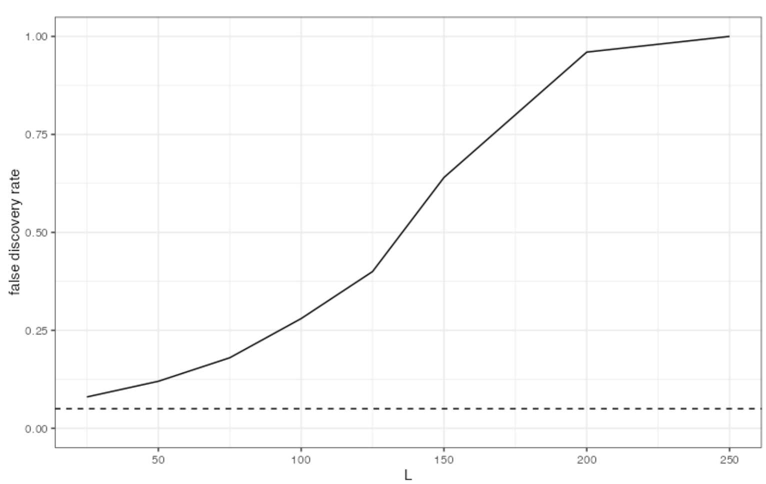

Figure 5: The type I error rate of **ExpertTest** as a function of L, with $n = 500$ and a critical threshold of $.05$. The horizontal dashed line corresponds to the nominal false discovery rate of $.05$

As we can see, the type I error increases sharply as a function of L, and **ExpertTest** incurs a false discovery rate of 100% at the largest possible value of $L = \frac{n}{2}$! This suggests that significant care should be exercised when choosing the value of L, particularly in small samples, and responsible use of **ExpertTest** will involve leveraging domain expertise to assess whether the pairs chosen are indeed 'similar' enough to provide type I error control.

G Pseudocode for ExpertTest

In this section we provide pseudocode for **ExpertTest**. Inputs $D_0, L, K, \alpha, F(\cdot), m(\cdot,\cdot)$ are as defined in Section 3.

ExpertTest$(D_0, L, K, \alpha, F(\cdot), m(\cdot,\cdot))$

$X_0 \leftarrow \{x \mid (x,\cdot,\cdot) \in D_0\}$ ▷ initialize set of remaining observations
$P \leftarrow \emptyset$ ▷ initialize set of paired predictions

for $\ell = 1 : L$ **do**
  $(x_{2\ell-1}, x_{2\ell}) \leftarrow \underset{(x,x')}{\operatorname{argmin}}\, m(x,x')$ ▷ find closest remaining pair, breaking ties arbitrarily
  $X_\ell \leftarrow X_{\ell-1} \setminus \{x_{2\ell-1}, x_{2\ell}\}$
  $P \leftarrow P \cup \{(\hat{y}_{2\ell-1}, \hat{y}_{2\ell})\}$ ▷ save predictions associated with closest remaining pair
end for

$f_0 \leftarrow F(D_0)$ ▷ calculate observed loss

for $k = 1 : K$ **do**
  $D_k \leftarrow \mathrm{swap}(D_0, P, \frac{1}{2})$ ▷ independently swap each $(\hat{y}_{2\ell-1}, \hat{y}_{2\ell}) \in P$ with equal probability
  $f_k \leftarrow F(D_k)$ ▷ calculate synthetic loss
end for

$\tau \leftarrow \frac{1}{K}\sum_k \mathbb{1}[f_k \lesssim f_0]$ ▷ calculate quantile of observed loss, breaking ties at random

if $\tau \leq \alpha$ **then** ▷ if $\tau \leq \alpha$, H_0 is rejected with p-value $\alpha + \frac{1}{K+1}$
  `REJECT`
end if
