Submitted to the Annals of Applied Statistics

HOLDOUT SETS FOR SAFE PREDICTIVE MODEL UPDATING

BY SAMI HAIDAR-WEHBE[1] SAMUEL R EMERSON [1] LOUIS JM ASLETT [1,a] AND JAMES LILEY [1,b]

[1]*Department of Mathematical Sciences, Durham University,* [a]*louis.aslett@durham.ac.uk;* [b]*james.liley@durham.ac.uk*

Predictive risk scores for adverse outcomes are increasingly crucial in guiding health interventions. Such scores may need to be periodically updated due to change in the distributions they model. However, directly updating risk scores used to guide intervention can lead to biased risk estimates. To address this, we propose updating using a 'holdout set' — a subset of the population that does not receive interventions guided by the risk score. Balancing the holdout set size is essential to ensure good performance of the updated risk score whilst minimising the number of held out samples. We prove that this approach reduces adverse outcome frequency to an asymptotically optimal level and argue that often there is no competitive alternative. We describe conditions under which an optimal holdout size (OHS) can be readily identified, and introduce parametric and semi-parametric algorithms for OHS estimation. We apply our methods to the ASPRE risk score for pre-eclampsia to recommend a plan for updating it in the presence of change in the underlying data distribution. We show that, in order to minimise the number of pre-eclampsia cases over time, this is best achieved using a holdout set of around 10,000 individuals.

1. Introduction. Risk scores estimate the probability of an event Y given predictors X. Their use has become routine in medical practice (Topol, 2019), where Y is typically a binary random variable representing an adverse event incidence and X various clinical observations. Once calculated, risk scores may be used to guide interventions, perhaps modifying X, with the aim of decreasing the probability of an adverse event. The ASPRE score (Akolekar et al., 2013), with which we will be working, evaluates risk of pre-eclampsia (PRE), a hypertensive complication of pregnancy, using predictors derived from ultrasound scans in early pregnancy, and can be used to prioritise prescription of aspirin (and other interventions) to at-risk pregnancies.

Risk scores are typically developed by regressing observations of Y on X. Should the conditional distribution $(Y|X)$ subsequently change, or 'drift', then risk estimates may become biased (Tsymbal, 2004). This can happen naturally over time, meaning that risk scores typically need to be updated periodically to maintain accuracy. In the complex settings in which such scores are typically used, interventions in response to risk scores may be multifaceted, and hence impractical to record.

Updating of the risk score will involve obtaining new observations of (X, Y), but crucially the distribution of (X, Y) may also have changed due to the effect of the risk score itself: that is, high predicted risk of an adverse event may trigger intervention to reduce that risk. The effect of such interventions may be impractical to infer or measure, and indeed the fact that intervention took place may be unrecorded. In the example above, this means individuals prescribed aspirin in response to higher ASPRE scores should have lower PRE risk than they would have if ASPRE was not used. Should a new risk score be fitted to observed (X, Y), the effect of hypertension on risk would be underestimated. This bias is worsened by heavier intervention resulting in risk scores becoming 'victims of their own success' (Lenert, Matheny and Walsh, 2019). This framework of directly updating a risk score on an 'intervened'

Keywords and phrases: Model updating, Performative prediction, Updating paradox.

population has been termed 'repeated risk minimisation' (Perdomo et al., 2020) in the context when such bias is accounted for, or 'naïve updating' (Liley et al., 2021) when it is not.

In Liley et al. (2021) we briefly noted that this bias could be avoided by splitting the population on which the score can be used into an 'intervention' set and a 'holdout' set, with an updated model trained on the latter. In this work, we formally develop this proposal for practical use in real-world predictive risk score updating, prove its suitability, and apply it to the ASPRE score. In particular, we address a vital tension in the choice of an optimal holdout size (OHS) for the holdout set: for the risk score to be accurate, the holdout set should not be too small; but any samples in the holdout set will not benefit from risk scores, so nor should it be too large. The holdout set must be actively generated; it is not sufficient to simply update a model using samples who received a risk score but were untreated.

We begin by introducing a motivating example in Section 1.1, reviewing relevant literature in Section 1.2. Our first question is whether a hold-out set is worth the cost, as opposed to simply continuing to use the existing score with degraded performance, or updating naïvely. We set out the problem and notation precisely in Section 2. In Section 3, we then develop theory which proves that under certain simplifying assumptions, as long as drift and intervention effects occur, the cost of holding out samples is generally justified and that the holdout set approach outperforms common alternatives. We then turn attention to the problem of selecting holdout set size in Section 4, constructing an optimisation problem to find an OHS. Therein we also set out why several apparent alternatives are not competitive. In Section 5, we describe two algorithms for OHS estimation, using a parametric model, and using Bayesian emulation. In Section 6, we support our findings with numerical demonstrations and resolve our motivating example by applying our methods to a risk score for pre-eclampsia (PRE) to estimate an OHS for updating it.

1.1. *Motivating example.* We consider the ASPRE score (Akolekar et al., 2013) for evaluating risk of pre-eclampsia (PRE), a hypertensive complication of pregnancy, on the basis of predictors derived from ultrasound scans in early pregnancy (we will not differentiate early- and late-stage PRE). Although treatable, PRE confers a serious risk to both the fetus and the mother. The risk of PRE is lowered by treatment with aspirin through the second and third trimesters (Rolnik et al., 2017a), but aspirin therapy itself confers a slight risk, contraindicating universal treatment, and suggesting prescription of aspirin only if the risk of PRE is sufficiently high or other indications are present (ACOG, 2016). The ASPRE score was developed to aid clinicians in estimating PRE risk and has been shown to be useful in prioritising patients for aspirin therapy (Rolnik et al., 2017b). Our aim is to develop a plan for updating the ASPRE predictive risk score, in such a way as to minimise the expected number of PRE cases per unit time. To update the ASPRE model, we presume covariate and outcome data is available on a set of pregnancies in a given time period.

Due to changing population demographics, we anticipate that the influence of risk factors is likely to change over time, and the ASPRE score will predict individual PRE risk less accurately over time (as compared to an optimal predictor), reducing the usefulness of the score in identifying high-risk pregnancies and the capacity of healthcare practitioners to anticipate and avoid PRE cases.

We illustrate why updating the ASPRE model is difficult using a somewhat exaggerated effect. Let us informally denote PRE incidence as Y, and ASPRE score covariates as X, and an indicator A for whether a patient is treated with aspirin. Suppose that we consider patients with particular set of covariates $X = x$, and that under normal healthcare (in the absence of an ASPRE score) such pregnancies have a PRE risk $P(Y|X = x) = 60\%$ (where some individuals are treated pre-emptively with aspirin). Suppose that if we were to alter clinical care to treat all (or almost all) of these patients with aspirin, the PRE risk would drop to 2%

(approximately the baseline rate). We write this as a *counterfactual* $P_{A\leftarrow 1}(Y|X=x) = 2\%$: that is, we force treatment $A = 1$.

Should we 'naively' re-fit the ASPRE score, we would learn that individuals with covariates x had a risk of approximately 2%. Such a risk score woud not generally warrant healthcare practitioners to focus additional attention on such patients, and may even serve as a false reassurance. Patients with covariates x now face a risk of (say) 70% of PRE, which is worse than having no risk score at all.

We propose avoiding this problem by maintaining a holdout set. For patients in this set, no ASPRE score would be calculated at first scan, and treatment would be according to best practice in the absence of a risk score. An updated ASPRE score can then be fitted to data from these patients. Patients in this holdout set go without the benefit of the ASPRE score, leading to a less accurate allocation of prophylactic treatment (aspirin) and consequently a higher risk of PRE (Rolnik et al., 2017b). However, an inappropriately small holdout set would lead to an inaccurate updated model, reducing the benefit of future use of the score. We make a simplifying assumption that ASPRE is only used for decisions on aspirin therapy, although in practice it could be used more generally.

1.2. *Review of related work.* In applied statistics, a vast amount of effort has been expended designing predictive scores in healthcare. Widespread collection of electronic health records has spurred development of new diagnostic and prognostic risk scores (Cook and Collins, 2015; Liley et al., 2024), which can allow detection of patterns too complex for humans to discover. Examples of such scores in widespread use include: EuroSCORE II, which predicts mortality risk at hospital discharge following cardiac surgery (Nashef et al., 2012); and the STS risk score from the United States, predicting risk of postoperative mortality (Shahian et al., 2018). Many such scores have demonstrable efficacy in clinical trials and in-vivo (Chalmers et al., 2013; Wallace et al., 2014; Hippisley-Cox, Coupland and Brindle, 2017).

An important general concern with these scores is continued accuracy of predictions. A 2011 review found that risk scores for hospital readmission perform poorly and highlighted issues with design of their trials (Kansagara et al., 2011). More recently, an analysis of a sepsis response score used during the COVID pandemic found increasing risk overestimation over time (Finlayson et al., 2020). Various efforts have been made to standardise procedures in risk score estimation to address these issues (Collins et al., 2015). A critical practical aspect of development of such scores is to determine how they must be updated: indeed, we claim that whenever a risk score is deployed for use, a plan should be made for its continued development, and it is to this general applied area that our work contributes.

Several algorithms have been developed to update models with new data in the presence of drift (Lu et al., 2018), which ideally leads to the best possible model performance after every update. However, adaptation of model updating to avoid naïve updating-induced bias requires explicit causal reasoning (Sperrin et al., 2019) and often further data collection (Liley, 2021). In a seminal paper, Perdomo et al. (2020) analyse asymptotic behaviour of repeated naïve updating, giving necessary and sufficient conditions under which successive predictions end up converging to a stable setting where they essentially predict their own effect. Other approaches to optimise a general loss function by modulating parameters of the risk score are developed in Mendler-Dünner et al. (2020); Drusvyatskiy and Xiao (2020); Li and Wai (2021) and Izzo, Zou and Ying (2021). These approaches seek to minimise a 'performative' loss to the population in the presence of an arbitrary risk score, whilst our approach seeks to target risk scores which reliably estimate the same quantity, namely $P(Y \mid X)$ in a 'native' system prior to risk score deployment. Our approach is well-suited to settings where the performative loss is essentially intractable, requiring cost estimates of risk scores only in limited settings.

We note that 'stability' is not necessarily desirable in terms of the distribution of interventions: in the QRISK3 setting, if an individual is at untreated risk of 50% and treated risk of 10%, with treatment distributed proportionally to assessed risk, a 'stable' risk score would assess risk as e.g. 30%, prompting a milder intervention than actual untreated risk would suggest, after which true risk remains at 30%, regardless of treatment cost.

We found no applied or theoretical literature directly addressing the focus in this paper: determining how large a holdout set should be. Similar problems do arise in clinical trial design: Stallard et al. (2017) estimate the optimal size of clinical trial groups for a rare disease in which individuals not in the trial stand to gain more than those in it, using a Bayesian decision-theoretic approach accounting for benefit to future patients in the population. Our problem is related to computation of a minimal training size for a clinical prediction model Riley et al. (2020), but rather than being limited by financial cost of obtaining training samples, we are limited by a cost to all individuals in the holdout set, allowing specification of an explicit tradeoff for larger sample sizes.

In previous work (Liley et al., 2021) we proposed, in addition to a holdout set, that the problems of updating a risk score in the presence of interventions could broadly be managed by complete causal modelling of the intervention or by explicitly specifying what interventions should be made. In applied work (Liley et al., 2024) we proposed a practical alternative in which we updated a risk score as a maximum of the existing risk score and a refitted score. We are concerned in this work with the setting in which none of these options are usable, which we consider to be a common setting. We detail in Supplement S3.2 why causal modelling is of limited use when we cannot record an intervention or when we deterministically plan an intervention, and in Supplement S3.3 why our approach of using the maximum of two scores is not generalisable.

OHS estimation requires quantification of expected material costs when using risk scores trained to holdout data sets of various sizes: that is, the cost of reduced accuracy from limiting the OHS, as well as the cost due to individuals in the holdout set not benefiting from a risk score. Such costs depend on the error in risk predictions. The relation of predictive error to training set size is well studied and known as the 'learning curve', which can sometimes be accurately parameterised (Amari, 1993). A recent review paper suggests a power-law is accurate for simple models (Viering and Loog, 2021).

1.3. *Legal and ethical considerations.* Use of a holdout set appears ethically tenuous. However, we argue that in many circumstances it is unethical *not* to use a holdout set, in that other options lead to worse outcomes. Essentially, use of a holdout set limits costs incurred from inaccuracies in prediction to individuals in that set, whereas all alternatives lead to risk score inaccuracy across the entire population. We formalise these arguments in Section 3, in particular showing that the updating paradox described above is inevitable for risk scores on complex systems intended to guide interventions.

Contributions in this area are important due to rapidly evolving legislation. Currently, the European Union treat each update of a risk model as a separate risk score requiring re-approval, but in the United States a proactive approach is taken with a 'total-life cycle' paradigm which allows practitioners to update risk models as necessary without requesting approval (USFDA et al., 2019). This approach could allow updating-induced biases to go undetected, and highlights the need for safe updating methods in risk score deployment. The use of holdout sets as examined in this work offers one potential solution.

2. Problem description.

2.1. *General setup.* We presume a random process X_t, $t \in \mathbb{R}^+$, representing the covariates of a single sample at time t. We let X_t have distribution μ_t where μ_t has constant support $\mathscr{X}$.

To demonstrate why we opt to use a holdout set approach, we define two functions dependent on t. We define $f_t(x)$ as the probability of an event if no risk score is in place, and $g_t(x;\rho)$ as the probability of an event when a risk score ρ is used to guide decisions. We allow ρ to be an arbitrary risk score (that is, $\rho \in R = \{r : \mathscr{X} \to [0,1]\}$ and $g : (\mathscr{X} \times R) \to [0,1]$) but will generally take it to be a risk score fitted to samples encountered prior to or at time t. We will not explicitly consider the occurence of adverse events as random variables (for the moment); rather we will simply consider f_t and g_t as functions (noting that $g_t(x,\rho)$ depends on the *function* ρ rather than only the value $\rho(x)$). We will generally omit the argument ρ from g_t, as it will usually be clear. We will use f, g to informally denote the sets of functions $\{f_t, t \in \mathbb{R}^+\}$, $\{g_t, t \in \mathbb{R}^+\}$ respectively.

We wish to estimate f_t rather than g_t, since it gives a risk of the event in question under standard practice: that is, *without* already using a risk score. An agent may opt to intervene in addition to standard practice if $f_t(x)$ is high.

We thus aim to generate risk scores $\rho_\cdot : \mathscr{X} \to [0,1]$ which estimate f_t for t on some interval. Risk scores will be fitted to samples observed over a time period $(e-s, e]$ for $s \leq e$. Denote by

$$\rho_h^{n,e,s}(\cdot) \tag{1}$$

a risk score fitted to n samples of (X,Y) with $t \sim U(e-s,e)$, $X \sim \mu_t$, and $Y \sim \text{Bern}(h(X))$ where U is a uniform distribution, $\text{Bern}(\cdot)$ a Bernoulli distribution, and h is either $f_t(\cdot)$ or $g_t(\cdot;\rho)$ for some ρ. We will measure deviation of a risk score ρ from a function f using (essentially) mean-square generalisation error:

$$\xi_t^2(\rho, f) = \mathbb{E}_{X_t \sim \mu_t}\left\{\left(\rho(X_t) - f(X_t)\right)^2\right\}. \tag{2}$$

We will consider four options for how best to decide on a series of risk scores to be used over a time period $[0,T]$. We will use the term 'epoch' to mean periods of time $(0,\delta]$, $(\delta, 2\delta]$, …, $(e, e+\delta]$ with $e \in \mathbb{N}\delta$ during which a particular risk score is used and a new risk score is (potentially) fitted. From Section 4 onwards, we will generally take $\delta = 1$ which we will be able to do without loss of generality.

Holdout: Amongst samples encountered while $t \in (e-s, e]$, $e \in \mathbb{N}\delta$, we withhold $n_\star$ randomly chosen samples from attaining risk scores, and fit a risk score to them. We thus obtain a risk score $\rho_f^{n_\star,e,s}$ for use on the epoch $(e, e+\delta]$. We call this approach the *holdout* strategy.

No-update: We use a risk score $\rho_0 = \rho_{f_0}^{n,0,0}$, fitted to a number n of samples at time 0, and continue using this score throughout the period. We call this the *no-update* strategy.

Naïve update: Whenever $t \in (e-s, e]$, $e \in \mathbb{N}\delta$, we fit a risk score to as many samples as possible (say n') on whom a risk score is already used. We thus attain a risk score $\rho_{g_t}^{n',e,s}$ for use on the epoch $(e, e+\delta]$. We call this approach the *naïve update* strategy.

We also consider the performance of an unspecified alternative which is 'less than asymptotically perfect', in that for some period of time the risk score is (on average) a slightly biased estimator of f_t.

Alternative: We consider an arbitrary alternative giving rise to a risk score ρ_t at time t for which for some value T:

$$\lim_{N_\rho \to \infty}\left(\mathbb{E}_{t \sim U[0,T], D}\left\{\xi_t^2(\rho_t, f_t)\right\}\right) = b > 0\,, \tag{3}$$

where D denotes any information used to fit risk scores ρ_t, N_ρ is the number of samples used to fit ρ, and b does not depend on any parameters of the strategy (e.g., update frequency δ), though may depend on T[1].

We will finally consider an 'oracle' option. Rather than perfect information (e.g. perfect knowledge of f_t for all t), our oracle can observe f_t acting on a number of samples:

Oracle: We presume that whenever $t \in [e-s, e)$, $e \in \mathbb{N}\delta$, we fit a risk score to n samples of (X, Y), for which $X \sim \mu_t$, $Y \sim \text{Bern}(f_t(X))$, and use the risk score $\rho_{f_t}^{n,e,s}$ on the epoch $(e, e+\delta]$. We call this approach the *oracle* strategy.

Our strategy for 'holdout' updating is illustrated with $\delta = 1$ as a causal graph in Figure 1. The ellipses $\{X_0\}$, $\{Y_0\}$ correspond to sets of observations of (X, Y) with $X \sim \mu_0$ and $(Y|X) \sim \text{Bern}(f_0(x))$ in epoch 0, representing initial training data, to which a risk score ρ_0 is fitted, where $\rho_0 \approx f_0$.

We use the shorthand $\{X_e^i\}$ ('i' for intervention) to mean a set of samples from X_t with $t \in (e, e+1]$, representing the set of samples on which the risk score is used, and $\{X_e^h\}$ ('h' for holdout) to mean sets of samples from X_t with t as close to $e+1$ as possible, representing the set of 'holdout' samples on which the next risk score ρ_e is fitted. We define corresponding sets $\{Y_e^i\}$ and $\{Y_e^h\}$ representing observations: for $x_t \in \{X_e^i\}$, we have the corresponding y_t in Y_e^i distributed as $y_t \sim \text{Bern}(g_t(x_t, \rho_{\lfloor t \rfloor}))$ and for $x_t \in \{X_e^h\}$, we have the corresponding y_t in Y_e^h distributed as $y_t \sim \text{Bern}(f_t(x_t))$.

Under a 'native' setting prior to deployment of a risk score, $\{X_0\}$ and $\{Y_0\}$ have a single causal link, modelled by risk score ρ_0 (leftmost epoch of Figure 1). Once ρ_0 is in use in the intervention set in epoch 1 (ellipses $\{X_1^i\}$, $\{Y_1^i\}$), a second causal pathway through ρ_0 is established from $\{X_1^i\}$ to $\{Y_1^i\}$, but there remains only one causal pathway from $\{X_1^h\}$ to $\{Y_1^h\}$ in the holdout set (middle epoch of Figure 1). The quantities in the shaded area do not causally depend on any quantities outside the shaded area; we will consider them together in Section 4. The updating process can be continued rightwards (ρ_1, ρ_2, ...).

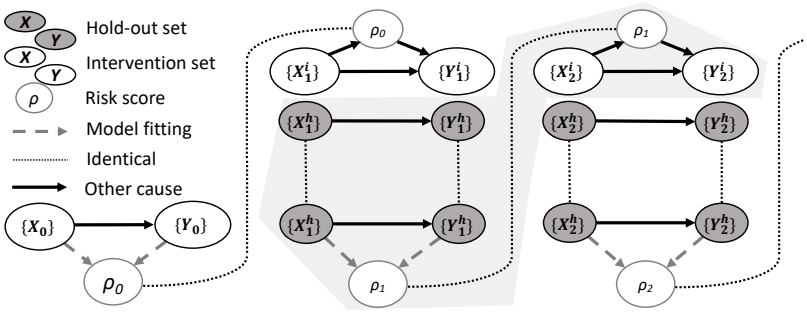


Fig 1: Dynamics of a risk model, ρ, across three epochs under holdout set updating. Each column corresponds to an epoch, denoted by subscripts $0, 1, 2$, representing consecutive intervals of continuous time $(0, 1]$, $(1, 2]$, and $(2, 3]$ respectively. Ellipses containing X or Y correspond to covariates and outcomes respectively, with superscripts i and h denoting the partition into intervention and holdout sets respectively.

[1] We will make assumptions under which the no-update strategy is of the alternative-strategy type. We differentiate them here to illustrate examples of such alternative strategies.

3. Motivation for use of a holdout set. We argue that use of a holdout set approach is generally necessary for tolerable total population costs when updating a risk score. In this sense, we argue that it is an ethical imperative to consider use of a holdout set in this context. We will now make and justify a series of assumptions (which will only be used directly in this section). We will discuss robustness to these assumptions in Section 3.1.

The motivation for updating a risk score is generally that the true risk f_t changes continuously with time t in a random way. If f_t did not change with t there would be no need to update a well-fitted risk score. Lipschitz-continuous change in the distribution of covariates and the risk function underlies the assumption that, for a risk score to be useful (regardless of updating strategy), it must remain reasonably accurate for a period of time after it is fitted.

ASSUMPTION A1 (Non-negligible drift). *For some $T > 0$ we have $\int_0^T \xi_t^2(f_t, f_0)dt > 0$*

ASSUMPTION A2 (Continuous drift in population distribution). *μ_t is uniformly Lipschitz continuous in t with respect to total variation distance.*

ASSUMPTION A3 (Continuous drift in target function). *For all x, $f_t(x)$ is α_1-Lipschitz continuous in t (where α_1 does not depend on x)*

We implicitly assume that drift in f_t is unpredictable: if f_t changed in a *predictable* way, we could simply update our estimate of f_t with time.

We presume that adding more samples to a training set for a risk score improves its error at an $O(n^{-1})$ rate (typical of, for instance, linear models). In subsequent sections, we will consider instead an arbitrary 'learning curve'.

ASSUMPTION A4 (Expected prediction error). *For $h \in \{f, g\}$, denoting*

$$H(x) = \mathbb{E}_{t\sim U(e-s,e)}\{h_t(x)\}$$

we have

$$\mathbb{E}\left\{\xi_u^2(\rho_h^{n,e,s}, H)\right\} = O(n^{-1})$$

for any time u, where the expectation is over the data used to fit $\rho_h^{n,e,s}$.

Since we assume prediction of an adverse event, we want to lower the chance of this event as much as possible. We define a cost per sample in terms of the reduction in the probability of an adverse event which can be achieved by intervention: that is, how much lower is g_t (probability of an adverse event when using a risk score to guide decisions) than f_t (probability of adverse event). For the moment, we are unconcerned with the scale of this cost, so we assume a unit coefficient for this difference. We shift this cost so that it is zero when we are doing 'as well as possible'; that is, when ρ is exactly f.

ASSUMPTION A5 (Costs). *For a sample encountered at time t, when a risk score ρ is in use, we define the expected total cost, $c_t(\rho)$, as:*

$$c_t(\rho) = \mathbb{E}_{X\sim\mu_t}\left\{g_t(X, \rho) - g_t(X, f_t)\right\};$$

that is, the change in expected difference in the probability of the outcome when using an estimate ρ_t of f_t and with perfect knowledge of f_t. We make the assumption that:

$$k_1 = \mathbb{E}_{X\sim\mu_t}\left\{f_t(X) - g_t(X, f_t)\right\} > 0\,. \tag{4}$$

The value of k_1 can be considered as the cost per sample when no risk score is used (since the risk of outcome is $f_t(X)$ in this case), so the assertion that $k_1 > 0$ is tantamount to a 'potentially useful' risk score: with perfect knowledge of f_t, it is possible to reduce the expected risk of the outcome below f_t itself.

We now make a key assumption in presuming that the risk score is 'useful', in that our propensity to make g_t smaller than f_t depends directly on the accuracy of the score. Taken together, A5 and A6 constrain the relationship between f_t, g_t, and ρ, helping to simplify the statement of our initial results, but we will show that we can substantially weaken this assumption in Section 3.1, and discuss our reasons for making this type of assumption in Section 5.1.

ASSUMPTION A6 (Usefulness of risk score). *If a risk score ρ_t is in use at time t, we have*

$$c_t(\rho_t) = k_2\, \xi_t^2(\rho_t, f_t)$$

for some constant $k_2 > 0$.

We note that it immediately follows that cost is minimised at 0 if and only if $\rho_t = f_t$ holds μ_t-almost everywhere, and that we cannot find a 'better' risk score than f_t itself (our cost may be considered an 'excess cost' over a setting with essentially perfect knowledge of f_t), and we allow that a sufficiently inaccurate risk score may incur a higher cost than no risk score, which could potentially occur in real settings. We give an example of a class of functions g_t satisfying Assumptions A5 and A6 in Supplement S2.4. Finally, we make simplifying assumptions of independence of samples at different times and a constant population size.

ASSUMPTION A7 (Population size and holdout number). *The times at which samples are observed occur uniformly randomly over time (that is, from a Poisson process) with a mean of N samples per unit of time. When using the holdout strategy, we use a constant number $n_\star$ of samples in each epoch (if there are fewer than $n_\star$ samples available in $(e-s, e]$, we use no risk score for the subsequent epoch $(e, e+\delta]$, but this will be rare). Samples of X_t need not be independent, although in many cases they will be nearly so: for the ASPRE score, for instance, treatment decisions will generally be made only once per pregnancy.*

We define the 'cost of a strategy' $C_{\text{strat}}[t_1, t_2]$ as the cost accrued for all samples encountered over time interval $[t_1, t_2]$ using a given strategy 'strat' (which may be: (h): holdout; (0): no-update; (n): naïve update; (a): alternative or (o): oracle). Given Assumptions A7 and A6, we have

$$\frac{1}{N(t_2 - t_1)} \mathbb{E}\{C_{\text{strat}}[t_1, t_2]\} = \mathbb{E}_{t \sim U[t_1, t_2], \rho_t}\{c_t(\rho_t)\} = \mathbb{E}_{t \sim U[t_1, t_2], \rho_t}\{k_2\, \xi_t^2(\rho_t, f_t)\}\,,$$

where the (potentially random) function ρ_t is determined by the strategy. We now state three results, the first of which establishes the growth rate of the holdout strategy over time.

THEOREM 3.1. *Suppose we use a holdout set with size $n_\star = \Theta(N^a)$, with $0 < a < 1$, an $s < 1$ of size $s = \Theta(N^{a+\epsilon-1})$ for some ϵ with $0 < \epsilon < 1 - a$, and an update frequency $\delta \leq 1$ which may vary with N. Under Assumptions A3, A4, A6 and A7, we have*

$$\mathbb{E}\left\{C_{(h)}[0, T]\right\} = \delta N T k_2 (2\alpha_1 + \alpha_1^2) + \delta^{-1} O(N^a) + O(N^{a+\epsilon}) + O(N^{1-a})\,.$$

For fixed δ, this suggests an optimal holdout set size $n_\star = \Theta(N^{1/2})$. If we choose $\delta = O(N^{-b})$ with $0 < b < 1 - (a + \epsilon)$ (given that we must have $s < \delta < 1$) then

$$\mathbb{E}\left\{C_{(h)}[0,T]\right\} = O(N^{a+b}) + O(N^{a+\epsilon}) + O(N^{1-a}) + O(N^{1-b})\,, \tag{5}$$

and we achieve an optimal sublinear asymptotic growth rate of $O(N^{\frac{2}{3}})$ if $a = \frac{1}{3}$, $b = \frac{1}{3}$.

We now consider the costs of the no-update and naïve-update strategies. We will show that for either strategy, costs must grow at least linearly in population size (thereby also showing that the updating paradox arises inevitably from Assumptions A5, A6):

THEOREM 3.2. *Suppose we choose s such that $s \to 0$ as $N \to \infty$. Under Assumptions A1–A7, for sufficiently small δ, we have for $(strat) \in \{(0), (n), (a)\}$:*

$$\mathbb{E}\left\{C_{(strat)}[0,T]\right\} = \Omega(N)\,. \tag{6}$$

We note that Theorem 3.2 immediately implies a dominance of the holdout-set strategy, since we cannot attain sublinear cost growth with the no-update, naïve-update, or alternative strategies.

Finally, we show that for fixed δ, the holdout strategy is essentially optimal in that its cost is asymptotically similar to that of the oracle strategy. This is not true for the no-update, alternative or naïve-update strategies, for which total cost arbitrarily exceeds that of the oracle.

THEOREM 3.3. *Consider use of each strategy in parallel with an 'oracle' procedure. Under Assumptions A1–A7, with sufficiently small fixed δ, holdout set size $n_\star = \Theta(N^{1/2})$, and $s < 1$ of size $s = \Theta(N^{\epsilon - 1/2})$ for some ϵ with $0 < \epsilon < 1/2$, we have for $(strat) \in \{(0), (n), (a)\}$:*

$$\lim_{N\to\infty}\left(\frac{\mathbb{E}\left\{C_{(h)}[0,T]\right\}}{\mathbb{E}\left\{C_{(o)}[0,T]\right\}}\right) = 1, \text{ and } \lim_{N\to\infty}\left(\frac{\mathbb{E}\left\{C_{(strat)}[0,T]\right\}}{\mathbb{E}\left\{C_{(o)}[0,T]\right\}}\right) > 1 \quad \left(= \Omega(\delta^{-2})\right)\,.$$

We prove Theorems 3.1, 3.2 and 3.3 in Supplement S2. Restrictions on δ in Theorems 3.2, 3.3 are required because for large δ (relative to Lipschitz constants in Assumptions A2 and A1) too much drift occurs between updates to guarantee sustained performance for the time a risk score is in use.

In general, our results show that any strategy for which alternative strategy condition (3) holds leads to higher costs than the holdout set approach. We claim that it is essentially impossible for a strategy to evade condition (3) (that is, be able to arbitrarily closely approximate f_t at almost all $t \in [0,T]$) without the use of a holdout set, and hence have costs competitive to the holdout-set strategy.

In our setting, without a holdout set, the only information available when we wish to update the model is a set of samples of X_t and $\text{Bern}\{g_t(X_t, \rho_t)\}$ for some ρ_t, where $g_t(\cdot, \rho_t) \neq f_t$ and $\rho_t \neq f_t$ (if $f_t = g_t$, the risk score is prompting no change in behaviour, which generally contradicts its usefulness). With only this information, we cannot hope to infer f_t without error. To see this, suppose we have some risk score ρ_t in place over an epoch and observe the function $g_t(\cdot, \rho_t)$. All we know about f_t comes from Assumptions A5 and A6, that is:

$$k_2\xi_t^2(f_t, \rho_t) = k_1 - \mathbb{E}_{X\sim\mu_t}\{f_t(X) - g_t(X, \rho_t)\}\,, \tag{7}$$

from which f_t is not identifiable (see Supplement S3.1 for details and an explicit example).

We may alternatively try to estimate f_t using our knowledge of f_u with $u < t$ (we will say $u = 0$). If $f_0 = f_t$, there is no drift, and the risk score need not be updated. To accurately infer f_t from f_0, we would need to know exactly how f_t changes with t, whereas typically drift in f_t is random.

It is more conceivable that f_t could be inferred from g_t using additional information (for instance, records of interventions). However, risk scores are usually used in complex settings (e.g. medicine, finance or law), and decisions are nuanced, so the extent to which a decision is due to a risk score is hard to quantify, even if decisions are explicitly recorded. In the most general setting where we have no additional information, we claim a holdout set provides the most principled statistical approach.

In the context of the ASPRE score, Assumption A1 corresponds to the claim that over time, the risk of PRE given ultrasound-derived predictors changes in a non-negligible way and Assumptions A2 and A3 assert that this change happens gradually. Assumption A5 and A6 state that our cost (number of PRE cases per unit time) should worsen according to the accuracy of our prediction of PRE risk. Assumption A7 states that we encounter pregnancies at approximately an even rate over time.

Presuming that we may not update the score arbitrarily frequently, Theorem 3.1 indicates we should choose a holdout set size scaling roughly as $\sqrt{N}$ to optimise costs, and Theorems 3.2 and 3.3 show that our costs, should we do so, will asymptotically be no worse than if we incurred no costs in the holdout set, but would be substantially worse should we update 'naïvely' without one, or not update the model at all.

3.1. *Robustness.* We briefly discuss the robustness of the findings in Theorems 3.1, 3.2 and 3.3 to Assumptions A1–A7. We first note that we can weaken Assumption A6 to

$$k_2^l \xi^{b_l} \leq c_t \leq k_2^u \xi^{b_u}, \tag{8}$$

where $\xi = \xi_t^2(\rho_t, f_t)$, for some constants $k_2^l, k_2^u > 0$ and $0 < b_u \leq 1 \leq b_l$, whilst retaining the correctness of Theorem 3.2, requiring only minor modifications to Theorems 3.1, 3.3, and Equation 5 (Supplement S2.4). Although we assume that k_1 and k_2 (or k_2^l, k_2^u) are constant over time, we may relax this to their being bounded-below over time by positive constants.

In the absence of drift (assumption A1), the no-update strategy is preferable to the holdout-set update, though they remain equivalent as $N \to \infty$ in the sense of Theorem 3.3. However, the naive-update strategy remains suboptimal in this case (Supplement S2.5).

We do not explicitly mention the possibility of 'latent' covariates which influence the risk of outcome f_t. Rather, we may simply assume that risk scores are considered as marginals over latent covariates, which we discuss in greater detail as part of Supplement S3.2. Additionally, we do not assume that drift is independent of our actions: that is, the choice of who to intervene on may affect μ_t in the future. However, we do not explicitly model any long-term effects of intervention. We presume that the intervention is not able to be recorded; we briefly discuss possibilities if the intervention is recorded (including use of a treatment indicator as a covariate in the predictive model) in Supplement S3.2.

If f_t or μ_t are not Lipschitz continuous for some t (Assumptions A3, A2 respectively), then a risk score approximating $f_{t-\epsilon}$ will not necessarily approximate $f_{t+\epsilon}$, even for small ϵ. Practically, this means that performance of a risk score fitted prior to t cannot be guaranteed after t. In practice, this could readily occur (for instance, a financial risk score fitted prior to an unexpected disaster), but such an event would nonetheless be better-managed using the holdout-set strategy than others, since a holdout set could be used to 'reset' the risk score after t.

Heuristically, when using a naïve update strategy, a less-biased risk score (that is, for which ρ_t is more similar to f_t) will initially result in lower costs during the first deployment epoch (by a mechanism analogous to Assumption A6), but then suffer higher costs after updating (since lower bias induces a larger difference between f_t and g_t by Assumption A5, worsening the consequences of using g_t to approximate f_t). Under Assumption A1, a non-updated model will accrue greater costs over time. We make the case that holdout sets enable good estimation of f_t for most of the sample population at all t.

These heuristics generally remain true when using a metric of similarity between ρ_t and f_t other than $\xi_t^2(f_t, \rho_t)$, or when Assumptions A4, A6, A5 and A7 only roughly hold. In such cases, the holdout set approach will still tend to be lower-cost than other approaches for sufficiently large N, although the precise statements of the Theorems may not hold. For example, if Assumption A4 fails perhaps due to model mis-specification so that accuracy decreases to a positive minimum, then costs will grow linearly with the holdout set strategy. However, it will be closer to the oracle strategy than no-update, naïve-update or alternative strategies. Although we specify a constant size (up to order) for the holdout set in Theorems 3.1 and 3.3, in practice lower costs may be attainable with variable holdout set sizes.

Nonetheless, we do acknowledge that our formulation is quite prescriptive. We roughly model a medical setting in which a physician sees a patient, assesses them, then intervenes and sends them away, only observing some outcome later. There are of course related settings for which this formulation is inappropriate, and we once again must defer to heuristic arguments in such situations.

3.2. *Note on ethics.* In medical settings (such as our motivating example) the use of holdout sets appears ethically tenuous, due to differential treatment of samples in and out of the holdout set. We argue, however, that the use of holdout sets in these settings should still be considered. Our main reason is the absence of a viable alternative: as above, in many settings it is not possible to attain costs comparable to an idealised oracle strategy without a holdout set. We recapitulate that, even in the event that risk-score guided interventions are recorded, it is *not* sufficient to use a 'natural' holdout set by simply considering individuals who received a risk score, but were untreated, and it is not necessarily possible to infer f_t (Supplement S3.2). The ethical questions surrounding holdout sets are discussed in more depth in Chislett et al. (2024).

We do note an important subtle assumption is that k_1 is finite (Assumption A5). Settings in which it is unacceptable for even one sample to not have a risk score correspond to an infinite k_1, and in such settings we must make do with a risk score for which performance is not close to optimal. By contrast, there are settings for which the absence of a risk score is less serious: for instance, if the outcome in question is not life-threatening and for which existing best practice is often adequate for identifying cases (e.g., tooth decay (Zukanović, 2013) or minor sexually transmitted infections (Kranzer et al., 2021)). In such settings the use of a holdout set is more ethically acceptable, since the cost to any given individual (even in the holdout set) is low.

Lastly, we underscore the importance of considering the adoption of a holdout set, when other ethical concerns do not take precedence. In situations where a holdout set is not feasible, and withholding any treatment is deemed unacceptable, the ability to effectively respond to a drifting ground truth is compromised. Consequently, updates would ultimately yield suboptimal risk scores for the entire population.

4. Choosing the size of a holdout set. For the remainder of this paper, we will be concerned with choosing an optimal size for the holdout set. We begin by somewhat simplifying our formulation, with a focus on total cost, and no longer using Assumptions A1–A7. We will retain k_1 and k_2 as having roughly their existing meanings.

Returning to Figure 1, our goal at time e is to nominate a size n of the holdout set $\big(\{X_e^h\}, \{Y_e^h\}\big)$, comprising samples encountered during $[e, e+1)$, to be used to fit a risk score to be used during $[e+1, e+2)$. During $[e+1, e+2)$, the risk score is used only on the intervention set $\big(\{X_{e+1}^i\}, \{Y_{e+1}^i\}\big)$, whose size will depend on the choice of holdout set size at the next update time $e+1$. To estimate the costs incurred on this intervention set, we take its size to be $N - n$: the same as the size of the intervention set in $[e, e+1)$, with N samples

encountered in total in $[e, e+1)$. We do not, however, intend that the same holdout size be used for successive epochs in general.

We denote the n samples in the 'holdout' set $\left(\{X_e^h\}, \{Y_e^h\}\right)$ as D_n (where 'D' indicates 'data'). Since we choose $\left(\{X_e^h\}, \{Y_e^h\}\right)$ to be as close in time to $\left(\{X_{e+1}^i\}, \{Y_{e+1}^i\}\right)$ as possible, we will presume that

$$X_e^h \sim \mu_{e+1}\,, \qquad \mathbb{E}(Y_e^h | X_e^h) = f_{e+1}(X_e^h)\,,$$
$$X_{e+1}^i \sim \mu_{e+1}\,, \qquad \mathbb{E}(Y_{e+1}^i | X_{e+1}^i) = g_{e+1}(X_{e+1}^i, \rho_e)\,.$$

A risk score ρ_e is fitted to D_n which approximates $f_{e+1}(x)$ and is used in the intervention set $\left(\{X_{e+1}^i\}, \{Y_{e+1}^i\}\right)$. We will presume that samples in D_n are pairwise independent, as are samples in $\left(\{X_{e+1}^i\}, \{Y_{e+1}^i\}\right)$, although samples in the latter depend on D_n through the fitted risk score.

We define $C_1(X)$ and $C_2(X; D_n)$ as random variables associated with the total 'cost' of an observation with covariates X in the holdout set in epoch e and intervention set in epoch $e+1$ respectively. Although the most natural cost (as in Assumption A5) may be the number of adverse events, we take the 'cost' in this case to represent only a quantity we aim to minimise through our use of the risk score. The cost could, for example, encompass the costs of managing adverse events and the costs of administering interventions.

The value of $C_2(X; D_n)$ depends on D_n only through the risk score fitted to D_n. We define the expected cost per observation in the holdout and intervention sets, respectively:

$$k_1 = \mathbb{E}_{X \sim \mu_{e+1}, C_1}\left\{C_1(X)\right\}\,, \qquad k_2(n) = \mathbb{E}_{X \sim \mu_{e+1}, C_2}\left[\mathbb{E}_{D_n}\left\{C_2(X; D_n)\right\}\right]\,. \tag{9}$$

Subscripted values C_1 and C_2 indicate variance in $C_1(X), C_2(X; D_n)$ independent of X, D_n. We now express total expected cost ℓ across all samples as a function of holdout set size n:

$$\ell(n) = \underbrace{k_1 n}_{\text{Holdout set}} + \underbrace{k_2(n)(N-n)}_{\text{Intervention set}}\,. \tag{10}$$

As for C_1, C_2, the meaning of ℓ is contextual dependent on the application; for instance, in QRISK3 it may mean total number of deaths for a fixed healthcare budget.

4.1. *Sufficient conditions for existence of an OHS.* In this Section, we consider conditions under which the cost $\ell(n)$ can be readily optimised. We discuss estimation of N, k_1 and $k_2(\cdot)$ in Section 5. We begin with the following assumptions:

ASSUMPTION B1. *k_1 does not depend on n: in a medical context, this means for example that treatment plans and outcomes for patients without risk scores do not depend on the number of such patients.*

ASSUMPTION B2. *$k_2(n)$ is monotonically decreasing in n: the more data available to train the risk score, the greater its clinical utility.*

ASSUMPTION B3. *There exists $M \in (0, N)$ such that $n \geq M \Leftrightarrow k_2(n) \leq k_1$: a good enough risk score will lead to better patient outcomes than baseline treatment, and a poor enough risk score fitted to small amounts of data leads to worse expected outcomes than baseline treatment.*

ASSUMPTION B4. *$\mathbb{E}\left\{k_2(i+1) - k_2(i)\right\} > \mathbb{E}\left\{k_2(j+1) - k_2(j)\right\}$ for $1 \leq i < j \leq N - 1$, with expectations over training data: the 'learning curve' for our risk score is convex; there are diminishing returns in the cost per patient from adding more samples to the training data.*

We may extend the domain of $k_2(\cdot)$, $\ell(\cdot)$ to the real interval $[0, N)$ such that both functions are smooth; and $k_2'(n) < 0$, given Assumption B2, $k_2''(n) > 0$, given Assumption B4. This leads to the following result, that there exists an optimal size for the holdout set minimizing the expected total cost. The proof is given in Supplement S4.

THEOREM 4.1. *Suppose Assumptions B1–B4 hold. Then there exists an OHS* $N_\star \in \{1, \ldots, N-1\}$ *with* $N \in \mathbb{N}$, *such that:* $\ell(i) \geq \ell(j)$ *for* $0 < i < j < N_\star$ *and* $\ell(i) \leq \ell(j)$ *for* $N_\star < i < j < N$

In the context of the ASPRE score, Assumption B1 indicates that if an ASPRE score is unavailable for a given patient, their care is independent of the number of people used in training the ASPRE score. Assumptions B2 and B4 state that when training the score, the usefulness of the score (that is, the frequency of PRE in patients on whom we use the score) improves at a diminishing rate as we increase the number of training samples. Assumption B3 states that a sufficiently good risk score is better for patients than no risk score at all. Theorem 4.1 then states that, to minimise overall cost of PRE cases, a holdout set of some non-zero size should be used to update the score.

We note that the OHS always exceeds the minimal training sample size required to match baseline treatment.

COROLLARY 1. *The value of* $N_\star$ *always exceeds the value of* M *in Assumption B3, since if* $N' < M$ *we have* $\ell(N') = k_1 N' + k_2(N')(N - N') > k_1 N' + k_1(N - N') = k_1 N \geq \ell(N_\star)$

Consequently Assumption B4 may be relaxed for $i, j < M$; we need only be concerned with the behaviour of $k_2(n)$ at realistically large values of n, rather than $n \in \{1, \ldots, M\}$. We also have

$$\lim_{N\to\infty} \lim_{n\to N} \ell'(n) = \lim_{N\to\infty} \lim_{n\to N} \left\{k_1 - k_2(n) + (N-n)k_2'(n)\right\} = k_1 - \lim_{n\to\infty} k_2(n)\,,$$

and since $k_1 > k_2(n) > 0$ for large n, we have that expected total costs $\ell(n)$ are increasing, but bounded by the per observation expected cost of baseline treatment k_1.

Let us suppose that the holdout set used to fit ρ_e is encountered at time $e+1$, and that we have Assumption A6 and a stronger form of Assumption A4 in that $\xi^2_{e+1}(\rho_e, f_{e+1}) = c_2 n^{-1} = O(n^{-1})$ for some constant c_2. The cost for a single sample with distribution μ_{e+1} (as in the intervention set), by Assumption A6 is given by $c_{e+1} = k_2(n) = k_2 \xi^2_{e+1}(\rho_e, f_{e+1}) = k_2 c_2 n^{-1}$ and $\ell(n) = k_1 n + k_2 c_2 n^{-1}(N-n)$. This is minimised at $n = \sqrt{k_2 c_2 N / k_1} = \Theta(N^{1/2})$, in line with the fixed-δ theoretical OHS in Section 3.

4.2. *Robustness to Assumptions B1–B4.* The applicability of Assumptions B1–B4 in real world settings requires careful consideration. We address violations of Assumptions B2 and B4 in Section 5.3 and Assumptions B1 and B3 here.

Assumption B1 is fundamental to the success of the holdout set concept. It may be violated if, for instance, agents who can make interventions learn the behaviour of a risk score and apply this to samples with no score. However, such violations are not of serious concern: if we presume that such changes in agents endure over time, then they can be considered as simply contributing to drift, which need not be independent of holdout set size.

If ethically appropriate, Assumption B1 could be assured by partitioning agents to manage only samples in holdout sets or only in intervention sets (e.g., cluster randomisation). This requires assuming that changes in agent behaviour as above *do not* endure until the following epoch.

If Assumption B3 fails because $k_2(0) \leq k_1$, we may show a weaker result (the presence of a non-trivial but potentially non-unique minimum loss) by replacing Assumptions B2, B3 and B4 with:

ASSUMPTION B5. *There exists an* $0 < M < N$ *such that* $\frac{N-M}{N}(k_1 - k_2(M)) > k_1 - k_2(0)$

This assumption is essentially stating that at some point the risk score will greatly outperform a risk score built with no data. This leads to the result that:

THEOREM 4.2. *Suppose Assumptions B1 and B5 hold, and* $k_2(0) \leq k_1$. *Then there exists an* $N_\star \in \{1, \ldots, N-1\}$ *such that:* $\ell(i) \geq \ell(N_\star)$ *for* $i \in \{1, \ldots, N-1\}$ *and* $\ell(i) > \ell(N_\star)$ *for* $i \in \{0, N\}$

In a setting in which $k_1 > k_2(0)$ but one or more of Assumptions B2–B4 do not hold, we have

THEOREM 4.3. *Suppose Assumption B1 holds,* $k_1 < k_2(0)$ *and there exists* $0 < M < N$ *such that* $k_2(M) < k_1$. *Then there exists an* $N_\star \in \{1, \ldots, N-1\}$ *such that:* $\ell(i) \geq \ell(N_\star)$ *for* $i \in \{1, \ldots, N-1\}$ *and* $\ell(i) > \ell(N_\star)$ *for* $i \in \{0, N\}$.

Both results are proved in Supplement S4. In the setting where $k_1 = k_2(0)$ and Assumption B1 does not hold, it may sometimes be reasonable to assign samples in the holdout set risk scores based on no data (for example risk scores generated entirely from expert opinion) and blind agents to holdout/intervention status. Under this setting we may have greater assurance of Assumption B1.

5. Estimation of OHS.

5.1. *Estimation of* $k_2(n)$. We are aiming to find a holdout set size which minimises costs during an epoch $e \geq 1$, $t \in [e, e+1)$ (noting that the holdout set will be used late in the epoch when $t \approx e+1$). This choice must be made during epoch $e-1$ (when $t < e$). We take it that we have the following:

1. An approximate number of samples on which the model will be used or refitted;
2. A cohort of samples (X, Y) with $X \sim \mu_{e-\epsilon} \approx \mu_e$, $Y|X \sim f_{e-\epsilon}(X) \approx f_e(X)$, with ϵ small.

In 2, the samples are from a holdout set if $e > 1$, or from initial training data if $e - 1 = 0$. We aim to estimate the cost function $\ell(n)$ at $t \in [e, e+1)$. Our approach is to estimate $\ell(n)$ for $t = e - \epsilon$ and assume that the OHS is approximately conserved from $t = e - \epsilon$ to $t = e + 1$, though in reality drift may occur in $\ell(n)$.

At time $t = e - \epsilon$, we need to estimate constants N, k_1, and the function $k_2(\cdot)$. The constants N and k_1 are straightforward: N, the total number of samples on which a predictive score can be fitted or used, will usually be known or specified (item 1 above); and k_1, the average cost per sample under baseline behaviour without a score, can be estimated from observed costs in the cohort in item 2 above. The function $k_2(\cdot)$ (Equation 9) is more difficult to estimate, as it involves quantifying costs of hypothetical risk scores. We may tractably estimate $k_2(\cdot)$ by assuming that

$$\mathbb{E}_{X \sim \mu_e, C_2}\{C_2(X; D_n)\} = \mathscr{L}\{\text{err}(\rho_{D_n})\}, \tag{11}$$

where ρ_{D_n} is a risk score fitted to D_n, $\text{err}(\cdot)$ is a measure of error, and $\mathscr{L}$ is some nondecreasing function. We claim that in general circumstances we may take $\mathscr{L}(\cdot)$ to be linear and $\text{err}(\cdot)$ to be expected mean-squared error (MSE) or a similar general loss. We derive this heuristically in Supplement S5 and derive expressions for $k_2(n)$ directly in a specific case in Section 7. Once $\mathscr{L}$ is known, this allows $k_2(n)$ to be estimated readily by establishing the 'learning curve' of the risk score using item 2 above.

Some direct estimates of $k_2(n)$ are necessary to determine $\mathscr{L}$. One option is to designate subcohorts of the intervention set in epoch $e-1$ to receive risk scores fitted to smaller subsamples of available training data, allowing direct observation of the costs of such risk scores. While simple, this approach may be ethically tenuous and expensive. Other options include estimating the function $\mathscr{L}$ through expert opinion or other outside information.

In summary, we recommend that $k_2(n)$ is estimated by jointly making a small number of estimates during epoch $e-1$, either directly or indirectly, to establish $\mathscr{L}$, and thereafter estimated by evaluating the error of a risk score fitted to n samples using the set in item 2 and transforming it according to the estimated $\mathscr{L}$.

5.2. *Parametric estimation of OHS.* A natural algorithm for estimating the OHS is immediately suggested by Theorem 4.1: assume k_2 is known up to parameters θ, and estimate N, k_1 and θ to estimate the OHS. Parameters θ of k_2 may be estimated from observations of pairs $\{n, k_2(n)\}$, potentially with error in $k_2(n)$. To minimize the number of estimates of $k_2(n)$ we iteratively add observations $(n, k_2(n))$ to an existing set of observations so as to greedily reduce expected error in the resultant OHS estimate.

We suggest a routine parametric algorithm (Algorithm 1) with estimation of asymptotic confidence intervals. Full details of theory, proofs and algorithms are given in Supplement S6.

Algorithm 1 Parametric OHS estimation overview

1: $\mathbf{n}, \mathbf{k_2}, \boldsymbol{\sigma}^2 \leftarrow$ some initial values $\mathbf{n} = \{n_1, \ldots, n_m\}$ with $(\boldsymbol{k_2})_i \approx k_2(n_i)$, $(\boldsymbol{\sigma}^2)_i = \text{var}(\hat{k}_2(n_i))$
2: **while** $|\mathbf{n}| <$ total iterations **do**
3: Find $\tilde{n}$ which minimises expected OHS confidence interval width (§S6, eq.48), and add to $\mathbf{n}$
4: Estimate $\hat{k}_2(\tilde{n}) \approx k_2(\tilde{n})$
5: $\mathbf{n} \leftarrow \{\mathbf{n} \cup \tilde{n}\}$, $\mathbf{k_2} \leftarrow \left\{\mathbf{k_2} \cup \hat{k}_2(\tilde{n})\right\}$, $\boldsymbol{\sigma}^2 \leftarrow \left\{\boldsymbol{\sigma}^2 \cup \text{var}\left\{\hat{k}_2(\tilde{n})\right\}\right\}$
6: **end while**
7: **return** Re-estimate OHS $n_\star^{\text{final}}$ from $\mathbf{n}, \mathbf{k_2}, \boldsymbol{\sigma}$

We use the shorthands $\Theta = (N, k_1, \theta)$ and $\Theta_0 = \mathbb{E}(\Theta)$. Consistency of Algorithm 1 depends on whether $\mathbf{n}$ eventually contains enough elements of sufficient multiplicity to estimate Θ_0 consistently. Sampling some positive proportion of values of $\mathbf{n}$ randomly from $\{1, \ldots, N\}$ guarantees that the multiplicity of all $n \in \mathbf{n}$ almost surely eventually exceeds any finite value, readily ensuring consistency. Finite-sample bias of $n_\star^{\text{final}}$ depends on $\nabla_\Theta n_\star$ and the variance of Θ. See Supplementary Figure S2 for typical forms of $\nabla_\Theta n_\star$.

5.3. *Semi-parametric (emulation) estimation of OHS.* Parametrization of $k_2(n)$ may be inappropriate if the learning curve of the risk score or the relation between the learning curve and $k_2(n)$ (from Section 5.1) are complex (Viering and Loog, 2021). We propose a second algorithm which is less reliant on assuming a parametric form for $k_2(n)$, using Bayesian optimisation (Brochu, Cora and De Freitas, 2010). We quantify the uncertainty in $k_2(n)$ through the construction of a Gaussian process emulator of ℓ . The prior mean function for this emulator takes a particular parametric form, but crucially can deviate from this prior function with the addition of data.

We take the n corresponding to the minimum cost function value (for the evaluated points) to be our OHS estimate. Values of n at which to estimate $\ell(n)$ are selected using an 'expected improvement' function $EI(\cdot)$, whereby if $EI(n) > \tau$ we roughly expect the minimum cost to decrease by at least τ from adding another estimate of $\ell(n)$ to our data. This also provides a natural stopping criterion. An outline procedure is given in Algorithm 2. Further algorithm details and proofs of consistency are in Supplement S7.

Algorithm 2 Emulation OHS estimation; minimum cost improvement τ

1: $\mathbf{n}, \mathbf{d} \leftarrow$ some initial values $\mathbf{n} = \{n_1, \ldots, n_m\}$ with $d_i = d(n_i) \approx \ell(n_i)$
2: Estimate mean and variance of Gaussian process $\ell(n)$ and function $EI(n)$
3: **while** $\max_{n\in\{1,\ldots,N\}}\{EI(n)\} > \tau$ **do**
4: $\quad \tilde{n} \leftarrow \arg\max_{n\in\{1,\ldots,N\}} EI(n)$
5: $\quad$ Estimate $d(\tilde{n}) \approx k_2(\tilde{n})$
6: $\quad \mathbf{n} \leftarrow (\mathbf{n} \cup \tilde{n})$; $\mathbf{d} \leftarrow (\mathbf{d} \cup d(\tilde{n}))$
7: $\quad$ Re-estimate mean and variance of Gaussian process $\ell(n)$ and function $EI(n)$
8: **end while**
9: **return** $n_\star^{\text{final}} = \arg\min_{n\in\{1,\ldots,N\}\cap\mathbf{n}} \left\{ \frac{1}{|\{j:n_j=n\}|} \sum_{j:n_j=n} d_j \right\}$

6. Simulations.

6.1. *Simulation showing dominance of holdout set approach.* We briefly illustrate the theory described in Section 2.1 using simulated data, similar to our motivating example, which satisfiess Assumptions A1–A5, A7 and the weaker form of Assumption A6 (Equation 8) with $\delta = 10$, $s = 1$, no drift in μ_t, $\alpha_1 \approx 0.32$, $k_1 = 0.023$, $k_2^l = 0.038$, $b_u = b_l = 1$ and $k_2^u = 0.22$ (details in Supplement S8.1). Figure 2 shows total costs accrued per sample during unit time periods of no-update ('none'), naïve-update ('naive') and holdout-update ('H.S') at two holdout sizes over time. As drift occurs in f_t, the costs associated with the no-update strategy grow due to increasingly poor approximation of f_t, and the costs of the naïve-update strategy increase dramatically due to intervention effects. The total costs of the holdout-set approaches remain low. Choice of the holdout set size aims to balance increased costs due to non-intervention in the holdout set (the 'spikes') against inaccuracy in fitted scores after drift. We demonstrate the natural emergence of an OHS in a simulated context in Supplement S8.2.

6.2. *Comparison of parametric and emulation algorithms.* In this section, we give circumstances in which one of Algorithms 1 or 2 may be preferable to the other. We consider two versions of the function $k_2(n)$:

$$k_2^p(n) = an^{-b} + c \quad , \quad k_2^{np}(n) = an^{-b} + c + \frac{10^4}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{n - 4\times 10^4}{8\times 10^3} \right)^2 \right);$$

where: 'p'/'np' denote 'parametric assumptions satisfied/not satisfied', and $\theta = (a, b, c) = (10000, 1.2, 0.2)$. We assume N and k_1 are known to be 10^5 and 0.4 respectively. For emulation, we use a kernel width $\zeta = 5000$ and variance σ_u^2 of 10^7.

The function $k_2^{np}(n)$ exhibits 'double-descent' behaviour (Supplementary Figures S5a, S5b), which is possible for learning curves (Viering and Loog, 2021) but violates Assumptions B2, B4.

We firstly show the distribution of estimates of OHS using both algorithms when k_2 takes either form above. To fit k_2, we use 200 randomly chosen values from $\{1, \ldots, N\}$ for $\mathbf{n}$, with values $\mathbf{k_2}$ independently sampled as $(\mathbf{k_2})_i \sim N\{k_2(n_i), \sigma_i^2\}$, where $\sigma_i \overset{\text{iid}}{\sim} U(0.001, 0.02)$.

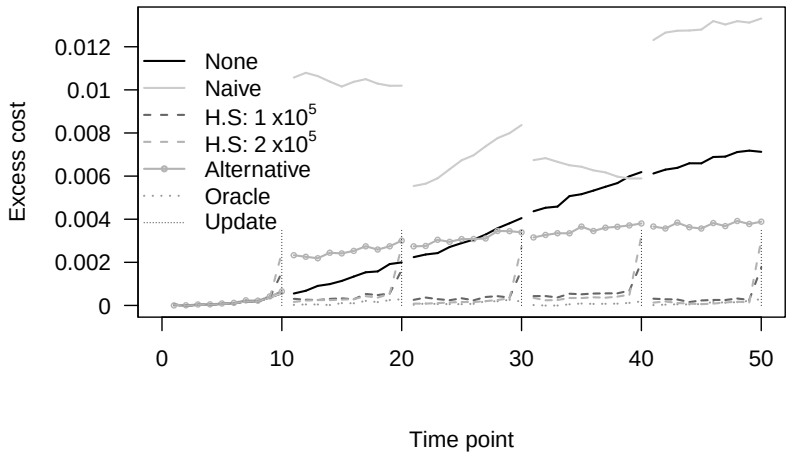


Fig 2: Cost per sample per unit of time for no-update, naïve-update, holdout-update, oracle, and alternative-update (using a treatment indicator as a covariate) strategies for a simulated example with population drift in risk. The costs associated with naïve updating will decrease during an epoch if f_t drifts towards the fitted risk score (that is, $\xi_t^2(f_t, \rho_t)$ decreases with t) and increase if it drifts away (that is, $\xi_t^2(f_t, \rho_t)$ increases with t). Supplementary Figure S1 shows the cumulative cost over time.

Supplementary Figure S5c shows the distributions and medians of OHS estimates using the parametric and emulation algorithms in settings with parametric assumptions either satisfied or not.

The results confirm expectations that the parametric OHS estimate is empirically unbiased and has less variance than the emulation estimate when parametric assumptions are satisfied, but is biased when they are not. Variance of OHS estimates using the emulation method is lower when parametric assumptions are not satisfied, because the true cost function has a sharper minimum in that case (see Supplementary Figure S5b). Since the cost function is 'flat' around the minimum in the setting where parametric assumptions are satisfied (Supplementary Figure S5b), the consequences of the high variance of the semi-parametric (emulation) estimator are minimal, as the cost is similar across a range of values near the OHS.

We next examine the convergence rates of OHS estimates when sampling the 'next' value of n, $\tilde{n}$, greedily, using Equation 48 for Algorithm 1 or EI for Algorithm 2, versus simply randomly selecting $\tilde{n}$ uniformly in $\{1, \ldots, N\}$. This is shown in Figure 3, which depicts medians and OHS estimates at various sizes of $|\mathbf{n}|$ under the different methods for selecting $\tilde{n}$.

Convergence is faster when next points are picked greedily rather than randomly and when using parametric estimates (though these are biased and inconsistent for k_2^{np}). This is highlighted by the smaller panels which show the root mean-square error between the total cost at the estimated optimal sizes and the total cost at the true OHS. In particular, observe that the non-parametric method shows bifurcation, detecting both local minima in the double descent setting, whilst the parametric method converges to a mid-point which is far from optimal in terms of total costs.

7. Application to ASPRE. We now return to our main motivating example. We are now in a position to address our main aim of developing an updating strategy for the ASPRE risk score.

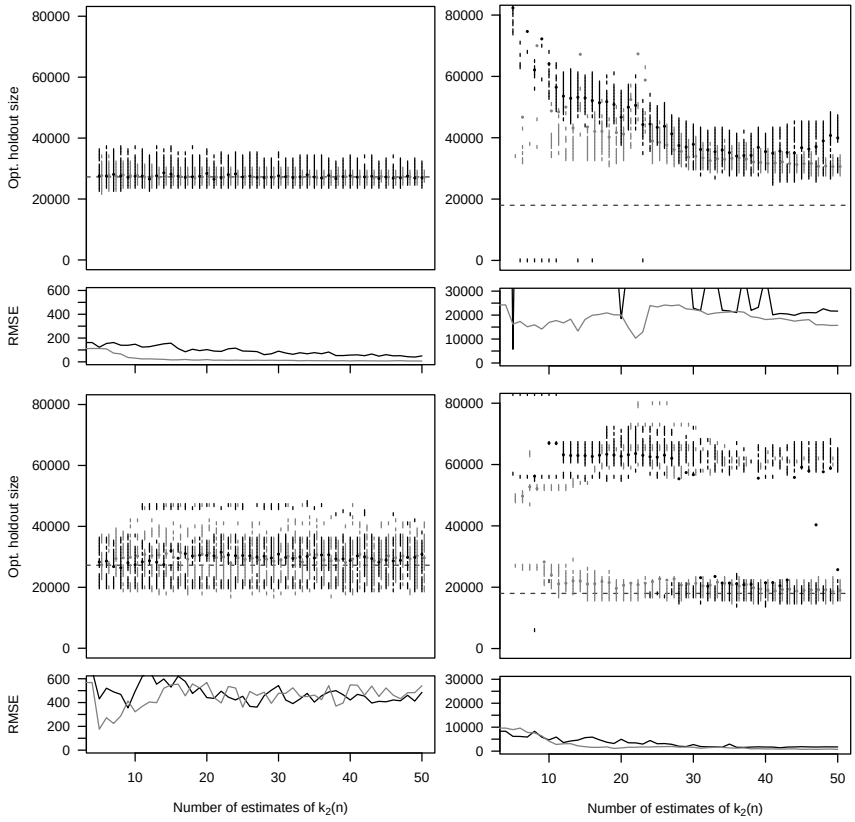


Fig 3: Convergence rates with parametric (top) and emulation (bottom) algorithms, using either a random (black) or greedy (gray) methods to select the next value, $\mathbf{n}$, with parametric assumptions satisfied (left) or unsatisfied (right). Simulations were run for 200 datasets from each underlying model. In larger panels, horizontal lines show true optimal holdout set (OHS) size; the OHS results from all simulation runs are discretised on a 1000-resolution grid, with vertical lines indicating OHS values that occurred in at least 2.5% of simulations (i.e., 5 occurrences). Smaller panels show root mean-square error between total costs from simulations and minimal total cost under random/greedy methods. Note variable axis scaling under the two models.

Supposing ASPRE is to be refitted every five years, the intervention set should include all individuals in the subsequent years before the model is refitted, and all individuals not used in the next refitting procedure. Suppose we refit ASPRE for use in a population of 5 million individuals, from which we have approximately 80,000 new pregnancies per year. The incidence of pregnancy per year is now $(8 \times 10^4)/(5 \times 10^6) = 1/125$ so we have $N \approx 5 \times 8 \times 10^4 = 400000$ (SE ≈ 1500). We must now estimate k_1 and $k_2(\cdot)$ from published data. Although this method is not especially generalisable, k_1 and $k_2(\cdot)$ will generally be more easily estimable given raw data, which is not publicly available.

We presume a simple clinical action in which a fixed proportion π of individuals at the highest assessed PRE risk are treated with aspirin. We assume $\pi = 10\% \approx 2707/25797$, the proportion of individuals assigned to the treatment group in Rolnik et al. (2017b) due to having an estimated risk of PRE $> 1\%$. We assume that if untreated with aspirin, a proportion π_0 of individuals designated to be 'low-risk' (lowest 90%) will develop PRE, as will a proportion π_1 of individuals designated high-risk.

To estimate π_0, π_1 and ultimately k_1, we considered the study reported in O'Gorman et al. (2017) assessing sensitivity and specificity of NICE and ACOG guidelines in assessing PRE risk. In this study, 8775 indivduals were assessed, amongst which 239 developed PRE, for an overall incidence of $239/8875 \approx 0.027$. We estimated the performance of a 'baseline' estimator of PRE risk (that is, in the absence of any ASPRE score) by linearly interpolating the points corresponding to 'ACOG aspirin', 'NICE' and 'ACOG' on ROC curves in Figure 1 of that paper. On this basis, a baseline estimator identifying the 10% of individuals at highest PRE risk (approximately 800) would correspond to the point (x, y) on the interpolated ROC curve with $239x + (8775 - 239)y = 0.1 \times 8875$ which occurs at roughly a 20% detection (true positive) rate and a 10% false positive rate, close to that of the NICE guidelines.

Since few women in the study were treated with aspirin, we assume that PRE rates in the highest-10% and lowest-10% risk groups assessed by baseline risk (NICE) are untreated risk (that is, if not treated with aspirin). At the inferred true and false positive rates, we would expect a PRE rate π_0 amongst the 10% of women designated highest-risk by the NICE guidelines and π_1 amongst the 90% designated lower risk, where

$$\pi_0 \approx \frac{(1 - \text{TPR}) \times (\text{Num. PRE})}{\text{Num. negative}} = \frac{0.8 \times 239}{0.9 \times 8875} \approx 0.024\,,$$
$$\pi_1 \approx \frac{\text{TPR} \times (\text{Num. PRE})}{\text{Num. positive}} = \frac{0.2 \times 239}{0.1 \times 8875} \approx 0.054\,,$$

with standard errors $SE(\pi_1) \approx 0.0076$ and $SE(\pi_0) \approx 0.0017$. We denote by α the relative reduction in PRE risk with aspirin treatment. Aspirin reduces PRE risk to approximately 63% (SE 0.09) of untreated risk (Rolnik et al., 2017a) so we take $\alpha = 1 - 0.63 = 0.37$. Now, treating errors in π_0, π_1 and α as pairwise independent, we have

$$k_1 = \pi_0(1 - \pi) + \pi_1\pi\alpha \approx 0.0235\,, \tag{12}$$

with $SE(k_1) = SE\left(\pi_0(1 - \pi) + \pi_1\pi\alpha\right) \approx 0.0016$. We estimate the population prevalence π_{PRE} of untreated PRE as the frequency observed in the original ASPRE data: $\pi_{\text{PRE}} = 1426/57974 \approx 2.4\%$. Note that, although this is approximately equal to π_0, they are different quantities: π_0 is the population frequency of PRE amongst individuals at the lowest 90% risk by NICE guidelines.

Denoting $\pi_1(n)$ as the untreated risk of PRE in the top 10% of individuals according to an ASPRE score trained on n individuals (and $\pi_0(n)$ correspondingly), we note that it is equal to the sensitivity (or TPR) of the risk score at the level where proportion π of individuals are designated high-risk. Thus for any training set size n, we have $\pi_0(n) = (\pi_{\text{PRE}} - \pi\pi_1(n))/(1 - \pi)$ so the average cost to an individual in the intervention set may be expressed in terms of $\pi_1(n)$:

$$k_2(n) = \pi_0(n)(1 - \pi) + \pi_1(n)\pi\alpha = \pi_{\text{PRE}} - \pi\pi_1(n)(1 - \alpha)\,.$$

We denote 'cost' as simply the number of cases of PRE in a population, so total expected cost per individual under 'baseline' treatment (clinical actions without the aid of a risk model) is

$$k_1 = \pi_0(1 - \pi) + \pi_1\pi\alpha \approx 0.02\,, \tag{13}$$

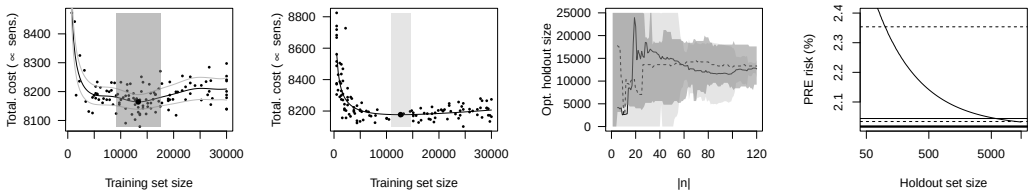
with standard error approximately 0.001. Note that this is not equal to the untreated PRE risk in the population, since some proportion of individuals are treated pre-emptively.

The data used to fit the initial ASPRE model could be used to estimate $\pi_1(n)$ and hence $k_2(n)$ for potential model updates. We do not have access to this dataset, but demonstrate estimation of a learning curve on synthetic data designed to resemble it. In this case, $k_2(n)$ is easy and fast to estimate, and is well-approximated by a power law, so we would favour use

of Algorithm 1. In order to mimic a real example where such estimation is time consuming or costly we use both algorithms and restrict ourselves to use only $|\mathbf{n}| = 120$ values of n, determined using either Algorithm 1 or 2. For both algorithms, we assumed a power-law form $k_2\{n; \theta = (a, b, c)\} = an^{-b} + c$.

Using the parametric algorithm, we found an OHS of 12684 (90% CI 10811-14556), with minimum cost (expected cases over five years) of 8177. Using the emulation algorithm, we found an OHS of 13313 with an expected cost of 8164, with holdout sizes of 9210-17619 having a probability > 0.1 of cost < 8164. Figure 4 shows estimated cost functions, OHSs, and error using the two algorithms. From our parametric approximation of $k_2(n)$, we estimate that if we use the suggested holdout set, the cost to a sample in the subsequent intervention set (that is, the PRE risk for a given pregnancy) is $k_2(12684) = 2.034\%$. This compares to a risk of $k_1 = 2.354\%$ in the current holdout set, and an overall risk of $8177/(5 \times 10^6) = 2.044\%$ across both sets.

We illustrate the practical mean risk per patient in Figure 4d. It is evident that the risk for patients in the holdout set (k_1, equivalent to use of a risk score fitted to 100 individuals), is only slightly higher than the risk to individuals in the intervention set, which is very close to the lowest possible risk and the average risk to all patients under the holdout set updating strategy (equivalent to a risk score fitted to around 6000 individuals).



(a) (b) (c) (d)

Fig 4: Estimation of cost functions (black lines), OHS (black dots), error using parametric (middle left) and emulation (leftmost; outer lines $\mu(n) + 3\sqrt{\Psi(n)}$) algorithms, change in estimated OHS and error with number of sample points $|\mathbf{n}|$ (middle right; solid lines: parametric, dashed lines: emulation), and overview of PRE risk per patient in final updating strategy (rightmost; solid black curve $k_2(n)$; heavy black line minimum possible risk $\min_n k_2(n)$; upper dashed line risk in holdout set; lower dashed line risk in intervention set; horizontal narrow solid line overall average risk). Note that the 'best' points (black dots) to optimize parametric estimation are spread-out to estimate θ well, but for emulation they are clustered for accurate local approximation. Error measures for OHS in parametric and emulation algorithms (red/blue shaded respectively) have different meanings and are not comparable.

In summary, we recommend that the ASPRE score use a holdout set of around 12,000 pregnancies in order to train an updated model, under the use assumptions outlined above. In keeping with our discussions in Section 3 and 4, we recommend that these samples are held-out late in the five-year period during which a given iteration of the model is in use. For held-out samples, the decision of whether to prescribe aspirin should be based on the best estimate of medical practitioners.

8. Concluding remarks. In this work we propose the use of a holdout set to safely update predictive models, and describe considerations in determining the optimal size of such a set. We establish theoretical properties of this optimal size under common conditions, and develop two algorithms for estimating it, evaluating their use in both a toy simulation and

a real-life motivated simulation. The holdout set approach comprises a practical and simple approach to an important problem in practical applied statistical modelling and machine learning, which will be increasingly important as risk scores start to be used ever more routinely to prompt intervention in real-world applications.

An appealing alternative for managing the effects of intervention without holdout sets is to attempt to explicitly infer parameters of an underlying causal structure (Alaa and van der Schaar, 2018; Sperrin et al., 2019). However, this approach cannot evade the difficulties we describe in Section 3: either we must be able to observe more detail (for instance, values of covariates after interventions which is often impractical) or make simplifying assumptions (such as absence of drift). In such relaxed settings, non-holdout-set options may allow a lower asymptotic and finite-sample cost than holdout set use. However, they will have higher asymptotic costs than holdout-set use in the more general setting considered here. A second potential non-holdout set option is to explicitly specify interventions which are to be made in response to risk scores, rather than leave them up to end-users (Ochs et al., 2019; Liley, 2021), but in typical complex settings in which we wish to use risk scores, we may wish to retain the autonomy of end users in decision-making. Indeed, prescriptive risk score based interventions would cause a model to fall under medical device regulation in the United Kingdom, so many risk scores there are developed for information only.

We believe that considerations for updating predictive models are under-studied in applied statistics. It is well-recognised that predictive models generally need to be updated Hippisley-Cox, Coupland and Brindle (2017); Kansagara et al. (2011); Wallace et al. (2014), but often this is planned to be done by simply refitting a predictive model to observed covariates and outcome data, corresponding to what we term 'naïve updating'. We note that the problem with naïve updating is not mentioned in the standard TRIPOD guideline (Collins et al., 2015). In a separate paper detailing the updating of a predictive risk score for emergency admissions in Scotland (Liley et al., 2024), we recognised the shortcomings of naïve updating, but in the absence of mature literature on safe updating methods or prior planning on managing updates, we were restricted in our choice of method. We proposed updating the risk score to the *maximum* of a previous risk score and a refitted risk score. This option will tend to lead to *overestimation* of risk, while avoiding *underestimation* of risk, and was satisfactory for a single update, though untenable in the longer term (Supplement S3.3).

Our methods may be extended in several ways. We do not consider the possibility of combining information over training iterations, which could reduce the number of training samples needed for a given prediction error. Nonetheless, retention of information is only useful up to a point: the main aim of using a holdout set is to correct for drift in f_t between times $e-1$ and e, which is described by the difference $f_e(x) - f_{e-1}(x)$ and not assisted by even perfect knowledge of $f_{e-1}(x)$. It is also possible that information from the intervention set could be used alongside the holdout set to partially infer the effect of interventions. We presume a setting in which a risk score is periodically updated, but continual or 'online' updating is also used (De Lange and Tuytelaars, 2021) and is also susceptible to intervention effects. A holdout set approach may also be usable in this setting. An implicit assumption of our work is that drift in μ_t, f_t and g_t are not influenced by our choice of strategy. Effectively, this means that the risk score affects the underlying system only through g_t, and does not affect the covariates or outcomes of samples for whom the risk score is not used. In complex settings such as medicine or finance, this assumption may be generally reasonable, but its relaxation is an important avenue for future research. For the ASPRE score, this corresponds to an expectation that treatment decisions made on the basis of the score will not causally influence the overall PRE incidence in the population (that is, f_t) or the characteristics of ultrasound scans or demographics (that is, μ_t).

Our simulations and theoretical findings show several non-obvious properties of the optimal holdout set size. We note that if the mean square error of the risk score decreases as $1/N$

(Assumption A4), then the OHS increases as $N^{1/2}$, with the immediate consequence that the OHS is a vanishing proportion of N, for large enough N. Moreover, in practical settings, the true OHS is fairly small, with rapidly diminishing returns to increasing risk score accuracy (Section 7). Interestingly, as demonstrated in Supplement S8.2, a more accurate risk score does not necessarily lead to lower OHS size. Given Corollary 1 (and as seen in Figure 4c), it is generally better to err on the higher end of the optimal holdout set size, since cost increases at most linearly, whereas it can increase faster for smaller holdout sets.

We propose two algorithms for estimating an optimal size for a holdout set. The parametric method is simple and converges rapidly, but the use of a Gaussian process emulator requires fewer assumptions. We advise use of the emulator method if the risk score is fitted using complex methods which may not lead to readily parametrisable k_2. A reasonable option if parametrisability of k_2 is uncertain is to use both estimation algorithms, and favour the emulation method if results disagree.

We strongly suggest planning an updating strategy for a risk model *before* it is deployed. This work illustrates one strategy in this direction and we hope stimulates both use of and extensions of such methods for safe predictive score updating.

Acknowledgments. The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

SH's contributions arose from an MSc dissertation for the MISCADA programme at Durham University. We thank Catalina Vallejos, Sebastian Vollmer and Bilal Mateen for helpful discussion.

Funding. LJMA was partially supported by a Health Programme Fellowship at the Alan Turing Institute. JL and LJMA were partially supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the 'Health' theme within that grant and The Alan Turing Institute; and by Health Data Research UK, an initiative funded by UKRI, Department of Health and Social Care (England), the devolved administrations, and leading medical research charities; SRE was funded by the EPSRC doctoral training partnership at Durham University, grant reference EP/R513039/1.

SUPPLEMENTARY MATERIAL

Supplement S1-S9
Contains proofs of Theorems 3.1, 3.2, 3.3, 4.1, 4.2, and 4.3, discussion of estimates of $k_2(n)$, details of algorithms, details of simulations, and details of the analysis of the ASPRE risk score.

Supplement S10
Supplementary Figures

Supplement S11
R scripts to reproduce figures and other output.

REFERENCES

ACOG (2016). Practice advisory on low-dose aspirin and prevention of preeclampsia: Updated recommendations. *American College of Obstetricians and Gynecologists (ACOG)*.

AKOLEKAR, R., SYNGELAKI, A., POON, L., WRIGHT, D. and NICOLAIDES, K. H. (2013). Competing risks model in early screening for preeclampsia by biophysical and biochemical markers. *Fetal diagnosis and therapy* **33** 8–15.

ALAA, A. M. and VAN DER SCHAAR, M. (2018). Autoprognosis: Automated clinical prognostic modeling via Bayesian optimization with structured kernel learning. *arXiv preprint arXiv:1802.07207*.

AMARI, S.-I. (1993). A universal theorem on learning curves. *Neural networks* **6** 161–166.

BROCHU, E., CORA, V. M. and DE FREITAS, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

CHALMERS, J., PULLAN, M., FABRI, B., MCSHANE, J., SHAW, M., MEDIRATTA, N. and POULLIS, M. (2013). Validation of EuroSCORE II in a modern cohort of patients undergoing cardiac surgery. *European Journal of Cardio-Thoracic Surgery* **43** 688–694.

CHISLETT, L., ASLETT, L. J. M., DAVIES, A. R., VALLEJOS, C. A. and LILEY, J. (2024). Ethical considerations of use of hold-out sets in clinical prediction model management. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00561-z

COLLINS, G. S., REITSMA, J. B., ALTMAN, D. G. and MOONS, K. G. (2015). Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *Circulation* **131** 211–219.

COOK, J. A. and COLLINS, G. S. (2015). The rise of big clinical databases. *British Journal of Surgery* **102** e93-e101. https://doi.org/10.1002/bjs.9723

DE LANGE, M. and TUYTELAARS, T. (2021). Continual prototype evolution: Learning online from non-stationary data streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 8250–8259.

DRUSVYATSKIY, D. and XIAO, L. (2020). Stochastic optimization with decision-dependent distributions. *arXiv preprint arXiv:2011.11173*.

FINLAYSON, S. G., SUBBASWAMY, A., SINGH, K., BOWERS, J., KUPKE, A., ZITTRAIN, J., KOHANE, I. S. and SARIA, S. (2020). The clinician and dataset shift in artificial intelligence. *The New England Journal of Medicine* 283–286.

HIPPISLEY-COX, J., COUPLAND, C. and BRINDLE, P. (2017). Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* **357**.

IZZO, Z., ZOU, J. and YING, L. (2021). How to Learn when Data Gradually Reacts to Your Model. *arXiv preprint arXiv:2112.07042*.

KANSAGARA, D., ENGLANDER, H., SALANITRO, A., KAGEN, D., THEOBALD, C., FREEMAN, M. and KRIPALANI, S. (2011). Risk prediction models for hospital readmission: a systematic review. *Jama* **306** 1688–1698.

KRANZER, K., SIMMS, V., DAUYA, E., OLARU, I. D., DZIVA CHIKWARI, C., MARTIN, K., REDZO, N., BANDASON, T., TEMBO, M., FRANCIS, S. C. et al. (2021). Identifying youth at high risk for sexually transmitted infections in community-based settings using a risk prediction tool: a validation study. *BMC Infectious Diseases* **21** 1–9.

LENERT, M. C., MATHENY, M. E. and WALSH, C. G. (2019). Prognostic models will be victims of their own success, unless.... *Journal of the American Medical Informatics Association* **26** 1645–1650.

LI, Q. and WAI, H.-T. (2021). State Dependent Performative Prediction with Stochastic Approximation. *arXiv preprint arXiv:2110.00800*.

LILEY, J. (2021). Stacking interventions for equitable outcomes. *arXiv preprint arXiv:2110.04163*.

LILEY, J., EMERSON, S. R., MATEEN, B. A., VALLEJOS, C. A., ASLETT, L. J. M. and VOLLMER, S. J. (2021). Model updating after interventions paradoxically introduces bias. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. *Proceedings of Machine Learning Research* **130** 3916–3924.

LILEY, J., BOHNER, G., EMERSON, S. R., MATEEN, B. A., BORLAND, K., CARR, D., HEALD, S., ODURO, S. D., IRELAND, J., MOFFAT, K., PORTEOUS, R., RIDDELL, S., ROGERS, S., CUNNINGHAM, N., HOLMES, C., PAYNE, K., VOLLMER, S. J., VALLEJOS, C. A. and ASLETT, L. J. M. (2024). Development and assessment of a machine learning tool for predicting emergency admission in Scotland. *npj Digital Medicine* (to appear).

LU, J., LIU, A., DONG, F., GU, F., GAMA, J. and ZHANG, G. (2018). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering* **31** 2346–2363.

MENDLER-DÜNNER, C., PERDOMO, J. C., ZRNIC, T. and HARDT, M. (2020). Stochastic optimization for performative prediction. *arXiv preprint arXiv:2006.06887*.

NASHEF, S. A., ROQUES, F., SHARPLES, L. D., NILSSON, J., SMITH, C., GOLDSTONE, A. R. and LOCKOWANDT, U. (2012). EuroSCORE II. *European journal of cardio-thoracic surgery* **41** 734–745.

OCHS, A., MCGURNAGHAN, S., BLACK, M. W., LEESE, G. P., PHILIP, S., SATTAR, N., STYLES, C., WILD, S. H., MCKEIGUE, P. M., COLHOUN, H. M. et al. (2019). Use of personalised risk-based screening schedules to optimise workload and sojourn time in screening programmes for diabetic retinopathy: a retrospective cohort study. *PLoS Medicine* **16** e1002945.

O'GORMAN, N., WRIGHT, D., POON, L., ROLNIK, D. L., SYNGELAKI, A., DE ALVARADO, M., CARBONE, I. F., DUTEMEYER, V., FIOLNA, M., FRICK, A. et al. (2017). Multicenter screening for pre-eclampsia by maternal factors and biomarkers at 11–13 weeks' gestation: comparison with NICE guidelines and ACOG recommendations. *Ultrasound in Obstetrics & Gynecology* **49** 756–760.

PERDOMO, J., ZRNIC, T., MENDLER-DÜNNER, C. and HARDT, M. (2020). Performative prediction. In *International Conference on Machine Learning* 7599–7609. PMLR.

RILEY, R. D., ENSOR, J., SNELL, K. I., HARRELL, F. E., MARTIN, G. P., REITSMA, J. B., MOONS, K. G., COLLINS, G. and VAN SMEDEN, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ* **368**.

ROLNIK, D. L., WRIGHT, D., POON, L. C., O'GORMAN, N., SYNGELAKI, A., DE PACO MATALLANA, C., AKOLEKAR, R., CICERO, S., JANGA, D., SINGH, M. et al. (2017a). Aspirin versus placebo in pregnancies at high risk for preterm preeclampsia. *New England Journal of Medicine* **377** 613–622.

ROLNIK, D. L., WRIGHT, D., POON, L., SYNGELAKI, A., O'GORMAN, N., DE PACO MATALLANA, C., AKOLEKAR, R., CICERO, S., JANGA, D., SINGH, M. et al. (2017b). ASPRE trial: performance of screening for preterm pre-eclampsia. *Ultrasound in obstetrics & gynecology* **50** 492–495.

SHAHIAN, D. M., JACOBS, J. P., BADHWAR, V., KURLANSKY, P. A., FURNARY, A. P., CLEVELAND JR, J. C., LOBDELL, K. W., VASSILEVA, C., VON BALLMOOS, M. C. W., THOURANI, V. H. et al. (2018). The Society of Thoracic Surgeons 2018 adult cardiac surgery risk models: part 1—background, design considerations, and model development. *The Annals of thoracic surgery* **105** 1411–1418.

SPERRIN, M., JENKINS, D., MARTIN, G. P. and PEEK, N. (2019). Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. *Journal of the American Medical Informatics Association* **26** 1675–1676.

STALLARD, N., MILLER, F., DAY, S., HEE, S. W., MADAN, J., ZOHAR, S. and POSCH, M. (2017). Determination of the optimal sample size for a clinical trial accounting for the population size. *Biometrical Journal* **59** 609–625.

TOPOL, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* **25** 44–56.

TSYMBAL, A. (2004). The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin* **106** 58.

USFDA et al. (2019). Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD).

VIERING, T. and LOOG, M. (2021). The Shape of Learning Curves: a Review. *arXiv preprint arXiv:2103.10948*.

WALLACE, E., STUART, E., VAUGHAN, N., BENNETT, K., FAHEY, T. and SMITH, S. M. (2014). Risk prediction models to predict emergency hospital admission in community-dwelling adults: a systematic review. *Medical care* **52** 751.

ZUKANOVIĆ, A. (2013). Caries risk assessment models in caries prediction. *Acta medica academica* **42**.

Holdout sets for safe predictive model updating
Supplementary material

Sami Haidar-Wehbe[1,⊥], Samuel R. Emerson[1,⊥], Louis J. M. Aslett[1,2,*], and James Liley[1,*]

[1]Department of Mathematical Sciences, Durham University, Durham, UK
[2]Alan Turing Institute, London, UK
[⊥]Equal contribution
[*]Corresponding

October 23, 2024

Contents

S1 General notation

This supplement requires several sets of notation, which will be introduced as needed. However, we note some commonalities. Throughout this document we will take X to generically mean 'covariates' and Y to mean 'outcomes'. The subscript t will be taken to mean 'time' in a continuous sense, and the subscript e to mean 'epoch', referring to consecutive episodes of time. The superscript h will correspond to the holdout set, and i to the intervention set. The number N will refer to the total number of samples on which a risk score may be trained or used during an epoch, and n to denote a holdout set size, usually taken to be variable. We denote the standard normal PDF and CDF by $\phi(\cdot)$, $\Phi(\cdot)$ respectively, the Bernoulli distribution with parameter p as $\text{Bern}(p)$, and the Poisson distribution with parameter λ as $\text{Pois}(\lambda)$.

S2 Proofs of Theorems 3.1, 3.2 and 3.3

S2.1 Theorem 3.1

Theorem 3.1. *Suppose we use a holdout set with size $n_\star = \Theta(N^a)$, with $0 < a < 1$, an $s < 1$ of size $s = \Theta(N^{a+\epsilon-1})$ for some ϵ with $0 < \epsilon < 1 - a$, and an update frequency $\delta \leq 1$ which may vary with N. Under assumptions A3, A4, A6 and A7, we have*

$$\mathbb{E}\left\{C_{(h)}[0,T]\right\} = \delta N T k_2(2\alpha_1 + \alpha_1^2) + \delta^{-1}O(N^a) + O(N^{a+\epsilon}) + O(N^{1-a})\,.$$

Proof. We begin by establishing an inequality on the quantity

$$\mathbb{E}\{\xi_t^2(\rho_f^{n_*,e,s}, f_t)\}\,, \tag{1}$$

where $t - e < \delta$ and the expectation is over the data used to fit $\rho_f^{n,e,s}$.

As in assumption A4, denote $F(x) = \mathbb{E}_{t\sim U(e-s,e)}\{f_t(x)\}$ and note that, by assumption A3, we have

$$f_e(x) - \alpha_1 s \leq F(x) \leq f_e(x) + \alpha_1 s\,. \tag{2}$$

To establish an upper bound, we use (in order) assumptions A3, inequality 2, and assumption A4. Taking these, along with noting that $|\rho(x) - f(x)| \leq 1$ for any x, ρ, we have:

$$\begin{aligned}
\mathbb{E}\left\{\xi_t^2(\rho_f^{n_*,e,s}, f_t)\right\} &= \mathbb{E}\left\{\int\left(\rho_f^{n_*,e,s}(x) - f_t(x)\right)^2 d\mu_t\right\} \\
&\leq \mathbb{E}\left\{\int\left(\left|\rho_f^{n_*,e,s}(x) - f_e(x)\right| + |f_t(x) - f_e(x)|\right)^2 d\mu_t\right\} \\
&\leq \mathbb{E}\left\{\xi_t^2\left(\rho_f^{n_*,e,s}, f_e\right) + 2\int|f_t(x) - f_e(x)|\,d\mu_t + \xi_t^2(f_t, f_e)\right\} \\
&\leq \mathbb{E}\left\{\int\left(\rho_f^{n_*,e,s}(x) - f_e(x)\right)^2 d\mu_t\right\} + 2\alpha_1(t-e) + \alpha_1^2(t-e)^2 \\
&\leq \mathbb{E}\left\{\int\left(\left|\rho_f^{n_*,e,s}(x) - F(x)\right| + |F(x) - f_e(x)|\right)^2 d\mu_t\right\} + 2\delta\alpha_1 + \delta^2\alpha_1^2 \\
&\leq \mathbb{E}\left\{\int\left(\left|\rho_f^{n_*,e,s}(x) - F(x)\right| + \alpha_1 s\right)^2 d\mu_t\right\} + \delta(2\alpha_1 + \alpha_1^2) \\
&\leq \mathbb{E}\left\{\xi_t^2\left(\rho_f^{n_*,e,s}, F\right)\right\} + (2\alpha_1 + \alpha_1^2)(\delta + s) \\
&= O(n_*^{-1}) + (2\alpha_1 + \alpha_1^2)(\delta + s)\,.
\end{aligned} \tag{3}$$

In any period $(e - \delta, e]$, the probability p_* of at least n_* samples being encountered in $(e - s, e)$, given $\frac{n_*}{Ns} = \Theta(N^{-\epsilon}) \to 0$ and $Ns = \Theta(N^{1-a+\epsilon}) \to \infty$, satisfies the (weak) condition:

$$p_* = P\left(\text{Pois}(Ns) \geq n_*\right) = 1 - O(N^{-2})\,. \tag{4}$$

Consider costs accrued in the period $(e - \delta, e]$ with $e > \delta$, under the holdout strategy. We encounter some total number of samples $n \sim \text{Pois}(N\delta)$. We assign at most n_* samples to the holdout set, with a total cost of at most $k_1 n_*$ (as a consequence of Assumption A5). The remaining samples each accrue a cost proportional to 3, as long as at least n_* samples were observed in $(e - \delta - s, e]$. We thus have

$$\mathbb{E}\left\{C_{(h)}[e-\delta, e]\right\} \leq \underbrace{k_1 n_*}_{\substack{\text{Cost for held} \\ \text{-out samples}}} + \mathbb{E}\{n\}\Bigg(\underbrace{(1-p_*)k_1}_{\substack{\text{Cost if } < n_* \\ \text{samples in} \\ (e-1-s, e]}} + \underbrace{p_*\mathbb{E}\left\{k_2\xi_t^2(\rho_f^{n_*,e,s}, f_t)\right\}}_{\substack{\text{Cost for non-held-} \\ \text{out samples otherwise}}}\Bigg) \tag{5}$$

$$
\begin{aligned}
&\leq k_1 n_* + N\delta \left(O(N^{-2}) + k_2 \mathbb{E}\left\{ \xi_t^2(\rho_f^{n_*,e,s}, f_t) \right\} \right) \\
&\leq k_1 n_* + O(\delta N^{-1}) + \delta N O(n_*^{-1}) + \delta N k_2 (2\alpha_1 + \alpha_1^2)(\delta + s) \\
&\leq \Theta(N^a) + \delta O(N^{1-a}) + \delta\Theta(Ns) + \delta^2 N k_2 (2\alpha_1 + \alpha_1^2)) \\
&= O(N^a) + \delta O(N^{a+\epsilon}) + \delta O(N^{1-a}) + \delta^2 N k_2 (2\alpha_1 + \alpha_1^2)) \,. \qquad (6)
\end{aligned}
$$

The total cost accrued over the T/δ total epochs is thus

$$
\begin{aligned}
\mathbb{E}\left\{ C_{(h)}[0,T] \right\} &= \frac{1}{\delta} \mathbb{E}\left\{ C_{(h)}[e-\delta, e] \right\} \\
&= \delta^{-1} O(N^a) + O(N^{a+\epsilon}) + O(N^{1-a}) + \delta N T k_2 (2\alpha_1 + \alpha_1^2)) \qquad (7) \\
&= \delta^{-1} O(N^a) + O(N^{a+\epsilon}) + O(N^{1-a}) + \delta O(N) \,, \qquad (8)
\end{aligned}
$$

where the bounds in $O(\cdot)$ depend only on k_1, k_2, α_1, T. We note that this result holds for *any* fixed T rather than only the T in assumption A1.

□

S2.2 Theorem 3.2

Theorem 3.2. *Suppose we choose s such that $s \to 0$ as $N \to \infty$. Under assumptions A1-A7, for sufficiently small δ, we have for (strat) $\in \{(0), (n), (a)\}$:*

$$\mathbb{E}\left\{C_{(strat)}[0,T]\right\} = \Omega(N)\,. \tag{9}$$

Proof. We will begin with a simple lemma which we will use repeatedly:

Lemma S1. *If, for all x, we have $0 \leq f(x), g(x), h(x) \leq 1$, then*

$$\xi_t^2(f,g) \geq \xi_t^2(f,h) + \xi_t^2(h,g) - 2\sqrt{\xi_t^2(h,g)}\,.$$

Proof. We have

$$\begin{aligned}
\xi_t^2(f,g) &= \int (f(x) - g(x))\, d\mu_t \\
&\geq \int (|f(x) - h(x)| - |h(x) - g(x)|)\, d\mu_t \\
&\geq \int (f(x) - h(x))^2\, d\mu_t + \int (h(x) - g(x))^2\, d\mu_t - 2\int |f(x) - h(x)||h(x) - g(x)| d\mu_t \\
&\geq \xi_t^2(f,h) + \xi_t^2(h,g) - 2\int |h(x) - g(x)| d\mu_t \qquad (10) \\
&\geq \xi_t^2(f,h) + \xi_t^2(h,g) - 2\sqrt{\int (h(x) - g(x))^2\, d\mu_t} \qquad (11) \\
&\geq \xi_t^2(f,h) + \xi_t^2(g,h) - 2\sqrt{\xi_t^2(h,g)}\,,
\end{aligned}$$

using the fact that $|f(x) - h(x)| \leq 1$ at step 10 and the Cauchy-Schwarz inequality at step 11. □

We secondly prove a short lemma to show that we need not consider the no-update strategy separately from the alternative strategy:

Lemma S2. *Strategy (0) is a special case of strategy (a).*

Proof. We note that from assumption A1 we have

$$\mathbb{E}_{t\sim U[0,T]}\{\xi_t^2(f_0, f_t)\} = \int_0^T \frac{1}{T}\xi_t^2(f_0, f_t)dt > 0\,.$$

As $N \to \infty$, since $s \to 0$, we have from assumption A4:

$$\mathbb{E}\left\{\xi_0^2\left(\rho_f^{N,0,s}, f_0\right)\right\} \to 0\,, \tag{12}$$

so we have (using Lemma S1)

$$\begin{aligned}
\lim_{N\to\infty}\left(\mathbb{E}_{t\sim U[0,T]}\left\{\xi_t^2(\rho_f^{N,0,s}, f_t)\right\}\right) &\geq \lim_{N\to\infty}\left(\mathbb{E}_{t\sim U[0,T]}\left\{\xi_t^2(f_0, f_t) - 2\sqrt{\xi_t^2(\rho_f^{N,0,s}, f_0)}\right\}\right) \\
&= \mathbb{E}_{t\sim U[0,T]}\{\xi_t^2(f_0, f_t)\} \\
&> 0\,, \qquad (13)
\end{aligned}$$

where the expectation is also over data used to fit $\rho_f^{N,0,s}$. Hence the no-update strategy is in the 'alternative' class of strategies, with $\rho_t = \rho_f^{N,0,s}$ for all t.

□

Recalling our definition of b as

$$\lim_{N\to\infty}\left(\mathbb{E}_{t\sim U[0,T],D}\left\{\xi_t^2(\rho_t,f_t)\right\}\right)=b>0\,,$$

we simply choose any ϵ with $0<\epsilon<b$, so for large enough N, we have

$$\begin{aligned}\mathbb{E}\left\{C_{(a)}[0,T]\right\}&=\mathbb{E}\{\text{Pois}(NT)\}\cdot k_2\mathbb{E}_{t\sim U(0,T),D}\left\{\mathbb{E}\left\{\xi_t^2(\rho_t,f_t)\right\}\right\} &(14)\\ &\geq k_2NT(b-\epsilon)\\ &=\Omega(N) &(15)\end{aligned}$$

as required, establishing the theorem for strat $=(0)$ and strat $=(a)$.

We now establish the rate of growth of the costs of the naive update strategy. The essential idea is

1. We consider two consecutive time periods $[(i-1)\delta,i\delta)$ and $[i\delta,(i+1)\delta)$, with $i\geq 2$.
2. We introduce an 'index' value Δ_0 which is the similarity of the risk score $\rho_g^{N,(i-1)\delta,s}$ used during period $[(i-1)\delta,i\delta)$ to the function $f_{(i-1)\delta}$ governing risk at the start of that period.
3. Given assumption A3, we establish that $f_{(i-1)\delta-s}$ is not very different from f_t with $t\in[(i-1)\delta,i\delta)$, so Δ_0 is similar to the difference between the risk score and the function f_t throughout the period $[(i-1)\delta,i\delta)$. We conclude that Δ_0 governs the total cost accrued during time period $[(i-1)\delta,i\delta)$, in that the larger Δ_0, the larger the total cost.
4. We then consider the similarity between g and f during the period $[i\delta-s,i\delta)$ during which the risk score $\rho_g^{N,i\delta,s}$ is fitted for use during period $[i\delta,(i+1)\delta)$. Since the 'cost' (the difference between f_t and g_t) and 'inaccuracy' (the difference between the risk score and f_t) are related by assumptions A5, A6, the difference between f_t and g_t is also governed by Δ_0, with a larger Δ_0 corresponding to a *smaller* difference.
5. Since $\rho_g^{N,i\delta,s}$ (the risk score for use in time period $[i\delta,(i+1)\delta]$ is fitted to g_t with $t\in[i\delta-s,i\delta)$, the similarity between f_t and g_t in this period is also the similarity between f_t and the new risk score. We establish that f_t is similar in time period $[i\delta-s,i\delta]$ and in time period $[i\delta,(i+1)\delta]$, so the difference between f_t and g_t, and hence the difference between f_t and the risk score is largely conserved. Since a small Δ_0 means a large difference between f_t and g_t for $t\in[i\delta-s,i\delta]$, it means a large difference between the risk score and f_t for $t\in[i\delta,(i+1)\delta]$.
6. Thus a small Δ_0 means a large cost in time period $[i\delta,(i+1)\delta]$ and a large Δ_0 means a large cost in $[(i-1)\delta,i\delta]$. We show that there is a non-negligible cost accrued overall. When summed across all such time periods, this results in a $\Omega(N)$ contribution to overall cost.

When using the naive update strategy, the costs during the time period $[i\delta,(i+1)\delta)$ depend on the similarity between f_t and g_t during the time period $[(i-1)\delta,i\delta)$. The risk score used during the time period $[(i-1)\delta,i\delta)$ under the naive update strategy is $\rho_g^{N,(i-1)\delta,s}$. We define

$$\Delta_0\triangleq\xi^2_{(i-1)\delta-s}\left(\rho_g^{N,(i-1)\delta,s},f_{(i-1)\delta}\right)\,.$$

Note that this is not an expectation; we will show a bound on the accrued costs which does not depend on Δ_0. Denote by α_2 the Lipschitz constant of μ_t in assumption A2. Given assumption A3, we have, for $t\in[(i-1)\delta,i\delta)$ (using a tighter bound than Lemma S1):

$$\xi_t^2\left(\rho_g^{N,(i-1)\delta,s},f_t\right)=\int\left(\rho_g^{N,(i-1)\delta,s}(x)-f_t(x)\right)^2d\mu_t$$

$$
\begin{aligned}
&\leq \int \left(\left| \rho_g^{N,(i-1)\delta,s}(x) - f_{(i-1)\delta}(x) \right| + \left| f_{(i-1)\delta}(x) - f_t(x) \right| \right)^2 d\mu_t \\
&\leq \xi_t^2 \left(\rho_g^{N,(i-1)\delta,s}, f_{(i-1)\delta} \right) + 2 \int \left| f_{(i-1)\delta}(x) - f_t(x) \right| d\mu_t \\
&\quad + \int \left(f_{(i-1)\delta}(x) - f_t(x) \right)^2 d\mu_t \\
&\leq \xi_{(i-1)\delta-s}^2 \left(\rho_g^{N,(i-1)\delta,s}, f_{(i-1)\delta} \right) + \alpha_2(t - ((i-1)\delta - s)) \\
&\quad + 2\alpha_1(t - (i-1)\delta) + \alpha_1^2\delta^2 \qquad \text{Asm. A3,A2} \\
&\leq \Delta_0 + (\alpha_2 + 2\alpha_1)\delta + \alpha_1^2\delta^2 + \alpha_2 s \\
&= \Delta_0 + m_s \,. \qquad (16)
\end{aligned}
$$

denoting, for brevity,

$$
m_s = (\alpha_2 + 2\alpha_1)\delta + \alpha_1^2\delta^2 + \alpha_2 s \,. \qquad (17)
$$

Note that by choosing a large enough N (and hence sufficiently small s) and a sufficiently small δ we may ensure m_s is arbitrarily small.

Using similar arguments, we have

$$
\begin{aligned}
\xi_t^2 \left(\rho_g^{N,(i-1)\delta,s}, f_t \right) &= \int \left(\rho_g^{N,(i-1)\delta,s}(x) - f_t(x) \right)^2 d\mu_t \\
&\geq \int \left(\left| \rho_g^{N,(i-1)\delta,s}(x) - f_{(i-1)\delta}(x) \right| - \left| f_{(i-1)\delta}(x) - f_t(x) \right| \right)^2 d\mu_t \\
&\geq \xi_t^2 \left(\rho_g^{N,(i-1)\delta,s}, f_{(i-1)\delta} \right) - 2 \int \left| f_{(i-1)\delta}(x) - f_t(x) \right| d\mu_t \\
&\geq \xi_{(i-1)\delta-s}^2 \left(\rho_g^{N,(i-1)\delta,s}, f_{(i-1)\delta} \right) - \alpha_2(t - ((i-1)\delta - s)) \\
&\quad - 2\alpha_1(t - (i-1)\delta) \qquad \text{Asms. A3,A2} \\
&= \Delta_0 - (\alpha_2 + 2\alpha_1)(t - (i-1)\delta) - \alpha_2 s \\
&\geq \Delta_0 - \delta(\alpha_2 + 2\alpha_1) - \alpha_2 s \\
&\geq \Delta_0 - m_s \,. \qquad (18)
\end{aligned}
$$

For any t in the time period $[i\delta - s, i\delta)$, we have, by assumptions A6 and A5:

$$
k_1 - \mathbb{E}_{X \sim \mu_t} \left\{ f_t(X) - g_t(X) \right\} = k_2 \xi_t^2 \left(\rho_g^{N,(e-1)\delta,s}, f_t \right) \,. \qquad (19)
$$

We now consider two cases.

Case 1. $\mathbb{E}_{X \sim \mu_\tau} \left\{ f_\tau(X) - g_\tau(X) \right\} < 0$ *for some* $\tau \in [i\delta - s, i\delta)$

Conceptually, in this case, use of the risk score $\rho_g^{N,(e-1)\delta,s}$ in fact makes the risk worse than would the use of no risk score at all at time τ. In this case, we have, by assumption:

$$
\xi_\tau^2 \left(\rho_g^{N,(e-1)\delta,s}, f_\tau \right) > \frac{k_1}{k_2} \,,
$$

so for any $t \in [(i-1)\delta, i\delta)$:

$$
\xi_t^2 \left(\rho_g^{N,(e-1)\delta,s}, f_t \right) \geq \xi_t^2 \left(\rho_g^{N,(e-1)\delta,s}, f_\tau \right) - 2\sqrt{\xi_t^2 \left(f_t, f_\tau \right)}
$$

$$\geq \xi_\tau^2\left(\rho_g^{N,(e-1)\delta,s}, f_\tau\right) - \alpha_2|t-\tau| - 2\alpha_1|t-\tau|$$
$$\geq \frac{k_1}{k_2} - (\alpha_2 + 2\alpha_1)\delta\,,$$

and the total expected cost accrued over the time period $[(i-1)\delta, (i+1)\delta]$ (during which we encounter Pois$(2N\delta)$ samples) is:

$$\begin{aligned}\mathbb{E}\left\{C_{(n)}[(i-1)\delta,(i+1)\delta]\right\} &\geq \mathbb{E}\left\{C_{(n)}[(i-1)\delta, i\delta]\right\} \\ &= \mathbb{E}\left\{\text{Pois}(N\delta)\right\} k_2 \mathbb{E}_{t\sim U((i-1)\delta, i\delta)}\left\{\xi_t^2\left(\rho_g^{N,(e-1)\delta,s}, f_t\right)\right\} \\ &\geq N\delta\left(k_1 - k_2(\alpha_2+2\alpha_1)\delta\right)\,,\end{aligned} \tag{20}$$

and for any ϵ we may choose δ dependent only on α_2, α_1, k_2 sufficiently small that

$$\mathbb{E}\left\{C_{(n)}[(i-1)\delta,(i+1)\delta]\right\} \geq N\delta\left(k_1 - \epsilon\right)\,. \tag{21}$$

Case 2. $\mathbb{E}_{X\sim\mu_t}\left\{f_t(X) - g_t(X)\right\} > 0$ *for all* $t \in [i\delta - s, i\delta)$.

The risk score used during the period $[i\delta, (i+1)\delta)$ is $\rho_g^{N,i\delta,s}$. Denote $G(x) = \mathbb{E}_{t\sim U(i\delta-s,i\delta)}\{g_t(x)\}$ (so $\rho_g^{N,i\delta,s}(x)$ estimates $G(x)$). We now have:

$$\begin{aligned}\xi_{i\delta}^2\left(f_{i\delta}, G\right) &= \int \left(f_{i\delta}(x) - G(X)\right)^2 d\mu_{i\delta} \\ &\geq \left(\int \left|f_{i\delta}(x) - G(x)\right| d\mu_{i\delta}\right)^2 \\ &\geq \left(\int \left(f_{i\delta}(x) - G(x)\right) d\mu_{i\delta}\right)^2 \\ &= \left(\int \frac{1}{s}\int_{i\delta-s}^{i\delta} (f_{i\delta}(x) - g_t(x)) dt\, d\mu_{i\delta}\right)^2 \\ &= \left(\frac{1}{s}\int_{i\delta-s}^{i\delta}\left(\int (f_t(x) - g_t(x)) d\mu_{i\delta} + \int (f_{i\delta}(x) - f_t(x)) d\mu_{i\delta}\right) dt\right)^2 \\ &\geq \left(\frac{1}{s}\int_{i\delta-s}^{i\delta} \max\left(0, \mathbb{E}_{X\sim\mu_{i\delta}}\left\{f_t(x) - g_t(x)\right\} - \alpha_1\delta\right) dt\right)^2 && \text{Asm. A3} \\ &\geq \left(\frac{1}{s}\int_{i\delta-s}^{i\delta} \max\left(0, \mathbb{E}_{X\sim\mu_t}\left\{f_t(x) - g_t(x)\right\} - \alpha_2 s - \alpha_1\delta\right) dt\right)^2 && \text{Asm. A2} \\ &\geq \left(\frac{1}{s}\int_{i\delta-s}^{i\delta} \max\left(0, k_1 - k_2\xi_t^2\left(\rho_g^{N,(e-1)\delta,s}, f_t\right) - \alpha_2 s - \alpha_1\delta\right) dt\right)^2 && \text{Asm. A5} \\ &\geq \max\left(0, k_1 - k_2\left(\Delta_0 + (\alpha_2+2\alpha_1)\delta + \alpha_1^2\delta^2 + \alpha_2 s\right) - \alpha_2 s - \alpha_1\delta\right)^2 && \text{Ineq. 16} \\ &\geq \max\left(0, k_1 - k_2\left(\Delta_0 + m_s\right) - m_s\right)^2\,, && (22)\end{aligned}$$

We consider the expectation over the data used to fit $\rho_g^{N,i\delta,s}$ to note that for $t \in [i\delta, (i+1)\delta)$, using Lemma S1:

$$\begin{aligned}\mathbb{E}\left\{\xi_t^2(\rho_g^{N,i\delta,s}, f_t)\right\} &\geq \xi_t^2(G, f_t) - 2\sqrt{\xi_t^2\left(\rho_g^{N,i\delta,s}, G\right)} \\ &\geq \xi_t^2(G, f_t) - O\left(N^{-\frac{1}{2}}\right)\end{aligned}$$

$$
\begin{aligned}
&\geq \xi_t^2(G, f_{i\delta}) - 2\sqrt{\xi_t^2(f_t, f_{i\delta})} - O\left(N^{-\frac{1}{2}}\right) \\
&\geq \xi_t^2(G, f_{i\delta}) - 2\alpha_1(t - i\delta) - O\left(N^{-\frac{1}{2}}\right) && \text{Asm. A3} \\
&\geq \xi_{i\delta}^2(G, f_{i\delta}) - \alpha_2(t - i\delta) - 2\alpha_1(t - i\delta) - O\left(N^{-\frac{1}{2}}\right) && \text{Asm. A2} \\
&\geq \xi_{i\delta}^2(G, f_{i\delta}) - \delta(\alpha_2 + 2\alpha_1) - O\left(N^{-\frac{1}{2}}\right) .
\end{aligned}
$$

The expected total cost accrued during the period $[(i-1)\delta, i\delta)$ (during which we encounter $\text{Pois}(N\delta)$ samples) is thus, from expression 18:

$$
\begin{aligned}
\mathbb{E}\left\{C_{(n)}[(i-1)\delta, i\delta]\right\} &= \mathbb{E}\left\{\text{Pois}(N\delta)\right\} k_2 \mathbb{E}_{t\sim U((i-1)\delta, i\delta)} \left\{\xi_t^2\left(\rho_g^{N,(i-1)\delta,s}, f_t\right)\right\} \\
&\geq N\delta k_2 \left(\Delta_0 - \delta(\alpha_2 + 2\alpha_1) - \alpha_2 s\right) \\
&\geq N\delta k_2 \left(\Delta_0 - m_s\right) .
\end{aligned} \tag{23}
$$

The total cost accrued during the period $[i\delta, (i+1)\delta)$ is, from expression 22

$$
\begin{aligned}
\mathbb{E}\left\{C_{(n)}[i\delta, (i+1)\delta]\right\} &= \mathbb{E}\left\{\text{Pois}(N\delta)\right\} k_2 \mathbb{E}_{t\sim U(i\delta, (i+1)\delta)} \left\{\xi_t^2\left(\rho_g^{N,i\delta,s}, f_t\right)\right\} \\
&\geq N\delta k_2 \left(k_1 - k_2(\Delta_0 - m_s) - m_s\right)^2 ,
\end{aligned} \tag{24}
$$

and hence the total cost over both periods is

$$
\mathbb{E}\left\{C_{(n)}[(i-1)\delta, (i+1)\delta]\right\} \geq N\delta k_2 \left(\Delta_0 - m_s + \left(k_1 - k_2(\Delta_0 - m_s) - m_s\right)^2\right) .
$$

For any ϵ we may choose δ (dependent only on k_1, k_2, α_2, α_1) sufficiently small that for large enough N:

$$
\mathbb{E}\left\{C_{(n)}[(i-1)\delta, (i+1)\delta]\right\} \geq N\delta \left(k_2 \min_{0\leq\Delta_0\leq 1} \left(\Delta_0 + (k_1 - k_2\Delta_0)^2\right) - \epsilon\right) .
$$

Recalling expression 21 for the earlier case, we denote

$$
c_{(n)} = \min\left(k_2 \min_{0\leq\Delta_0\leq 1} \left(\Delta_0 + (k_1 - k_2\Delta_0)^2\right), k_1\right) > 0 , \tag{25}
$$

so, in either case:

$$
\mathbb{E}\left\{C_{(n)}[(i-1)\delta, (i+1)\delta]\right\} \geq N\delta \left(c_{(n)} - \epsilon\right) .
$$

We now finally consider the cost accrued over the entire time period $[0, T]$, where $T > 2\delta$. We have

$$
\begin{aligned}
\mathbb{E}\left\{C_{(n)}[0, T]\right\} &= \mathbb{E}\left\{\sum_{i=0}^{T/\delta-1} C_{(n)}[i\delta, (i+1)\delta]\right\} \\
&\geq \frac{1}{2}\mathbb{E}\left\{\sum_{i=1}^{T/\delta-1} C_{(n)}[(i-1)\delta, (i+1)\delta]\right\} \\
&\geq \frac{1}{2}\left(\frac{T}{\delta} - 1\right) N\delta(c_{(n)} - \epsilon) \\
&= \Omega(N)
\end{aligned}
$$

as required.

□

S2.3 Theorem 3.3

Theorem 3.3. *Consider use of each strategy in parallel with an 'oracle' procedure. Under assumptions A1-A7, with sufficiently small fixed δ, holdout set size $n_* = \Theta(N^{2/3})$, and $s < 1$ of size $s = \Theta(N^{\epsilon-1/3})$ for some ϵ with $0 < \epsilon < 1/3$, we have for (strat) $\in \{(0), (n), (a)\}$:*

$$\lim_{N\to\infty}\left(\frac{\mathbb{E}\left\{C_{(h)}[0,T]\right\}}{\mathbb{E}\left\{C_{(o)}[0,T]\right\}}\right) = 1, \text{ and } \lim_{N\to\infty}\left(\frac{\mathbb{E}\left\{C_{(strat)}[0,T]\right\}}{\mathbb{E}\left\{C_{(o)}[0,T]\right\}}\right) > 1 \quad \left(= \Omega(\delta^{-2})\right) .$$

Proof. We begin with the following lemma:

Lemma S3. *For the T in assumption A1:*

$$I(T) \triangleq \sum_{i=0}^{T/\delta-1} \mathbb{E}_{t\sim U(i\delta,(i+1)\delta)}\left\{\xi_t^2(f_{i\delta}, f_t)\right\} > 0\,. \tag{26}$$

Proof. We employ assumptions A1- A3. As above, we denote α_2 as the Lipschitz constant of μ_t. We have $\int_0^T \xi_t^2(f_t, f_0)dt > 0$. There must be some $u < T$ with $j\delta < u < (j+1)\delta$ such that

$$d \triangleq \xi_u^2(f_u, f_{j\delta})dt > 0\,.$$

Let τ be a number satisfying

$$0 < \tau < \min\left(\frac{d}{\alpha_2 + 2\alpha_1}, \frac{u}{j\delta}\right) \tag{27}$$

so $d - (\alpha_2 + 2\alpha_1)\tau > 0$ and $u - \tau > j\delta$. We now have

$$\begin{aligned}
I(T) &\geq \mathbb{E}_{t\sim U(j\delta,(j+1)\delta)}\left\{\xi_t^2(f_{j\delta}, f_t)\right\} \\
&= \int_{j\delta}^{(j+1)\delta} \frac{1}{\delta}\xi_t^2(f_{j\delta}, f_t)dt \\
&\geq \frac{1}{\delta}\int_{u-\tau}^{u} \xi_t^2(f_{j\delta}, f_t)dt \\
&= \frac{1}{\delta}\int_{u-\tau}^{u}\int \left(f_t(x) - f_{j\delta}(x)\right)^2 d\mu_t dt \\
&\geq \frac{1}{\delta}\int_{u-\tau}^{u}\left(\int \left(|f_u(x) - f_{j\delta}(x)| - |f_t(x) - f_u(x)|\right)^2 d\mu_u - \alpha_2(u-t)\right)dt \\
&\geq \frac{1}{\delta}\int_{u-\tau}^{u}\left(\xi_t^2(f_u, f_{j\delta}) + \xi_u^2(f_t, f_u) - 2\sup_x(|f_t(x) - f_u(x)|)\right) - \alpha_2(u-t)dt \\
&\geq \frac{1}{\delta}\int_{u-\tau}^{u}\left(\xi_u^2(f_u, f_{j\delta}) - \alpha_2(u-t) - 2\alpha_1(u-t)\right)dt \\
&\geq \frac{1}{\delta}(d - 2\alpha_2\tau - \alpha_1\tau) \\
&> 0\,.
\end{aligned}$$

□

We firstly consider $C_{(h)}[0,T]$. We have, by Lemma S3, and recalling expressions 4 and 5:

$$\lim_{N\to\infty}\left(\frac{\mathbb{E}\left\{C_{(h)}[0,T]\right\}}{\mathbb{E}\left\{C_{(o)}[0,T]\right\}}\right) = \lim_{N\to\infty}\left(\frac{\sum_{i=0}^{T/\delta-1} C_{(h)}[i\delta,(i+1)\delta]}{\sum_{i=0}^{T/\delta-1} C_{(o)}[i\delta,(i+1)\delta]}\right)$$

$$
\begin{aligned}
&= \lim_{N\to\infty}\left(\frac{\sum_{i=0}^{T/\delta-1}\left(k_1 n_* + N\delta\left(p_* + (1-p_*)k_2\mathbb{E}_{t\sim U(i\delta,(i+1)\delta)}\left\{\xi_t^2(\rho_f^{n_*,i\delta,s}, f_t)\right\}\right)\right)}{\sum_{i=0}^{T/\delta-1}\left(N\delta k_2\mathbb{E}_{t\sim U(i\delta,(i+1)\delta)}\left\{\xi_t^2(\rho_f^{N,i\delta,s}, f_t)\right\}\right)}\right)\\
&= \lim_{N\to\infty}\left(\frac{\frac{T}{\delta}k_1\Theta\left(N^{\frac{1}{3}}\right) + N\delta k_2\sum_{i=0}^{T/\delta-1}\left(\mathbb{E}_{t\sim U(i\delta,(i+1)\delta)}\left\{\xi_t^2(f_{i\delta}, f_t)\right\}\right)}{N\delta k_2\sum_{i=0}^{T/\delta-1}\mathbb{E}_{t\sim U(i\delta,(i+1)\delta)}\left\{\xi_t^2(f_{i\delta}, f_t)\right\}}\right)\\
&= \lim_{N\to\infty}\left(\frac{\frac{T}{\delta}k_1\Theta\left(N^{\frac{1}{3}}\right) + N\delta k_2 I(T)}{N\delta k_2 I(T)}\right)\\
&= 1\,. \qquad (28)
\end{aligned}
$$

We next consider $C_{(n)}[0,T]$ and $C_{(a)}[0,T]$, recalling from Lemma S2 that we need only consider the latter. We note that

$$
\begin{aligned}
I(T) &= \sum_{i=0}^{T/\delta-1}\int_{i\delta}^{(i+1)\delta}\xi_t^2(f_{i\delta}, f_t)\frac{1}{\delta}dt\\
&\le \frac{1}{\delta}\sum_{i=0}^{T/\delta-1}\int_{i\delta}^{(i+1)\delta}\int(f_t - f_{i\delta})^2 d\mu_t dt\\
&\le \frac{1}{\delta}\sum_{i=0}^{T/\delta-1}\int_{i\delta}^{(i+1)\delta}\int(\alpha_1\delta)^2 d\mu_t dt\\
&= T\delta\alpha_1^2\,. \qquad (29)
\end{aligned}
$$

Recalling our definition of b as

$$
\lim_{N\to\infty}\left(\mathbb{E}_{t\sim U[0,T],D}\left\{\xi_t^2(\rho_t, f_t)\right\}\right) = b > 0\,,
$$

we choose any ϵ with $0 < \epsilon < b$, so for large enough N, we have for small enough $\delta < \frac{\sqrt{b}}{\alpha_1}$:

$$
\begin{aligned}
\lim_{N\to\infty}\left(\frac{\mathbb{E}\left\{C_{(a)}[0,T]\right\}}{\mathbb{E}\left\{C_{(o)}[0,T]\right\}}\right) &= \lim_{N\to\infty}\left(\frac{NTk_2\mathbb{E}_{t\sim U(0,T)}\left\{\xi_t^2(\rho_t, f_t)\right\}}{N\delta k_2 I(T)}\right)\\
&\ge \lim_{N\to\infty}\left(\frac{T(b-\epsilon)}{T\delta^2\alpha_1^2}\right)\\
&= \Omega(\delta^{-2})\\
&> 1\,.
\end{aligned}
$$

We next consider $C_{(n)}[0,T]$. We recall from expression 25 in the proof of Theorem 3.2 that for any $\epsilon > 0$ we may choose δ (dependent only on k_1, k_2, α_2, α_1) sufficiently small that for large enough N we have:

$$
\mathbb{E}\left\{C_{(n)}[(i-1)\delta, (i+1)\delta]\right\} \ge N\delta\left(c_{(n)} - \epsilon\right)\,,
$$

where $c_{(n)}$ does not depend on δ. Thus, for sufficiently small δ:

$$
\lim_{N\to\infty}\left(\frac{\mathbb{E}\left\{C_{(n)}[0,T]\right\}}{\mathbb{E}\left\{C_{(o)}[0,T]\right\}}\right) \ge \lim_{N\to\infty}\left(\frac{\frac{1}{2}\sum_{i=1}^{T/\delta-1}C_{(n)}[(i-1)\delta, (i+1)\delta]}{\sum_{i=0}^{T/\delta-1}C_{(o)}[i\delta, (i+1)\delta]}\right) \qquad (30)
$$

$$
\begin{aligned}
&\geq \lim_{N\to\infty} \left(\frac{\frac{1}{2} N\delta \left(\frac{T}{\delta} - 1\right) (c_{(n)} - \epsilon)}{N\delta k_2 I(T)} \right) \\
&\geq \lim_{N\to\infty} \left(\frac{(T-\delta)(c_{(n)} - \epsilon)}{2k_2 T\delta^2 \alpha_1^2} \right) \\
&= \Omega(\delta^{-2}) \\
&> 1
\end{aligned}
$$

as required.

□

S2.4 Relaxation of assumption A6

Our assumption A6 is prescriptive, necessitating a strict relationship between a particular conception of 'accuracy' of the risk score (as measured by ξ_t^2) and the amount by which we can reduce f_t in expectation (assumption A5). Assumption A6 allows some flexibility in the form of g_t: given f_t, an explicit example of a function $g_t(x,\rho)$ satisfying assumptions A5 and A6 can be constructed by considering constants $p_1, p_2 > 0$ and functions $q_1(x)$, $q_2(x)$ satisfying $\mathbb{E}_{X\sim\mu_t}\{q_1(X)\} = \mathbb{E}_{X\sim\mu_t}\{q_2(X)\} = 0$, and considering any form of $g_t(x,\rho)$ satisfying:

$$g_t(x,\rho) = \underbrace{f_t(x)}_{\text{T1}} - \underbrace{p_1 + q_1(x)}_{\text{T2}} + \underbrace{\xi_t^2(\rho, f_t)(p_2 + q_2(x))}_{\text{T3}},$$

where term T1 is the underlying risk (in the absence of a risk score), term T2 is the maximum potential reduction in risk (covariate-dependent, through $q_1(\cdot)$) and term T3 arises due to imperfect prediction: for instance, representing under-treatment or over-treatment due to $\rho(\cdot)$ differing from $f_t(\cdot)$, in a covariate-dependent manner through $q_2(\cdot)$. This can be seen to satisfy assumption A5 as:

$$\begin{aligned} k_1 &= \mathbb{E}_{X\sim\mu_t}\left\{f_t(X) - g_t(X, f_t)\right\} \\ &= \mathbb{E}_{X\sim\mu_t}\left\{p_1 - q_1(X) + \xi_t^2(f_t, f_t)(p_2 + q_2(x))\right\} \\ &= p_1 > 0\,, \end{aligned}$$

since $\xi_t^2(f_t, f_t) = 0$, and since:

$$\begin{aligned} c_t(\rho_t) &= \mathbb{E}_{X\sim\mu_t}\left\{g_t(X,\rho_t) - g_t(X, f_t)\right\} \\ &= \mathbb{E}_{X\sim\mu_t}\left\{\xi^2(\rho_t, f_t)\left(p_2 + q_2(X)\right) - \xi^2(f_t, f_t)\left(p_2 + q_2(X)\right)\right\} \\ &= \xi^2(\rho_t, f_t)\mathbb{E}_{X\sim\mu_t}\left\{p_2 + q_2(X)\right\} \\ &= p_2\xi^2(\rho_t, f_t)\,, \end{aligned}$$

the function g_t satisfies assumption A6 with $k_2 = p_2$.

We may substantially weaken assumption A6 and maintain our results, although the statement of Theorem 3.1 becomes a little more complex. We consider the alternative assumption (using the conventions u, l for 'upper' and 'lower'):

Assumption A6-alt. *Suppose a risk score ρ_t is in use at time t, and denote $\xi = \xi_t^2(\rho_t, f_t)$. We have*

$$k_2^l \xi^{b_l} \le c_t \le k_2^u \xi^{b_u}\,,$$

for some constants $k_2^l, k_2^u > 0$, and constants b_l, b_u satisfying $0 < b_u \le 1 \le b_l$.

along with a slight strengthening we will need for Theorem 3.3 (stating that cost must be determined by a function of ξ, not merely bounded)

Assumption A6-add. *With ρ_t, $\xi = \xi_t^2(\rho_t, f_t)$ as per assumption A6-alt and c_t as per assumption A5, we have $c_t = k_2(\xi)$, where k_2 is a continuous function.*

and note the following:

Remark 1. *Suppose we replace A6 with A6-alt in the statement of Theorems 3.1, 3.2 and 3.3. Then, replacing the statement of Theorem 3.1 with*

$$\mathbb{E}\left\{C_{(h)}[0,T]\right\} = \delta^{b_u} N T k_2^u (2\alpha_1 + \alpha_1^2)^{b_u} + \delta^{-1} O(N^a) + O(N^{1+b_u(a+\epsilon-1)}) + O(N^{1-b_u a})\,,$$

Theorems 3.1 and 3.2 still hold. If we additionally make assumption A6-add, then the statement of Theorem 3.3 still holds, with an asymptotic growth rate of $\Omega(\delta^{-2b_u})$ rather than $\Omega(\delta^{-2})$ for strategies (0), (a), (n).

Before proceeding to the proof, we note that in analogy to the main manuscript, we may specify an asymptotic optimal holdout size and corresponding cost in terms of N, although this time we must allow dependency on b_u.

For a fixed δ, a holdout set size of $n_* = \Theta\left(N^{\frac{1}{2}}\right)$ (regardless of b_u) will give an optimal asymptotic cost of

$$\mathbb{E}\left\{C_{(h)}[0,T]\right\} = \delta^{b_u} N k_2^u (2\alpha_1 + \alpha_1^2)^{b_u} + O\left(N^{1+\left(\epsilon-\frac{1}{2}\right)b_u}\right),$$

and if we choose

$$\delta = O\left(N^{\frac{1}{2+b_u}}\right), \quad n_* = \Theta\left(N^{\frac{1}{2+b_u}}\right), \quad s = \Theta\left(N^{\frac{1}{2+b_u}+\epsilon-1}\right)$$

we may achieve an optimal growth rate (for sufficiently small ϵ) of:

$$\mathbb{E}\left\{C_{(h)}[0,T]\right\} = O\left(N^{\frac{2}{2+b_u}}\right).$$

Proof. **Theorem 3.1**. For Theorem 3.1 we proceed in the same way until expression 5, where we instead have (employing the observations that $\mathbb{E}\{\xi^{b_u}\} \leq \mathbb{E}\{\xi\}^{b_u}$ and $(a+b)^{b_u} \leq a^{b_u} + b^{b_u}$ for $a, b, \xi > 0$, since $b_u \leq 1$)

$$\begin{aligned}
\mathbb{E}\left\{C_{(h)}[e-\delta,e]\right\} \leq & \underbrace{k_1 n_*}_{\substack{\text{Cost for held}\\ \text{-out samples}}} + \mathbb{E}\{n\}\Bigg(\underbrace{(1-p_*)k_1}_{\substack{\text{Cost if } < n_* \\ \text{samples in} \\ (e-1-s,e]}} + \underbrace{p_* \mathbb{E}\left\{k_2^u \left(\xi_t^2(\rho_f^{n_*,e,s}, f_t)\right)^{b_u}\right\}}_{\substack{\text{Cost for non-held-} \\ \text{out samples otherwise}}}\Bigg) \\
& \leq k_1 n_* + N\delta\left(O(N^{-2}) + k_2^u \mathbb{E}\left\{\xi_t^2(\rho_f^{n_*,e,s}, f_t)\right\}^{b_u}\right) \\
& \leq k_1 n_* + N\delta\left(O(n_*^{-1}) + (2\alpha_1 + \alpha_1^2)(\delta + s)\right)^{b_u} \\
& \leq k_1 n_* + N\delta O(n_*^{-1})^{b_u} + N\delta(2\alpha_1 + \alpha_1^2)_u^b(\delta^{b_u} + s^{b_u}) \\
& \leq \Theta(N^a) + \delta O(N^{1-b_u a}) + \delta\Theta(N s^{b_u}) + \delta^{1+b_u} N k_2^u (2\alpha_1 + \alpha_1^2) \\
& = O(N^a) + \delta O(N^{1-b_u a}) + \delta O(N^{1+b_u(a+\epsilon-1)}) + \delta^{1+b_u} N k_2^u (2\alpha_1 + \alpha_1^2), \qquad (31)
\end{aligned}$$

with a total cost per unit time accrued over the T/δ total epochs is thus

$$\mathbb{E}\left\{C_{(h)}[0,T]\right\} = \delta^{-1} O(N^a) + O(N^{1+b_u(a+\epsilon-1)}) + O(N^{1-b_u a}) + \delta^{b_u} N T k_2^u (2\alpha_1 + \alpha_1^2)^{b_u}).$$

Theorem 3.2. To establish Theorem 3.2 for (strat) $\in \{(0), (a)\}$ we pick up at equation 14 to instead note

$$\begin{aligned}
\mathbb{E}\left\{C_{(a)}[0,T]\right\} & \geq \mathbb{E}\{\text{Pois}(NT)\} \cdot \mathbb{E}_{t\sim U(0,T),D}\left\{\mathbb{E}\left\{k_2^l \left(\xi_t^2(\rho_t, f_t)\right)^{b_l}\right\}\right\} \\
& \geq NT k_2^l \mathbb{E}\left\{\xi_t^2(\rho_t, f_t)\right\}^{b_l} \\
& \geq NT k_2^l (b-\epsilon)^{b_l} \\
& = \Omega(N).
\end{aligned}$$

For (strat) $= (n)$ we begin by rephrasing equation 19 as

$$k_2^u \xi_t^2 \left(\rho_g^{N,(e-1)\delta,s}, f_t\right)^{b_u} \geq k_1 - \mathbb{E}_{X\sim\mu_t}\left\{f_t(X) - g_t(X)\right\} \geq k_2^l \xi_t^2 \left(\rho_g^{N,(e-1)\delta,s}, f_t\right)^{b_l} .$$

We must rework case 1; we now have, by assumption:

$$\begin{aligned} & k_1 - k_2^u \xi_u^2 \left(\rho_g^{N,(e-1)\delta,s}, f_u\right)^{b_u} < 0 \\ \Leftrightarrow \quad & \xi_\tau^2 \left(\rho_g^{N,(e-1)\delta,s}, f_u\right) > \left(\frac{k_1}{k_2^u}\right)^{\frac{1}{b_u}} , \end{aligned}$$

so for any $t \in [(i-1)\delta, i\delta)$:

$$\begin{aligned} \xi_t^2 \left(\rho_g^{N,(e-1)\delta,s}, f_t\right) &\geq \xi_t^2 \left(\rho_g^{N,(e-1)\delta,s}, f_\tau\right) - 2\sqrt{\xi_t^2\left(f_t, f_\tau\right)} \\ &\geq \xi_u^2 \left(\rho_g^{N,(e-1)\delta,s}, f_\tau\right) - \alpha_2|t-\tau| - 2\alpha_1|t-\tau| \\ &\geq \left(\frac{k_1}{k_2^u}\right)^{\frac{1}{b_u}} - (\alpha_2 + 2\alpha_1)\delta , \end{aligned}$$

and the total expected cost accrued over the time period $[(i-1)\delta, (i+1)\delta]$ (during which we encounter Pois$(2N\delta)$ samples) is:

$$\begin{aligned} \mathbb{E}\left\{C_{(n)}[(i-1)\delta, (i+1)\delta]\right\} &\geq \mathbb{E}\left\{C_{(n)}[(i-1)\delta, i\delta]\right\} \\ &\geq \mathbb{E}\left\{\text{Pois}(N\delta)\right\} k_2^l \mathbb{E}_{t\sim U((i-1)\delta, i\delta)} \left\{\xi_t^2 \left(\rho_g^{N,(e-1)\delta,s}, f_t\right)^{b_l}\right\} \\ &\geq N\delta k_2^l \left(\left(\frac{k_1}{k_2^u}\right)^{\frac{1}{b_u}} - (\alpha_2 + 2\alpha_1)\delta\right)^{b_l} , \end{aligned}$$

hence we may choose δ dependent only on α_2, α_1, k_2 sufficiently small that

$$\mathbb{E}\left\{C_{(n)}[(i-1)\delta, (i+1)\delta]\right\} \geq N\delta m_1 \tag{32}$$

for some positive constant m_1.

For case 2, we again appeal to expression 18 to rewrite expression 23 (using the fact that $\mathbb{E}\{\xi^{b_l}\} \geq \mathbb{E}\{\xi\}^{b_l}$ since $b_l > 0$)

$$\begin{aligned} \mathbb{E}\left\{C_{(n)}[(i-1)\delta, i\delta]\right\} &\geq \mathbb{E}\left\{\text{Pois}(N\delta)\right\} k_2^l \mathbb{E}_{t\sim U((i-1)\delta, i\delta)} \left\{\xi_t^2 \left(\rho_g^{N,(i-1)\delta,s}, f_t\right)^{b_l}\right\} \\ &\geq N\delta k_2^l \mathbb{E}_{t\sim U((i-1)\delta, i\delta)} \left\{\xi_t^2 \left(\rho_g^{N,(i-1)\delta,s}, f_t\right)\right\}^{b_l} \\ &\geq N\delta k_2^l \left(\Delta_0 - \delta(\alpha_2 + 2\alpha_1) - \alpha_2 s\right)^{b_l} \\ &\geq N\delta k_2^l \left(\Delta_0 - m_s\right)^{b_l} . \end{aligned}$$

We may recalculate the final three lines of derivation 22 as:

$$\xi_{i\delta}\left(f_{i\delta}, G\right) \geq \left(\frac{1}{s}\int_{i\delta-s}^{i\delta} \max\left(0, \mathbb{E}_{X\sim\mu_t}\left\{f_t(x) - g_t(x)\right\} - \alpha_1\delta\right)\right)^2$$

$$\geq \left(\frac{1}{s} \int_{i\delta - s}^{i\delta} \max\left(0, k_1 - k_2^u \xi_t^2 \left(\rho_g^{N,(e-1)\delta,s}, f_t\right)^{b_u} - \alpha_1 \delta\right)\right)^2$$
$$\geq \max\left(0, k_1 - k_2^u \left(\Delta_0 + (\alpha_2 + 2\alpha_1)\delta + \alpha_1^2\delta^2 + \alpha_2 s\right)^{b_u} - \alpha_1\delta\right)^2 \tag{33}$$

and hence rewrite expression 24 as

$$\begin{aligned}
\mathbb{E}\left\{C_{(n)}[i\delta,(i+1)\delta]\right\} &\geq \mathbb{E}\left\{\text{Pois}(N\delta)\right\} k_2^l \mathbb{E}_{t\sim U(i\delta,(i+1)\delta)} \left\{\xi_t^2 \left(\rho_g^{N,i\delta,s}, f_t\right)^{b_l}\right\} \\
&\geq N\delta k_2^l \mathbb{E}_{t\sim U(i\delta,(i+1)\delta)} \left\{\xi_t^2 \left(\rho_g^{N,i\delta,s}, f_t\right)\right\}^{b_l} \\
&\geq N\delta k_2^l \left(k_1 - k_2^u(\Delta_0 - m_s)^{b_u} - m_s\right)^{2b_l} ,
\end{aligned}$$

and hence the total cost over both periods is

$$\mathbb{E}\left\{C_{(n)}[(i-1)\delta,(i+1)\delta]\right\} \geq N\delta k_2 \left((\Delta_0 - m_s)^{b_l} + (k_1 - k_2^u(\Delta_0 - m_s)^{b_u} - m_s)^{2b_l}\right) .$$

Analogous to the proof of Theorem 3.2, given ϵ, we may choose δ (dependent only on k_1, k_2, α_2, α_1) sufficiently small that for large enough N:

$$\mathbb{E}\left\{C_{(n)}[(i-1)\delta,(i+1)\delta]\right\} \geq N\delta \left(k_2 \min_{0\leq\Delta_0\leq 1} \left(\Delta_0^{b_l} + (k_1 - k_2^u \Delta_0^{b_u})^{2b_l}\right) - \epsilon\right) .$$

Recalling expression 32 for the earlier case, we denote

$$c_{(n)} = \min\left(k_2^l \min_{0\leq\Delta_0\leq 1} \left(\Delta_0^{b_l} + (k_1 - k_2^u \Delta_0^{b_u})^{2b_l}\right), k_1\right) > 0 , \tag{34}$$

so, in either case

$$\mathbb{E}\left\{C_{(n)}[(i-1)\delta,(i+1)\delta]\right\} \geq N\delta \left(c_{(n)} - \epsilon\right) ,$$

and the proof proceeds as above.

Theorem 3.3. For Theorem 3.3, we must additionally make assumption A6-add, since without it there is no guarantee that costs of various strategies will be similar even if the associated risk scores are equally similar to f_t.

We require a slightly modified form of Lemma S3; namely that for sufficiently large T we have

$$I(T) \triangleq \sum_{i=0}^{T/\delta-1} \mathbb{E}_{t\sim U(i\delta,(i+1)\delta)} \left\{k_2\left(\xi_t^2(f_{i\delta}, f_t)\right)\right\} > 0 ,$$

which can proved in essentially the same way as the original lemma by first noting that since $k_2(\xi) \geq k_2^l \xi^{b_l}$, we have

$$I_2(T) \geq k_2^l \sum_{i=0}^{T/\delta-1} \mathbb{E}_{t\sim U(i\delta,(i+1)\delta)} \left\{\xi_t^2(f_{i\delta}, f_t)^{b_l}\right\} .$$

With this, we may trudge back to derivation 28, delete any occurrences of the *constant* k_2, replace $I(\cdot)$ with $I_2(\cdot)$, and replace any instances of $\xi_t^2(f_{i\delta}, f_t)$ with $k_2\left(\xi_t^2(f_{i\delta}, f_t)\right)$, and see that the derivation holds.

Given that $k_2(\xi) \leq k_2^u \xi^{b_u}$ we may amend expression 29 to note:

$$I_2(T) \leq T k_2^u \alpha_1^{2b_u} \delta^{2b_u - 1}\,,$$

so for sufficiently small δ:

$$\begin{aligned}
\lim_{N\to\infty}\left(\frac{\mathbb{E}\left\{C_{(a)}[0,T]\right\}}{\mathbb{E}\left\{C_{(o)}[0,T]\right\}}\right) &= \lim_{N\to\infty}\left(\frac{NTk_2\mathbb{E}_{t\sim U(0,T)}\left\{k_2\left(\xi_t^2(\rho_t,f_t)\right)\right\}}{N\delta I_2(T)}\right)\\
&\geq \lim_{N\to\infty}\left(\frac{Tk_2^l\mathbb{E}_{t\sim U(0,T)}\left\{\xi_t^2(\rho_t,f_t)^{b_l}\right\}}{T\delta k_2^u\alpha_1^{b_u}\delta^{2b_u-1}}\right)\\
&\geq \lim_{N\to\infty}\left(\frac{k_2^l\mathbb{E}_{t\sim U(0,T)}\left\{\xi_t^2(\rho_t,f_t)\right\}^{b_l}}{k_2^u\alpha_1^{b_u}\delta^{2b_u}}\right)\\
&\geq \lim_{N\to\infty}\left(\frac{k_2^l(b-\epsilon)^{b_l}}{k_2^u\delta^{2b_u}\alpha_1^{2b_u}}\right)\\
&= \Omega(\delta^{-2b_u})\\
&> 1\,.
\end{aligned}$$

For (strat) = (n), allowing for the new definition of $c_{(n)}$ in expression 34, we have (picking up at derivation 30)

$$\begin{aligned}
&\lim_{N\to\infty}\left(\frac{\mathbb{E}\left\{C_{(n)}[0,T]\right\}}{\mathbb{E}\left\{C_{(o)}[0,T]\right\}}\right) \geq \lim_{N\to\infty}\left(\frac{\frac{1}{2}\sum_{i=1}^{T/\delta-1}C_{(n)}[(i-1)\delta,(i+1)\delta]}{\sum_{i=0}^{T/\delta-1}C_{(o)}[i\delta,(i+1)\delta]}\right)\\
&\geq \lim_{N\to\infty}\left(\frac{\frac{1}{2}N\delta\left(\frac{T}{\delta}-1\right)(c_{(n)}-\epsilon)}{N\delta I_2(T)}\right)\\
&\geq \lim_{N\to\infty}\left(\frac{(T-\delta)(c_{(n)}-\epsilon)}{2k_2^uT\delta^{2b_u}\alpha_1^{2b_u}}\right)\\
&= \Omega(\delta^{-2b_u})\\
&> 1\,,
\end{aligned}$$

as required. □

S2.5 Updating in the absence of drift

We briefly consider performance of updating strategies (section 2.1) when no drift occurs. In some such settings researchers may still want to update risk scores: for instance, in order to make use of a new or better training algorithm. Essentially we claim that *either* the risk score should not be updated, or a holdout set should be used to update it, even in this case: the naive-update strategy will still lead to suboptial costs.

We firstly consider a fixed time period $[0, T]$ for which there is no $T_1 \leq T$ for which assumption A1 holds. We note that the proof of Theorem 3.2 for $(\text{strat}) = (n)$ makes no use of assumption A1, so if naive updating is used, cost accrues at a rate $\Omega(N)$. Essentially, the costs accrued due to use of naive updating arise from the tension that a difference between f_t and g_t both indicates that the risk score is a 'success' in reducing incidence of adverse events, but means that it is inevitably inaccurate once updated.

The upper bound in Theorem 3.1 similarly does not require assumption A1, so we may attain $O(N)$ cost for fixed δ or $O(N^{2/3})$ cost if we may choose δ. The costs accrued using the no-update strategy ($(\text{strat}) = (0)$) are 0 if we assume full knowledge of f_0, since $f_t = f_0$ almost everywhere with respect to μ_t for all $t \leq T$.

We now turn our attention to an asymptotic regime with an infinite time horizon. Since $\sup_{x,\rho} |f_t(x) - g_t(\rho, x)| \leq 1$, we have, for fixed N:

$$\int_0^T \xi_t^2(f_t f_0)dt = O(T)\,.$$

If we had:

$$\int_0^T \xi_t^2(f_t f_0)dt = o(T)\,,$$

then the mean deviation of f_t from f_0 would converge to 0 as $T \to \infty$, as would the time-averaged total cost if the no-update strategy is used with fixed N:

$$\frac{1}{T}C_{(0)}[0, T] = \frac{1}{T}k_2 \int_0^T \xi_t^2(f_t, f_0)dt = \frac{o(T)}{T} \to 0\,.$$

If either the naive-update or holdout-set update strategies are followed with a fixed update frequency δ, the time-averaged cost as above will not converge to 0, since a new cost is added each time the model is updated. We note that over any finite time horizon over which some drift occurs, our usual formulations of Theorems 3.1, 3.2 and 3.3 govern the relative costs of each strategy.

If we have

$$\int_0^T \xi_t^2(f_t f_0)dt = \Theta(T)\,,$$

then for almost all finite time horizons T, assumption A1 will hold, and Theorems 3.1, 3.2 and 3.3 describe the performance of each strategy.

S3 Notes on ethics and alternative options

S3.1 Non-identifiability of f_t from g_t and ρ_t

As in the main paper, suppose that we are a researcher at time t, at which a risk score ρ_t is in place, and we have at hand *only* a set of sample covariate values x and values $g_t(x)$ (or rather, samples from $\text{Bern}(g_t(x, \rho_t))$). It is immediate that we cannot infer f_t from this: we do not directly observe $f_t(x)$ or values $\text{Bern}(f_t(x))$, and without further assumptions on the relationship between f_t and g_t, we cannot judge what $f_t(x)$ is.

Intuitively, however, we may expect that f_t will be similar to g_t if we have little confidence in the risk score (and hence take little risk-score guided action) and f_t will be different from g_t if we are more confident that the risk score resembles ρ_t. These intuitions are quantified by assumptions A5 and A6, in that we have (all expectations over $X \sim \mu_t$):

$$
\begin{aligned}
c_t(\rho_t) &= k_2 \xi_t^2(f_t, \rho_t) && \text{Assumption A6} \\
&= \mathbb{E}\left\{g_t(X, \rho_t) - g_t(X, f_t)\right\} && \text{Assumption A5} \\
&= \mathbb{E}\left\{g_t(X, \rho_t) - f_t(X) + f_t(X) - g_t(X, f_t)\right\} \\
&= \mathbb{E}\left\{g_t(X, \rho_t) - f_t(X)\right\} + \mathbb{E}\left\{f_t(X) - g_t(X, f_t)\right\} \\
&= \mathbb{E}\left\{g_t(X, \rho_t) - f_t(X)\right\} + k_1 && \text{Assumption A5} \\
&= k_1 - \mathbb{E}\left\{f_t(X) - g_t(X, \rho_t)\right\}
\end{aligned}
$$

Therefore, as stated in the main paper,

$$
k_2 \xi_t^2(f_t, \rho_t) = k_1 - \mathbb{E}_{X \sim \mu_t}\{f_t(X) - g_t(X, \rho_t)\}\,.
$$

We show in this section that this is (unsurprisingly) not enough to uniquely identify f_t. Our construction is elementary, but somewhat laborious. Let $h_t(x)$ be a function and consider the function $f_t^h(x) = f_t(x) + h(x) - h$ as follows:

$$
\begin{aligned}
&k_2 \xi(f_t^h, \rho_t) - \Big(k_1 - \mathbb{E}\{f_t^h(X) - g_t(X, \rho_t)\}\Big) \\
&= k_2 \mathbb{E}\{(f_t(X) + (h_t(X) - h) - \rho_t(X))^2\} - (k_1 - \mathbb{E}\{f_t(X) + h_t(X) - h - g_t(X, \rho_t)\}) \\
&= k_2 \xi_t^2(f_t, \rho_t) + k_2 \mathbb{E}\{(h_t(X) - h)(f_t(X) - \rho_t(X))\} + 2k_2 \mathbb{E}\{(h_t(X) - h)^2\} \\
&\quad - (k_1 - \mathbb{E}\{f_t(X) - g_t(X, \rho_t)\}) + \mathbb{E}\{h_t(X) - h\} \\
&= \mathbb{E}\{(h_t(X) - h)(k_2(f_t(X) - \rho_t(X) + h_t(X) - h) + 1)\} \\
&= \mathbb{E}\{(h_t(X) - h)(k_2(f_t(X) - \rho_t(X) + h_t(X) - h) + 1)\} \\
&= Ah^2 + \mathbb{E}\{B\}h + \mathbb{E}\{C\}\,,
\end{aligned} \tag{35}
$$

where

$$
\begin{aligned}
A &= k_2 \\
B &= -\left(2k_2(f_t(X) + h_t(X) - \rho_t(X)) + 1\right) \\
C &= \left(k_2\left(h_t(X)^2 + 2h_t(X)(f_t(X) - \rho_t(X))\right) + h_t(X)\right)\,,
\end{aligned}
$$

and all expectations are over $X \sim \mu_t$. We note that

$$
\mathbb{E}\{B^2 - 4AC\} = \mathbb{E}\{B^2\} - 4A\mathbb{E}\{C\} = \left(2k_2(f_t(X) - \rho_t(X)) + 1\right)^2 > 0\,. \tag{36}
$$

If we choose $h_t(X)$ close to $\rho_t(X) - f_t(X)$, we can ensure $\mathrm{var}(B)$ is small enough that

$$\mathrm{var}(B) = \mathbb{E}\{B^2\} - \mathbb{E}\{B\}^2 < \mathbb{E}\{B^2 - 4AC\}\,,$$

and hence $\mathbb{E}\{B\}^2 - 4A\{C\} > 0$ and equation 35 has a root h (call it h_t^0). Thus, for *arbitrary* $h_t(X)$ close to $\rho_t(X) - f_t(X)$, we can find an h_t^0 such that $f_t^h(x) = f_t(x) + h_t(x) - h_t^0$ also satisfies

$$k_2 \xi_t^2(f_t^h, \rho_t) = k_1 - \mathbb{E}_{X\sim\mu_t}\{f_t^h(X) - g_t(X)\}\,,$$

so f_t is not distinguishable from f_t^h, even with knowledge of k_1, k_2, and hence is not identifiable from g_t, μ_t, ρ_t alone.

S3.2 Use of natural hold-out sets, recorded interventions, and maximum-of-two updating

Ethical objections to the use of holdout sets are in a sense due to the need to *actively* with-hold some samples from access to a risk score. A brief discussion is merited on potential use of 'natural' hold-out sets, in which some proportion of a population may be assumed to behave like a holdout set without actively requiring withholding of scores. We will use our motivating example (section 1.1) to illustrate this idea.

An obvious setting in which a natural holdout set would be appropriate is in the case where an effectively random subset of samples already do not have access to a risk score. In our example, this may comprise a set of individuals under the care of a medical practitioner who chose not to use the ASPRE score. As long as the distribution of covariates of such individuals is typical of the population distribution (that is, μ_t is conserved) and the behaviour of such practitioners is typical of the general behaviour of practitioners in the absence of a risk score (that is, f_t is conserved) then such a subset of individuals can be used as a holdout set.

S3.2.1 Recorded interventions

Risk scores are generally used to simplify information in complex settings, including medicine or finance. This complexity is generally also present in the range and effect of intervention: it is difficult to determine the optimal intervention in a given context. We generally wish to make use of expertise of domain experts (for instance, doctors or financial analysts) in making interventions, and hence do not generally consider direct recommendations of intervention to be practical in this circumstance. Moreover, even *recording* of interventions may be difficult: a doctor responding to a high disease risk may see a patient more often, reduce thresholds for further investigation, or consider new treatments, all of which may be difficult to record.

However, we consider here options for a setting in which the intervention is simple and recorded (for the case of PRE, whether an individual was given aspirin). The set of individuals for whom no intervention was recorded (that is, did not get prescribed aspirin) do not constitute a 'natural' holdout set in our sense: the function f_t measures the probability of an event *under normal care without a risk score*, and patients without a risk score may still get prescribed aspirin.

Suppose we are working at a fixed time t (that is, disregard dependence on time) and denote by X a set of covariates included in a risk score, A an indicator for whether aspirin was prescribed or not, L a 'latent' covariate (taking values in finite set $\mathscr{L}$ for simplicity) representing patient characteristics visible to a medical practitioner but which are not amongst covariates X, $Y \in \{0, 1\}$ the event in question (in this case, the incidence of PRE) and G the event of whether the practitioner had access to a risk score ($G = 1$ for yes, $G = 0$ for no).

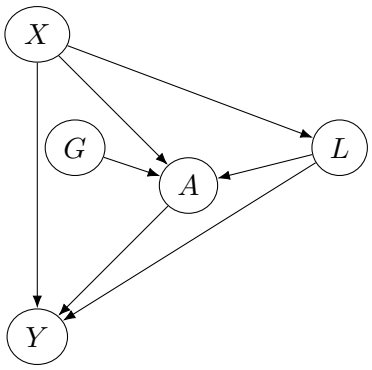
We firstly note that recording A does not necessarily help in meaningfully 'updating' the risk score, essentially because the same considerations as in Figure 1 continue to hold. In particular, the treatment decision $P(A|X)$ depends on the risk score currently in place. Should we estimate $\rho_{e+1} = P(Y|A, X)$ while a risk score ρ_e is in use, and replace ρ_e with ρ_{e+1}, then ρ_{e+1} will now not generally estimate $P(Y|A, X)$, since $P(A|X)$ has changed.

It is worthwhile to consider a more general estimation problem. Faced with the problem of whether to allocate treatment to a particular patient, given covariates x, we ideally wish to estimate the counterfactuals

$$P_{A\leftarrow 0}(Y|X) \quad \text{and} \quad P_{A\leftarrow 1}(Y|X)\,; \tag{37}$$

that is, the probability of the outcome on a patient with $X = x$ if we forcibly assign a treatment $A = 0$ or $A = 1$. The counterfactual quantities are computed assuming that random variables

X, G, Y, A, L have the following causal structure:



in which L are partly dependent on X, the decision A to prescribe aspirin is made based on X and L with the decision rule modulated by G, and Y is dependent on A, X and L. To reconcile L with our earlier notation we have

$$f_t(x) = P(Y = 1|X = x, G = 0) = \sum_{\lambda \in \{0,1\}} P(Y = 1|X = x, G = 0, L = \lambda)P(L = \lambda|X = x)$$
$$g_t(x) = P(Y = 1|X = x, G = 1) = \sum_{\lambda \in \{0,1\}} P(Y = 1|X = x, G = 1, L = \lambda)P(L = \lambda|X = x)\,,$$

taking L and G as conditionally independent given X.

The data we may use to estimate these counterfactuals are direct estimates of the conditionals

$$P(Y|X = x, A = 1) \quad \text{and} \quad P(Y|X = x, A = 1)\,, \tag{38}$$

whose values are dependent on the risk score through $P(A|X = x)$. If we make a *deterministic* decision on T on the basis of x, so $P(T|X) \in \{0, 1\}$, we can estimate one of the values 38 and it is equal to the corresponding counterfactual 37, but we lose all ability to estimate the other.

In more generality, we consider two circumstances:

1. We make a deterministic decision on the basis of X, L. This leaves us in much the same circumstance: faced with a patient with $(X, L) = (x, \lambda)$, we can estimate only one of $P_{A\leftarrow 1}(Y|X = x, L = \lambda)$ or $P_{A\leftarrow 0}(Y|X = x, L = \lambda)$.

2. We have a degree of randomness, so some patients with $X = x, L = \lambda$ are treated and some are not.

We have problems in both cases. In the first, we know only one of the counterfactuals 37, so we have only information on what happens if we take a specific, forced course of action. We claim that knowledge of f_t: the probability of the outcome *under normal clinical care without a risk score* is more useful.

In the second case, we claim that:

- Because we do not measure L, we cannot readily integrate over it, and cannot directly estimate counterfactuals 37;

- Since estimation of 37 would therefore generally require modelling assumptions, any estimates would not generally be consistent, whereas estimates of f_t made from a holdout set could be reasonably expected to be consistent;

- Recalling that L represents everything that a clinician can see, this scenario effectively constitutes randomising patients to a treatment. If this were done *consciously and independently of* L for some set of patients, then we avoid the previous two problems - but we are left with something even worse than a holdout set, in that we 'hold-out' a set of patients only to force them to receive one treatment or another!

In summary, we claim that recording interventions may enable estimation of counterfactual quantities 37, but that this is not straightforward, that estimates would be unlikely to be consistent, and that this may be ethically little better than use of a holdout set.

In general, the choice of updating method is essentially a consideration of the lesser of several evils. The case where interventions can be recorded is similar in that no option is clearly best, although there are several more options, and use of a holdout set may still be indicated. We do not consider it in our analysis of the ASPRE score, but it is a potential avenue for future work.

S3.3 Maximum-of-two updating

In a parallel applied work Liley et al. (2024) we recognised the potential problems with naive updating as applied to a real-world risk score. The score in question predicted yearly individual risk of emergency hospital admission for the majority of the Scottish population, with risk scores deployed monthly to Scottish general practitioners for the patients in their care. We developed the fourth version of the risk score, the third version being currently in use.

In the absence of a holdout set (and the lack of precedent) we chose to first re-fit the risk score naively: in the notation of section 3, taking $t = 1, 2, 3, 4$ as the times when the first through fourth versions of the risk score were fitted, if ρ_3 is the third version of the risk score, our naive estimate ρ'_4 was

$$\rho'_4(x) \approx g_4(x, \rho_3)\,.$$

We then proposed to designate the updated risk score ρ_4 as:

$$\rho_4 = \max\left(\rho_3(x), \rho'_4(x)\right)\,.$$

Our intent in doing so was to try and ensure that $\rho_4(x) \geq f_4(x)$, given that we could not aim for $\rho_4(x) = f_4(x)$. We consider that this was the best option available to us at the time. However, this approach has many drawbacks compared to a holdout set approach, and we do not consider that it is viable in general or in the long term. Our reasons for this are as follows:

1. Overestimation of risk, while possibly less costly than underestimation of risk, is not without cost. In practical terms, overestimation of risk could lead to overtreatment, or diversion of care away from patients in greater need.

2. There is no clear extension of this approach to updating the model a second time (say, to ρ_5). Given $\rho'_5(x) \approx g_t(x, \rho_4)$, we could take

   $$\rho_5(x) = \max\left(\rho_4, \rho'_5\right) = \max\left(\rho_3, \rho'_4, \rho'_5\right)\,,$$

   but this will lead to more severe overestimation of f_5, ultimately lowering confidence in the risk score as above. It is also more susceptible to drift than our $\rho_3 \to \rho_4$ update, since we are relying on an estimate made two epochs ago rather than just one. This is not avoided if we take $\rho_5(x) = \max\left(\rho'_4, \rho'_5\right)$, since ρ_3 has a causal influence on ρ'_4 as per Figure 1.

3. As per Section 4, should we use a holdout set, we would generally try and ensure the holdout set is as close as possible to $t = 4$, and thus reduce the effect of drift between $t = 3$ and $t = 4$. Our approach in this case is more susceptible to drift, because we use the risk score ρ_3 directly in our estimate of ρ_4.

S4 Proofs of Theorems 4.1, 4.2, and 4.3

Theorem 4.1. *Suppose assumptions B1-B4 in the main manuscript hold. Then there exists a* $N_* \in (0, N)$ *with* $N \in \mathbb{N}$*, which we call the optimal holdout set size, such that:*

$$\ell(i) \geq \ell(j) \text{ for } 0 < i < j < N_*$$
$$\ell(i) \leq \ell(j) \text{ for } N_* < i < j < N\,.$$

Proof. As discussed in the main manuscript, we may impose that

$$\frac{\partial}{\partial n} k_2(n) < 0\,. \tag{39}$$

Since both $k_2(n)$ and $(N-n)$ are positive and monotonically decreasing in n, so is $k_2(n)(N-n)$. Now

$$\begin{aligned}
\ell'(n) &= \frac{\partial}{\partial n}\left(k_1 n + k_2\left(n\right)\left(N-n\right)\right) && (40)\\
&= k_1 + k_2'\left(n\right)\left(N-n\right) - k_2\left(n\right) && (41)\\
&= \left(k_1 - k_2\left(n\right)\right) + k_2'\left(n\right)\left(N-n\right)\,. && (42)
\end{aligned}$$

By assumption B3 in the main manuscript, $k_1 < k_2(0)$, and, from equation 39, $k_2'(0) < 0$, so both terms in equation 42 are negative when $n = 0$ and $\ell'(0) < 0$. When $n = N$, the second term vanishes while the first one is positive, as $k_1 > k_2(N)$ by assumption B3 in the main manuscript. We thus have $\ell'(N) > 0$. By assumption, ℓ is smooth, so by Bolzano's Theorem, there must exist at least one point n_* for which $\ell'\left(n_*\right) = 0$, which is an extremum of ℓ.

We now prove that this extremum is unique and a minimum. First, by assumption B4 in the main manuscript, we may impose that

$$\frac{\partial^2}{\partial n^2} k_2(n) > 0\,. \tag{43}$$

Taking the second derivative of ℓ:

$$\begin{aligned}
\frac{\partial^2}{\partial n^2}\ell(n) &= \frac{\partial^2}{\partial n^2}\left(k_1 n + k_2(n)(N-n)\right) && (44)\\
&= k_2''(n)(N-n) - 2k_2'(n)\,; && (45)
\end{aligned}$$

and using equations 39 and 43, we see that $\ell''(n)$ is strictly positive, and, as a consequence, $\ell'(n)$ is monotonically increasing. Therefore, the extremum of $\ell(n)$ at n_* we found earlier is unique and, as $\ell''(n) > 0$, it is a minimum.

If $n_* \in 1..(N-1)$, let $N_* = n_*$. If $n_* \notin \mathbb{N}$, let N_* be the closest natural number to either side of n_*. From assumption B3 in the main manuscript, N_* cannot be 0 or N. In both scenarios, this completes the proof.

□

For the proofs of Theorems 4.2 and 4.3, we will use the following lemma:

Lemma S4. *Suppose assumption 1 holds and there exists* $0 < M < N$ *such that* $k_2(M) < k_1$*. Then* $\ell(M) < \ell(N)$*.*

Proof.

$$\begin{aligned} k_2(M) < k_1 &\implies (N-M)k_2(M) < (N-M)k_1 \\ &\implies (N-M)k_2(M) + Mk_1 < Nk_1 \\ &\implies \ell(M) < \ell(N)\,. \end{aligned}$$

□

We now have

Theorem 4.2. *Suppose assumptions B1 and B5 hold, and $k_2(0) \leq k_1$. Then there exists an $N_* \in \{1, \ldots, N-1\}$ such that: $\ell(i) \geq \ell(N_*)$ for $i \in \{1, \ldots, N-1\}$ and $\ell(i) > \ell(N_*)$ for $i \in \{0, N\}$*

Proof. All that is needed to show that there exists a holdout set size M where $\ell(M) < \ell(0)$ and $\ell(M) < \ell(N)$. This will be the M in assumption B5.

It immediately follows from assumption B5 that $k_2(M) < k_1$, so by Lemma S4 $\ell(M) < \ell(N)$. If $k_1 = k_2(0)$ then we are done as $\ell(N) = \ell(0)$. If $k_1 > k_2(0)$ then from assumption B5 we have

$$\begin{aligned} k_1 - k_2(0) < \frac{N-M}{N}(k_1 - k_2(M)) &\implies N(k_1 - k_2(0)) < (N-M)(k_1 - k_2(M)) \\ &\implies Nk_1 - Nk_2(0) < Nk_1 - (N-M)k_2(M) - Mk_1 \\ &\implies \ell(N) - \ell(0) < \ell(N) - \ell(M) \\ &\implies \ell(M) < \ell(0)\,, \end{aligned}$$

as needed. □

Theorem 4.3. *Suppose assumption B1 holds, $k_1 < k_2(0)$ and there exists $0 < M < N$ such that $k_2(M) < k_1$. Then there exists an $N_* \in \{1, \ldots, N-1\}$ such that: $\ell(i) \geq \ell(N_*)$ for $i \in \{1, \ldots, N-1\}$ and $\ell(i) > \ell(N_*)$ for $i \in \{0, N\}$.*

Proof. Note

$$k_1 < k_2(0) \implies Nk_1 < Nk_2(0) \implies \ell(N) < \ell(0)$$

so all that is needed is $\ell(M) < \ell(N)$, which is given by Lemma S4. □

S5 Estimation of $k_2(n)$

We argue in this section that $k_2(n)$ can generally be a modelled as a linear function of expected mean squared error of the risk score. Suppose that we are interested in making predictions at a time t. We consider a risk score $\rho(X)$ which is an inexact approximation of $f_t(X)$ (recalling the definition of f_t from section 3 as the probability of $Y = 1$ given X at time t when no risk score is used). We will write $c_2(x, \rho) = \mathbb{E}_{C_2}\{C_2(x; d_n)\}$, where $\mathbb{E}_{C_2}$ indicates expectation over randomness in actual cost for a given sample with a given risk score, and ρ a risk score fitted to samples d_n.

If it is reasonable to assume that $c_2(x, \rho)$ has a straightforward form in terms of ρ, f_t, then a corresponding form for k_2 may be immediate. For instance, if one of

$$\begin{aligned} c_2(x, \rho) &= c^0 + c^1|\rho - f_t(x)| \\ c_2(x, \rho) &= c^0 + c^1(\rho - f_t(x))^2 \\ c_2(x, \rho) &= c^0 + c^1 f_t(x)(\rho - f_t(x))^2 \end{aligned}$$

holds, for some constants c^0, c^1, then k_2 will be linear in

$$\begin{aligned} \mathbb{E}_{D_n}\left\{\mathbb{E}_X\left[|\rho(X) - f_t(X)|\right]\right\} \\ \mathbb{E}_{D_n}\left\{\mathbb{E}_X\left[(\rho(X) - f_t(X))^2\right]\right\} \\ \mathbb{E}_{D_n}\left\{\mathbb{E}_X\left[f_t(X)(\rho(X) - f_t(X))^2\right]\right\} \end{aligned} \tag{46}$$

respectively. As discussed in the main manuscript, this reduces the estimation of $k_2(n)$ to estimating the 'learning curve' of a risk score, and expectations of risk score accuracy measures such as those above over X and D_n can be readily estimated for small n given training samples X, Y.

In more general cases where simple forms of $c_2(x, \rho)$ cannot be assumed, we claim that if $c_2(x, \rho)$ is smooth in ρ, we should generally expect $k_2(n)$ to be approximately linear in the expected mean-square error of the risk score.

We work from the following heuristic:

> For any given sample and a range of possible risk scores for that sample, the intervention taken will minimise the expected cost for the risk score which is unbiased.

This is equivalent to

$$\arg\min_{\rho} c_2(x, \rho) = f_t(x)$$

for all x in the domain of X. We suppose firstly that $c_2(x, \rho)$ is smooth in ρ, and write

$$c_2(x, \rho) = c_2\left(x, f_t(x)\right) + \frac{1}{2}\frac{\partial^2 c_2}{\partial \rho^2}\left(x, f_t(x)\right)\left(\rho - f_t(x)\right)^2 + O\left(\left(\rho - f_t(x)\right)^3\right),$$

noting that $\frac{\partial c_2}{\partial \rho}\left(x, f_t(x)\right) = 0$ by assumption.

The value $\frac{\partial^2 c_2}{\partial \rho^2}\left(x, f_t(x)\right)$ represents the curvature with respect to ρ of the function $c_2(x, \rho)$ about $\rho = f_t(x)$. Practically, this corresponds to the tolerance or robustness of the intervention: the amount of cost incurred due to a given deviation of the risk score from $f_t(x)$. We claim that this quantity will thus have relatively low variation across values of x, as the degree of robustness should be roughly constant.

Given this, we have

$$\mathbb{E}_X\left[c_2(X; \rho)\right] = \mathbb{E}_X\Big[c_2\left(x, f_t(x)\right)$$

$$
\begin{aligned}
&+ \frac{\partial^2 c_2}{\partial \rho^2}\left(x, f_t(x)\right)\left(\rho(X) - f_t(X)\right)^2 \\
&+ O\left(\sup_x (\rho(x) - f_t(x))^3\right)\Big] \\
&\approx \mathbb{E}_X\left[c_2\left(x, f_t(x)\right)\right] \\
&+ \mathbb{E}_X\left[\frac{\partial^2 c_2}{\partial \rho^2}\left(x, f_t(x)\right)\right]\mathbb{E}_X\left[\left(\rho(X) - f_t(X)\right)^2\right] \\
&+ O\left(\sup_x (\rho(x) - f_t(x))^3\right) \\
&\stackrel{\text{def}}{=} c_0 + c_2 \,\mathrm{MSE}(\rho) + O\left(\sup_x (\rho(x) - f_t(x))^3\right),
\end{aligned}
$$

where c_0, c_2 are independent of ρ and D_n, and $\mathrm{MSE}(\rho)$ is the standard mean-square error of ρ. Hence

$$
\begin{aligned}
k_2(n) &= \mathbb{E}_{D_n}\left\{\mathbb{E}_X\left[c_2(X;\rho)\right]\right\} \\
&\approx c_0 + c_2 \mathbb{E}_{D_n}\left\{MSE(\rho)\right\},
\end{aligned}
$$

so $k_2(n)$ is approximately linear in $\mathbb{E}_{D_n}\left\{MSE(\rho)\right\}$.

Suppose that we have a simple setting where we have a single intervention which we may use, which has a proportional effect on the risk of $Y = 1$ (that is, $g_t(x) = (1-\alpha) f_t(x)$, with $\alpha < 1$). We intervene on a sample if their risk score exceeds a particular threshold ρ_0. The cost function $c_2(x, \rho)$ is now discontinuous, but we may simply derive the form of k_2 in terms of risk score performance.

We assume that the intervention has a fixed cost γ_i, and that an event $Y = 1$ has a fixed cost γ_y. Then

$$
\begin{aligned}
c_2(x, \rho) &= \mathbb{E}_{C_2}\left\{(\text{Cost of an event}) + (\text{Cost of intervening})\right\} \\
&= \begin{cases} \gamma_y f_t(x) + 0 & \text{if } \rho < \rho_0 \\ \gamma_y \alpha f_t(x) + \gamma_i & \text{if } \rho \geq \rho_0 \end{cases},
\end{aligned}
$$

disregarding potential baseline costs common to all samples. We may now apply the heuristic above more directly, by presuming that the threshold ρ_0 is chosen so as to minimise the expectation of $c_2(X, \rho)$ over X under the assumption that $\rho(X) = f_t(X)$. In other words, we choose the threshold that gives us the best outcome assuming the risk score is correct.

This implies that for $f_t(x) < \rho_0$, we have $\gamma_y f_t(x) < \gamma_y \alpha f_t(x) + \gamma_i$ (if true risk is below the threshold, it is cheaper not to intervene) and for $f_t(x) \geq \rho_0$, we have $\gamma_y f_t(x) \geq \gamma_y \alpha f_t(x) + \gamma_i$ (if true risk is above the threshold, it is cheaper to intervene).

We now have:

$$
\begin{aligned}
c_2(x, \rho) - c_2(x, f_t(x)) &= \begin{cases} \gamma_y f_t(x) & \text{if } \rho < \rho_0, f_t(x) < \rho_0 \\ \gamma_y \alpha f_t(x) + \gamma_i - \gamma_y f_t(x) & \text{if } \rho \geq \rho_0, f_t(x) < \rho_0 \\ \gamma_y f_t(x) - (\gamma_y \alpha f_t(x) + \gamma_i) & \text{if } \rho < \rho_0, f_t(x) \geq \rho_0 \\ \gamma_y \alpha f_t(x) + \gamma_i & \text{if } \rho \geq \rho_0, f_t(x) \geq \rho_0 \end{cases} \\
&= 1_{(\rho < \rho_0)\, XOR\, (f_t(x) < \rho_0)} \left|\gamma_y (\alpha - 1) f_t(x) + \gamma_i\right|,
\end{aligned}
$$

and, denoting $\delta_\rho(x) = (\rho(x) < \rho_0)\, XOR\, (f_t(x) < \rho_0)$ and presuming $\lim_{n\to\infty} k_2(n)$ exists and is finite, we have

$$
k_2(n) = k_2(n) - \lim_{n\to\infty} k_2(n) + \lim_{n\to\infty} k_2(n)
$$

$$
\begin{aligned}
&= \lim_{n\to\infty} k_2(n) + \mathbb{E}_{D_n}\left\{\mathbb{E}_X\left[c_2(X,\rho) - c_2(X,f_t)\right]\right\} \\
&= c^0 + \mathbb{E}_{D_n}\left\{\mathbb{E}_X\left[1_{\delta_\rho(X)}|c^1 f_t(x) + c^2|\right]\right\}\ ,
\end{aligned}
\tag{47}
$$

where c^0, c^1, c^2 are constant, and readily estimated from several observations of k_2. This form is unsurprising: if a risk score $\rho(X)$ is such that the sign of $\rho(X) - \rho_0$ agrees with the sign of $f_t(X) - \rho_0$, it will have identical cost to a risk score $\rho(X)$ which agrees with $f_t(X)$ everywhere.

S6 Parametric OHS estimation

In this section, we describe estimation of optimal holdout set sizes by explicit parametrisation of the function $k_2(n)$. As in section 5.2 in the main paper, we take $k_2(n) = k_2(n;\theta)$. We will take k_2', k_2'', ℓ' to mean partial derivatives with respect to n, and the shorthand $\Theta = (N, k_1, \theta)$ and $\Theta_0 = \mathbb{E}(\Theta)$. We will also write $n_* = n_*(\Theta)$, $\ell(n_*) = \ell\{n_*(\Theta);\Theta\}$, $n_0 = n_*(\Theta_0)$ and $\ell(n_0) = \ell\{n_*(\Theta_0),\Theta_0\}$ for brevity. We presume that Θ is an unbiased estimate of Θ_0, so Θ_0 corresponds to 'true' parameter values.

We firstly develop asymptotic confidence intervals for parametric OHS estimates to link error in parameter estimates to error in optimal size. The sample-size m used in the following denotes a proxy for effort expended in estimating Θ_0.

Theorem S1. *Assume that $k_2''(n;\theta)$, $k_2'(n;\theta)$ and $\nabla_\theta k_2(n;\theta)$ are continuous in n and θ in some neighbourhood of (n_0,Θ_0), and that Θ_0 parametrizes a setting satisfying assumptions B1-B4. Suppose that Θ behaves as a mean of of m appropriately-distributed samples in satisfying $\sqrt{m}(\Theta-\Theta_0) \to N(0,\Sigma)$ in distribution where Θ_0 does not depend on m, that an estimate $\widehat{\Sigma}$ of Σ is available which is independent of Θ and satisfies $||\widehat{\Sigma}-\Sigma||_2 \to 0$ in distribution, and that n_0 is finite and unique as above. Then denoting*

$$\beta_\Theta = \frac{\partial^2\ell}{\partial n\partial\Theta_i}\Big/\frac{\partial^2\ell}{\partial n^2}, \qquad \gamma_\Theta = \frac{\partial\ell}{\partial\Theta_i}$$

we may uniquely define $n_0 = \{n : \ell'(n;\Theta_0) = 0\}$ and we have

$$\sqrt{m}(n_* - n_0) \to N\left(0, \beta_{\Theta_0}^t \Sigma \beta_{\Theta_0}\right), \qquad \sqrt{m}\left\{\ell(n_*) - \ell(n_0)\right\} \to N\left(0, \gamma_{\Theta_0}^t \Sigma \gamma_{\Theta_0}\right)$$

in distribution, and denoting $z_\alpha = \Phi^{-1}(1-\alpha/2)$, the confidence intervals

$$I_\alpha(\Theta,\hat{\Sigma}) = \left[n_*(\Theta) \pm z_\alpha\sqrt{\frac{\beta_\Theta^t\widehat{\Sigma}\beta_\Theta}{m}}\right], \qquad J_\alpha(\Theta,\hat{\Sigma}) = \left[\ell(n_*) \pm z_\alpha\sqrt{\frac{\gamma_\Theta^t\widehat{\Sigma}\gamma_\Theta}{m}}\right]$$

satisfy $P\left\{n_0 \in I_\alpha(\Theta,\hat{\Sigma})\right\} \to 1-\alpha$ and $P\left\{\ell(n_0) \in J_\alpha(\Theta,\hat{\Sigma})\right\} \to 1-\alpha$ as $m \to \infty$.

The proof is given in Supplement S6.2 below. A consequence is that for sufficiently accurately estimated costs, the OHS will be a non-trivial size:

Corollary S1. *Under assumptions of Theorems 4.1, S1, $P\{1 < n_*(\Theta) < N\} \to 1$ as $m \to \infty$.*

In light of the proportionality assumption in Section 5.1, and the tendency of the accuracy of a risk score with number of training samples ('learning curve') to follow a power-law form (Viering and Loog, 2021), we recommend considering such a parametric form for k_2 (i.e. $k_2(n;\theta) = an^{-b} + c$ with $\theta = (a,b,c)$), and provide explicit asymptotic confidence intervals for this setting in Supplement S6.1. Examples of variation in n_* and $\ell(n_*)$ with a power-law form for k_2, are shown in Supplementary Figures S2, S3.

Note that confidence intervals must be interpreted with with care: if the sampling distributions for k_1 and θ admit the possibility that assumptions of Theorem 4.1 are violated such that $P\left[k_1 < \liminf_{n\to\infty}\{k_2(n,\theta)\}\right] > 0$ then the standard error of n_* does not exist, as n_* can be undefined. Finite-sample confidence intervals may be constructed by bootstrapping (see function `ci_ohs()` in our R package `OptHoldoutSize`).

Our parametric algorithm assumes Θ is estimated from a multiset $\mathbf{n}$ of values in $\{1,\ldots,N\}$ and estimates $\mathbf{d}$ of $k_2(n)$ for each $n \in \mathbf{n}$ with known finite sampling variances $\boldsymbol{\sigma}^2$. For certain

multisets $\mathbf{n}$, estimates of Θ will not converge; for instance, if $\mathbf{n}$ contains only a single value repeated. The value m in Theorem S1 should be interpreted as an 'effective' population size, such that $\sqrt{m}\{\Theta(\mathbf{n}) - \Theta_0\} \to N(0, \Sigma)$ in distribution.

Given that our eventual aim to estimate the OHS with minimal error, we suggest the following way to iteratively select a new value $\tilde{n}$ at which an estimate $\hat{k_2}(\tilde{n})$ of $k_2(\tilde{n})$ should be made, given a set $\mathbf{n}$ of points at which estimates $\mathbf{k_2}$ of $k_2(\mathbf{n})$ have been made already. We denote by $\Theta(\mathbf{n}, \mathbf{k_2}, \boldsymbol{\sigma})$, $\hat{\Sigma}(\mathbf{n}, \mathbf{k_2}, \boldsymbol{\sigma})$ and $I_\alpha(\mathbf{n}, \mathbf{k_2}, \boldsymbol{\sigma})$ respectively the estimates of Θ_0, $\lim_{m\to\infty} \mathrm{var}\left[\sqrt{m}\{\Theta(\mathbf{n}, \mathbf{k_2}, \boldsymbol{\sigma}) - \Theta_0\}\right]$ and the width of the confidence interval $I_\alpha\left\{\Theta(\mathbf{n}, \mathbf{k_2}, \boldsymbol{\sigma}), \hat{\Sigma}(\mathbf{n}, \mathbf{k_2}, \boldsymbol{\sigma})\right\}$. Suppose we have the option of estimating $d(n)$ for one value of $n \in \{1, \ldots, N\}$ with known variance $\mathrm{var}\{d(n)\} = \sigma^2$. We select $\tilde{n}$ as:

$$\tilde{n} = \arg\min_n \mathbb{E}_{d(n)\sim N[k_2\{n, \Theta(\mathbf{n}, \mathbf{k_2}, \boldsymbol{\sigma})\}, \sigma^2]}\left[I_\alpha\left\{\mathbf{n} \cup n, \mathbf{k_2} \cup \hat{k_2}(n), \boldsymbol{\sigma} \cup \sigma\right\}\right], \tag{48}$$

that is, 'select the $\tilde{n}$ which will minimize the expected OHS confidence interval width if added to our set $\mathbf{n}$, with expectation computed with respect to our current parameter estimates'. If no minimum exists, $\tilde{n}$ is selected uniformly from $1, \ldots, N$. Algorithm 1, an expanded version of algorithm 1 in the main manuscript, shows our full parametric estimation procedure.

Algorithm 1: Parametric OHS estimation; with n_{add} estimates of $k_2(\cdot)$

$\mathbf{n}, \mathbf{k_2}, \boldsymbol{\sigma}^2 \leftarrow$ some initial values $\mathbf{n}$ with $(\boldsymbol{k_2})_i = \hat{k_2}(n_i) \approx k_2(n_i)$, $(\boldsymbol{\sigma}^2)_i = \mathrm{var}(\hat{k_2}(n_i))$;
while $|\mathbf{n}| < n_{add}$ **do**
  Find best new value $\tilde{n}$ to add to $\mathbf{n}$ as per formula 48 ;
  Estimate $\hat{k_2}(\tilde{n}) \approx k_2(\tilde{n})$;
  $\mathbf{n} \leftarrow (\mathbf{n} \cup \tilde{n})$, $\mathbf{k_2} \leftarrow \left\{\mathbf{k_2} \cup \hat{k_2}(\tilde{n})\right\}$, $\boldsymbol{\sigma}^2 \leftarrow \boldsymbol{\sigma}^2 \cup \mathrm{var}\left\{\hat{k_2}(\tilde{n})\right\}$;
end
Re-estimate OHS $n_*^{final} = n_*\{\Theta(\mathbf{n}, \mathbf{k_2}, \boldsymbol{\sigma})\}$;
return n_*^{final}

S6.1 Explicit partial derivatives for n^*, ℓ with power-law parametrisation

If we assume a power-law form of k_2, parametrised by $\theta = (a, b, c, k_1, N)$;

$$k_2(n;\theta) = an^{-b} + c\,,$$

then we have

$$\begin{aligned}
\frac{\partial n_*}{\partial a} &= \frac{1}{a}\left(\frac{bNn_* - (b-1)n_*^2}{b(b+1)N - b(b-1)n_*}\right)\\
\frac{\partial n_*}{\partial b} &= \frac{Nn_*(b\log(n_*) - 1) - n_*^2\left((b-1)\log(n_*) - 1\right)}{b(b+1)N - b(b-1)n_*}\\
\frac{\partial n_*}{\partial c} &= \frac{1}{a}\left(\frac{n_*^{b+2}}{b(b+1)N - b(b-1)n_*}\right)\\
\frac{\partial n_*}{\partial k_1} &= \frac{1}{a}\left(\frac{-n_*^{b+2}}{b(b+1)N - b(b-1)n_*}\right)\\
\frac{\partial n_*}{\partial N} &= \frac{bn_*}{b(b+1)N - b(b-1)n_*}\,,
\end{aligned}$$

and, more simply

$$\begin{aligned}
\frac{\partial}{\partial a}\ell(n_*;\theta) &= (N - n_*)n_*^{-b}\\
\frac{\partial}{\partial b}\ell(n_*;\theta) &= -\log(n_*)(N - n_*)an_*^{-b}\\
\frac{\partial}{\partial c}\ell(n_*;\theta) &= N - n_*\\
\frac{\partial}{\partial k_1}\ell(n_*;\theta) &= n_*\\
\frac{\partial}{\partial N}\ell(n_*;\theta) &= an_*^{-b} + c\,.
\end{aligned}$$

S6.2 Proof of Theorem S1

Theorem S1. *Assume that $k_2''(n;\theta)$, $k_2'(n;\theta)$ and $\nabla_\theta k_2(n;\theta)$ are continuous in n and θ in some neighbourhood of (n_0,Θ_0), and that Θ_0 parametrizes a setting satisfying assumptions B1-B4. Suppose that Θ behaves as a mean of of m appropriately-distributed samples in satisfying $\sqrt{m}(\Theta-\Theta_0)\to N(0,\Sigma)$ in distribution where Θ_0 does not depend on m, that an estimate $\widehat{\Sigma}$ of Σ is available which is independent of Θ and satisfies $||\widehat{\Sigma}-\Sigma||_2\to 0$ in distribution, and that n_0 is finite and unique as above. Then denoting*

$$\beta_\Theta=\frac{\partial^2\ell}{\partial n\partial\Theta_i}\Big/\frac{\partial^2\ell}{\partial n^2},\qquad \gamma_\Theta=\frac{\partial\ell}{\partial\Theta_i},$$

we may uniquely define $n_0=\{n:\ell'(n;\Theta_0)=0\}$ and we have

$$\sqrt{m}(n_*-n_0)\to N\left(0,\beta_{\Theta_0}^t\Sigma\beta_{\Theta_0}\right),\qquad \sqrt{m}\left\{\ell(n_*)-\ell(n_0)\right\}\to N\left(0,\gamma_{\Theta_0}^t\Sigma\gamma_{\Theta_0}\right) \tag{49}$$

in distribution, and denoting $z_\alpha=\Phi^{-1}(1-\alpha/2)$, the confidence intervals

$$I_\alpha(\Theta,\hat{\Sigma})=\left[n_*(\Theta)\pm z_\alpha\sqrt{\frac{\beta_\Theta^t\widehat{\Sigma}\beta_\Theta}{m}}\right],\qquad J_\alpha(\Theta,\hat{\Sigma})=\left[\ell(n_*)\pm z_\alpha\sqrt{\frac{\gamma_\Theta^t\widehat{\Sigma}\gamma_\Theta}{m}}\right]$$

satisfy $P\left\{n_0\in I_\alpha(\Theta,\hat{\Sigma})\right\}\to 1-\alpha$ and $P\left\{\ell(n_0)\in J_\alpha(\Theta,\hat{\Sigma})\right\}\to 1-\alpha$ as $m\to\infty$.

As above, we consider n_* as a function of parameters $\Theta=(N,k_1,\theta)$ (where $k_2(\cdot)=k_2(\cdot;\theta)$), write $n_*=n_*(\Theta)$, set $\Theta_0=E(\Theta)$, $n_0=n_*(\Theta_0)$ and $\ell(n_0)=\ell(n_0;\Theta_0)$ and $\ell(n_*)=\ell(n_*(\Theta),\Theta)$. As discussed above, n_* and $\ell(n_*)$ do not generally have means or standard errors.

Proof. From $\ell(n)=k_1n+k_2(n;\theta)(N-n)$ and $n_*=\{n:\ell'(n;\Theta)=0\}$, where such n_* is unique, we have (as per section 5.2)

$$(\nabla n_*)_i=\frac{\partial n_*}{\partial\Theta_i}=\frac{\frac{\partial^2 l}{\partial n\partial\Theta_i}}{\frac{\partial^2\ell}{\partial n^2}}=(\beta_{\Theta_0})_i$$
$$\left(\nabla\ell(n_*)\right)_i=\frac{\partial\ell}{\partial\Theta_i}=(\gamma_{\Theta_0})_i$$

for all components Θ_i of Θ. Thus partial derivatives of n_* exist as long as

$$\frac{\partial^2\ell}{\partial n^2}>0\,.$$

By assumption, $\ell(\cdot;\Theta_0)$ has a minimum at n_0. Since

$$\frac{\partial^2\ell}{\partial n^2}=\frac{\partial^2}{\partial n^2}k_2(n;\theta)-2\frac{\partial}{\partial n}k_2(n;\theta)\,,$$

where both terms are continuous in a neighbourhood of n_0, Θ_0 by assumption, the value of $\frac{\partial^2\ell}{\partial n^2}$ must be positive in some (possibly smaller) neighbourhood R_δ of (n_0,Θ_0) of width 2δ, and hence all partial derivatives of n_* and $\ell(n_*)$ are defined (and indeed continuous) in R_δ. Within R_δ we have

$$n_*(\Theta)=n_*(\Theta_0)+(\nabla n_*|_{\Theta=\Theta_0})\cdot(\Theta-\Theta_0)+O\left(||\Theta-\Theta_0||_2\right)$$

$$= n_0 + \beta^t_{\Theta_0} \cdot (\Theta - \Theta_0) + O\left(||\Theta - \Theta_0||_2\right) \tag{50}$$

$$\begin{aligned}\ell(n_*) &= \ell(n_*(\Theta_0); \Theta_0) + (\nabla \ell(n_*)|_{\Theta=\Theta_0}) \cdot (\Theta - \Theta_0) + O\left(||\Theta - \Theta_0||_2\right) \\ &= \ell(n_0) + \gamma^t_{\Theta_0} \cdot (\Theta - \Theta_0) + O\left(||\Theta - \Theta_0||_2\right) ,\end{aligned} \tag{51}$$

from which, given the assumption of asymptotic normality of Θ, assertions 49 follow. We note that despite this convergence in distribution, n_* and $\ell(n_*)$ do not generally have first or second moments for finite m.

We now have

$$\begin{aligned}P\left(n_0 \geq n_*(\Theta) + z_\alpha \sqrt{\frac{\beta_\Theta \widehat{\Sigma} \beta^t_\Theta}{m}}\right) &= P\left(\frac{\sqrt{m}}{z_\alpha}(n_0 - n_*(\Theta)) \geq \sqrt{\beta^t_\Theta \widehat{\Sigma} \beta_\Theta}\right) \\ &= P\Bigg(\frac{\sqrt{m}}{z_\alpha}(n_0 - n_*(\Theta)) \geq \big(\beta^t_{\Theta_0} \Sigma \beta_{\Theta_0} + \\ &\qquad \beta^t_\Theta \left(\widehat{\Sigma} - \Sigma\right) \beta_\Theta + \\ &\qquad (\beta_\Theta - \beta_{\Theta_0})^t \Sigma \left(\beta_\Theta + \beta_{\Theta_0}\right)\big)^{\frac{1}{2}}\Bigg) \\ &\to P\left(\frac{\sqrt{m}}{z_\alpha}(n_0 - n_*(\Theta)) \geq \sqrt{\beta^t_{\Theta_0} \Sigma \beta_{\Theta_0}}\right) \\ &= \frac{\alpha}{2} ,\end{aligned}$$

since, by the assumption of convergence of $\widehat{\Sigma}$

$$\begin{aligned}\left|\beta^t_\Theta \left(\Sigma - \widehat{\Sigma}\right) \beta_\Theta\right| &\leq ||\beta_\Theta||_2 ||\Sigma - \widehat{\Sigma}||_2 \\ &\to_p 0 ,\end{aligned}$$

and, since $P(\Theta \in R_\delta) \to 1$ by the asymptotic normality of Θ, we have from 50

$$\begin{aligned}\left|(\beta_\Theta - \beta_{\Theta_0})^t \Sigma \left(\beta_\Theta + \beta_{\Theta_0}\right)\right| &= O\left(||\beta_\Theta - \beta_{\Theta_0}||_2\right) \\ &\to_p 0 .\end{aligned}$$

Thus, combining with the corresponding limit for the lower end of $I_\alpha(\Theta, \hat{\Sigma})$:

$$P(n_0 \in I_\alpha(\Theta, \hat{\Sigma})) \to 1 - \alpha$$

as required. An identical argument holds for $J_\alpha(\Theta, \hat{\Sigma})$.

□

S7 Estimation of OHS by Bayesian Emulation

Our second algorithm for estimation of optimal holdout sizes uses Bayesian emulation (Brochu et al., 2010). In many cases, it may be difficult or unrealistic to provide a precise parametric form for the function $k_2(n)$. The function depends both on the 'learning curve' of the risk score, which may be complex (Viering and Loog, 2021), and the relationship of the risk score accuracy to the accrued cost, which may be nonlinear. Here, we propose a second algorithm which is less reliant on assuming a particular parametric form for $k_2(n)$.

As in the main manuscript, we approximate the cost function ℓ as an 'emulator' modelled as a Gaussian process, and take the minimum of its posterior mean over n as our OHS estimate. It is worth noting that whilst gaining an accurate approximation of the cost function is important, the main goal is to ascertain the minimum of this function, not provide a universally effective approximation at all points. Therefore, we aim to choose the location of design points $\mathbf{n}$ in order to efficiently obtain the minimum of the cost function, and hence the OHS.

First we must construct an emulator which approximates the cost function. We begin with an initial set of design points $\mathbf{n}$ and their corresponding observed noisy cost estimates $\mathbf{d}$. The prior for our emulator is, following Vernon et al. (2018), $\ell(n) = m(n, \Theta) + u(n)$ with mean function $m(n, \Theta) = k_1 n + k_2(n; \theta)(N - n)$, given some initial estimate of $\Theta = (N, k_1, \theta)$, and $u(n)$ a zero-mean Gaussian process

$$u(n) \sim \mathcal{GP}\left\{0, k(n, n')\right\} \qquad k(n, n') = \sigma_u^2 \exp\left\{-\left(\frac{n - n'}{\zeta}\right)^2\right\}$$

where k is chosen to enforce smoothness in $\ell(n)$, though other covariance functions having varying degrees of smoothness could be used. The hyperparameters θ, σ_u and ζ are problem-specific and must be specified; however, we will show that for sufficiently large $|\mathbf{n}|$ mis-specification of θ, σ_u and θ is overcome.

Since $\mathbf{n}$ may be a multiset, we take $\mathbf{n}^1$ as the set of unique values in $\mathbf{n}$, with $\mathbf{d}^1$, $\boldsymbol{\sigma}^1$ defined correspondingly with d_i^1 as the mean of $\{d_j : n_j = n_i^1\}$ with sample standard error σ_i^1, noting that σ_i^1 may change with i. We state $d_i^1 = \ell(n_i^1) + \mathfrak{e}$ where $\mathfrak{e} \sim N\left\{0, (\sigma_i^1)^2 = \frac{(\tilde{\sigma}_i^1)^2}{|j : n_j = n_i^1|}\right\}$ and $(\tilde{\sigma}_i^1)^2$ is the sample variance of a single evaluation at n_i^1. Alternatively, we may account for the variation in $\mathbf{d}$ through 'inactive' variables and opt to use a 'nugget' term; this approach is described in detail in Supplement S7.1.

Now with input n, with an unevaluated loss value, our emulator specifies that the joint distribution of $\ell(n)$ and our observed output values $\mathbf{d}^1$ is:

$$\begin{bmatrix} \ell(n) \\ \mathbf{d^1} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(n, \Theta) \\ m(\mathbf{n^1}, \Theta) \end{bmatrix}, \begin{bmatrix} k(n, n) & k(n, \mathbf{n^1}) \\ k(\mathbf{n^1}, n) & k(\mathbf{n^1}, \mathbf{n^1}) + \mathrm{diag}\{(\boldsymbol{\sigma^1})^2\} \end{bmatrix}\right),$$

where $m(\mathbf{n^1}, \Theta)_i^T = m(n_i^1, \Theta)$, $k(n, \mathbf{n^1})_i = k(\mathbf{n^1}, n)_i^T = k(n, n_i^1)$, $k(\mathbf{n}, \mathbf{n})_{ij} = k(n_i^1, n_j^1)$, $\mathrm{diag}\{(\boldsymbol{\sigma^1})^2\}_{ij} = (\sigma_i^1)^2 1_{i=j}$. By obtaining the conditional posterior distribution $\pi_{\mathbf{n}} = \pi\{\ell(n) \mid n, \mathbf{n^1}, \mathbf{d^1}, \boldsymbol{\sigma^1}\}$ and taking the expectation and variance we gain the Bayes linear update equations (Vernon et al., 2018):

$$\begin{aligned} \mu(n) &= \mathbb{E}_{\pi_{\mathbf{n}}}\{\ell(n)\} = m(n, \Theta) + k(n, \mathbf{n^1})\left[k(\mathbf{n^1}, \mathbf{n^1}) + \mathrm{diag}\{(\boldsymbol{\sigma^1})^2\}\right]^{-1}\{\mathbf{d^1} - m(\mathbf{n^1}, \Theta)\} \\ \Psi(n) &= \mathrm{var}_{\pi_{\mathbf{n}}}\{\ell(n)\} = k(n, n) - k(n, \mathbf{n^1})[k(\mathbf{n^1}, \mathbf{n^1}) + \mathrm{diag}\{(\boldsymbol{\sigma^1})^2\}]^{-1} k(\mathbf{n^1}, n)\,. \end{aligned}$$

In algorithm 1, selection of new design points should generally favour well-spaced points across $\{1, \ldots, N\}$ for both exploration and exploitation. Here, since we wish both to estimate the OHS

accurately but also locally approximate ℓ well, we choose the next n in a way which predominantly but not completely favours exploitation. We use the 'expected improvement', which measures discrepancy between the emulator at a certain design point and the known minimum $EI(\cdot)$ (Brochu et al., 2010):

$$EI(n) = \{d^- - \mu(n)\}\Phi\left(\frac{d^- - \mu(n)}{\sqrt{\Psi(n)}}\right) + \sqrt{\Psi(n)}\phi\left(\frac{d^- - \mu(n)}{\sqrt{\Psi(n)}}\right),$$

where $d^- = \min_i\{\mathbf{d^1}_i\}$, and

$$\tilde{n} = \arg\max_{n\in\{1,\ldots,N\}} EI(n) = \arg\max\left(\mathbb{E}_{\pi_{\mathbf{n}}}\left[\max\{0, d^- - \ell(n)\}\right]\right).$$

We see that by formulating the problem in terms of $EI(\cdot)$, there is a natural stopping criterion on the size of $\mathbf{n}$: setting a threshold $EI(\tilde{n}) > \tau$ allows us to specify that for each iteration that we expect total cost to improve by at least τ over our current known minimum d^-. Examples of $\mu(\cdot), \Psi(\cdot), EI(\cdot)$ are shown in Supplementary Figure S4. This leads to algorithm 2 for OHS estimation by Bayesian Emulation (a more precise version of algorithm 2 in the main manuscript).

Algorithm 2: Emulation OHS estimation; minimum cost improvement τ

$\mathbf{n}, \mathbf{d} \leftarrow$ some initial values $\mathbf{n}$ with $(\boldsymbol{d})_i = d(n_i) \approx \ell(n)$;
Coalesce $\mathbf{n}, \mathbf{d}$ into $\mathbf{n^1}, \mathbf{d^1}$ and obtain $\boldsymbol{\sigma^1}$ as above ;
Estimate functions $\mu(n)$, $\Psi(n)$, $EI(n)$, with $\Theta = \Theta(\mathbf{n^1}, \mathbf{d^1}, \boldsymbol{\sigma^1})$;
while $\max_{n\in\{1,\ldots,N\}}\{EI(n)\} > \tau$ **do**
    $\tilde{n} \leftarrow \arg\max_{n\in\{1,\ldots,N\}} EI(n)$;
    Estimate $d(\tilde{n}) \approx k_2(\tilde{n})$;
    $\mathbf{n} \leftarrow (\mathbf{n} \cup \tilde{n})$; $\mathbf{d} \leftarrow (\mathbf{d} \cup d\{\tilde{n}\})$;
    Coalesce $\mathbf{n}, \mathbf{d}$ into $\mathbf{n^1}, \mathbf{d^1}$ and obtain $\boldsymbol{\sigma^1}$;
    Re-estimate functions $\mu(n)$, $\Psi(n)$, $EI(n)$, with $\Theta = \Theta(\mathbf{n^1}, \mathbf{d^1}, \boldsymbol{\sigma^1})$;
end
return $n_*^{final} = \arg\min_{n_i\in\mathbf{n}^1}\left\{d_i^1\right\}$

Various results on the consistency of the expected improvement algorithm have been proved, albeit in differing settings; either with noiseless observations $\mathbf{d}$ (Locatelli, 1997; Vazquez and Bect, 2010; Bull, 2011) or with noisy observations with known variance (Ryzhov, 2016). We prove the following consistency results specifically for the setting of this work in Supplement S7.2.

Theorem S2. *If $\ell(n)$, $\tilde{\boldsymbol{\sigma}}^{\mathbf{1}}$, and $m(n, \Theta)$ are almost surely bounded and $d_i^1 = \ell(n_i^1) + \mathfrak{e}$ where $\mathfrak{e} \sim N\left\{0, \frac{(\tilde{\sigma}_i^1)^2}{|j:n_j=n_i^1|}\right\}$ then for every $n \in \{1, \ldots, N\}$, as the multiplicity of n in $\mathbf{n}$ tends to ∞ we have $\mu(n) \longrightarrow \ell(n)$ and $\Psi(n) \to 0$ almost surely with respect to variation in $\mathbf{d}$.*

also noting the following simple result, proved in Supplement S7.3:

Theorem S3. *Given the conditions of Theorem S2, for every $n \in \{1, \ldots, N\}$, as the multiplicity of n in $\mathbf{n}$ tends to ∞,*

$$EI(n) \longrightarrow 0$$

almost surely with respect to randomness in $\mathbf{d}$

These results assert that $\mu(n)$ can eventually approximate any loss function sufficiently well given enough estimates of ℓ at all values of n. It is not obvious that this is guaranteed by algorithm 2, although we show that this generally does occur in the following, the proof of which is given in Supplement S7.4:

Theorem S4. *If $\ell(n)$, $\tilde{\boldsymbol{\sigma}}^{\mathbf{1}}$, and $m(n, \Theta)$ are almost surely bounded and $d_i^1 = \ell(n_i^1) + \mathfrak{e}$ where $\mathfrak{e} \sim N\left\{0, \frac{(\tilde{\sigma}_i^1)^2}{|j:n_j = n_i^1|}\right\}$ then under algorithm 2 with $\tau = 0$, the value $\mu(\tilde{n})$ converges almost surely to $\ell(\tilde{n})$ for every $\tilde{n} \in \{1, \ldots, N\}$.*

We characterize the error in n_* using 'the number of values of n for which the probability of the true cost at holdout set size n is less than the estimated minimum cost exceeds $1 - \alpha$', or formally: $\{n : pr_{\mathbf{n}_\pi}\{\ell(n) < \mu(n_*)\} \geq 1 - \alpha\}$, although this should not be interpreted as a credible set for n_*. This is implemented in our R package `OptHoldoutSize`, available on CRAN.

S7.1 Emulation of cost function with nugget term

Rather than explaining the variation of values in $\mathbf{d}$ corresponding to a design point in $\mathbf{n}^{\mathbf{1}}$ as approximation error of a deterministic loss function, we can explain this variation as the result of not including active variables, being the data (X, Y). Note that as a consequence we are now not emulating a deterministic function $\ell(n)$ as we are not generalising the loss through expectations, we are generalising the loss through omission of the data which generated $\mathbf{d}$. To clarify this distinction we replace the loss function $\ell(n)$ with the stochastic function $\mathcal{E}(n)$.

Now we may specify variation in $\mathbf{d}$ using a 'nugget' term $w(n)$, following Bower et al. (2010):

$$\mathcal{E}(n) = m(n) + u(n) + w(n)\,,$$

where $m(n)$ and $u(n)$ are as before but now $w(n)$ represents our nugget term, which we again specify as a Gaussian process:

$$w(n) \sim \mathcal{GP}(0, \kappa(n, n'))\,,$$

with

$$\kappa(n, n') = \begin{cases} \kappa(n) & \text{if } n = n' \\ 0 & \text{otherwise.} \end{cases} \tag{52}$$

Since there is less variance in risk scores fitted to larger datasets, we expect less variance in $\mathcal{E}(n)$ for larger n, so we specify $\kappa(n)$ as a monotonically decreasing function in n.

The joint distribution between $\mathcal{E}(n)$ and $\mathbf{d}^{\mathbf{1}}$ is now:

$$\begin{bmatrix} \mathcal{E}(n) \\ \mathbf{d}^{\mathbf{1}} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(n) \\ m(\mathbf{n}^{\mathbf{1}}) \end{bmatrix}, \begin{bmatrix} k(n,n) + \kappa(n) & k(n, \mathbf{n}^{\mathbf{1}}) \\ k(\mathbf{n}^{\mathbf{1}}, n) & k(\mathbf{n}^{\mathbf{1}}, \mathbf{n}^{\mathbf{1}}) + \mathrm{diag}(\kappa(\mathbf{n}^{\mathbf{1}})) \end{bmatrix}\right).$$

This then gives our Bayes linear update equations in terms of $\pi_{\mathbf{n}} = \pi(\mathcal{E}(n)|n, \mathbf{n}^{\mathbf{1}}, \mathbf{d}^{\mathbf{1}})$ as

$$\begin{aligned} \mu(n) &= \mathbb{E}_{\pi_{\mathbf{n}^{\mathbf{1}}}}(\mathcal{E}(n)) \\ &= m(n) + k(n, \mathbf{n}^{\mathbf{1}})[k(\mathbf{n}^{\mathbf{1}}, \mathbf{n}^{\mathbf{1}}) + \mathrm{diag}(\kappa(\mathbf{n}^{\mathbf{1}}))]^{-1}(\mathbf{d}^{\mathbf{1}} - m(\mathbf{n}^{\mathbf{1}})) \\ \Psi(n) &= \mathrm{var}_{\pi_{\mathbf{n}^{\mathbf{1}}}}(\mathcal{E}(n)) \\ &= k(n,n) + \kappa(n) - k(n, \mathbf{n})[k(\mathbf{n}^{\mathbf{1}}, \mathbf{n}^{\mathbf{1}}) + \mathrm{diag}(\kappa(\mathbf{n}^{\mathbf{1}}))]^{-1} k(\mathbf{n}^{\mathbf{1}}, n)\,. \end{aligned} \tag{53}$$

Note that this differs only slightly from the emulator constructed in section 5.3, with the main difference being we now attribute uncertainty in the loss values as an inherent behaviour of our

emulator and not in the procedure to obtain these loss values. As a result $\kappa(n)$ does not decrease as the multiplicity of elements of $\mathbf{n}$ increases, which represents a major disadvantage to the uncertainty representation in section 5.3.

One may then be sceptical of the benefit of duplicating design points for this method, and whilst it is possible to use this method without duplication (i.e $\mathbf{n} = \mathbf{n^1}$), the consequence of this would be that we are heavily reliant on a singular sample to locate the minimum which could be misleading. Averaging various samples at the same design point mitigates this potential problem, as does replacing d^- with $\mu^- = \min_i\{\mu(\mathbf{n^1}_i)\}$ as detailed in Brochu et al. (2010). Taking the median of samples instead of a weighted mean is more appropriate here as we are not seeking to accurately approximate an expectation, instead we only wish to avoid extreme samples misleading our search for the minimum.

S7.2 Proof of Theorem S2

Theorem S2. *If $\ell(n)$, $\tilde{\boldsymbol{\sigma}}^{\mathbf{1}}$, and $m(n,\Theta)$ are almost surely bounded and $d_i^1 = \ell(n_i^1) + \mathfrak{e}$ where $\mathfrak{e} \sim N\left\{0, \frac{(\tilde{\sigma}_i^1)^2}{|j:n_j=n_i^1|}\right\}$ then for every $n \in \{1,\dots,N\}$, as the multiplicity of n in $\mathbf{n}$ tends to ∞ we have $\mu(n) \longrightarrow \ell(n)$ and $\Psi(n) \to 0$ almost surely with respect to variation in $\mathbf{d}$.*

Proof. Assume W.L.O.G that $(\mathbf{n}^1)_1 = n$. Since $\tilde{\boldsymbol{\sigma}}^{\mathbf{1}}$ is bounded, we have (from the distributional assumption of noise in evaluations) $\text{var}\left((\mathbf{d}^1)_1\right) = (\boldsymbol{\sigma}^{\mathbf{1}})_1^2 \to 0$, so $(\mathbf{d}^1)_1 \longrightarrow \ell(n)$ almost surely. We now prove that $k(n,\mathbf{n^1})[k(\mathbf{n^1},\mathbf{n^1}) + \text{diag}((\boldsymbol{\sigma}^{\mathbf{1}})^2)]^{-1} = (1,0,\dots,0)$ when $(\boldsymbol{\sigma}^{\mathbf{1}})_1 = 0$. Now:

$$\begin{aligned} k(n,\mathbf{n^1})[k(\mathbf{n^1},\mathbf{n^1}) + \text{diag}((\boldsymbol{\sigma}^{\mathbf{1}})^2)]^{-1} &= (1,0,\dots,0) \\ \Leftrightarrow \qquad k(n,\mathbf{n^1}) &= (1,0,\dots,0) * [k(\mathbf{n^1},\mathbf{n^1}) + \text{diag}((\boldsymbol{\sigma}^{\mathbf{1}})^2)]\,, \end{aligned}$$

and $k(n,\mathbf{n^1}) = (1,0,\dots,0) * k(\mathbf{n^1},\mathbf{n^1} + \text{diag}((\boldsymbol{\sigma}^{\mathbf{1}})^2)_{-1})$ is true by definition as the first row of $k(\mathbf{n^1},\mathbf{n^1}) + \text{diag}((\boldsymbol{\sigma}^{\mathbf{1}})^2)$ is $k(n,\mathbf{n^1})$. Therefore,

$$\mu(n) = m(n,\Theta) + (1,0,\dots,0)(\mathbf{d^1} - m(\mathbf{n^1},\Theta)) = m(n,\Theta) + d(n) - m(n,\Theta) = d(n) = \ell(n)$$

almost surely, and

$$\Psi(n) = k(n,n) - (1,0,\dots,0)k(\mathbf{n^1},n) = k(n,n) - k(n,n) = 0$$

in the limit.

□

S7.3 Proof of Theorem S3

Theorem S3. *Given the conditions of Theorem S2, for every $n \in \{1,\dots,N\}$, as the multiplicity of n in $\mathbf{n}$ tends to ∞,*

$$EI(n) \longrightarrow 0$$

almost surely with respect to randomness in $\mathbf{d}$

Proof. From Theorem S2 we have that $\mu(n) \longrightarrow \ell(n) < \infty$ and $d_i^1 \longrightarrow \ell(n_i^1)$, so therefore in the limit we can state $pr_{\mathbf{d}}(-\infty < d^- - \mu(n) \leq 0) = 1$. Indeed, let j be the index such that $n_j^1 = n$. If in the limit $d^- > \mu(n) = d(n)$ then this implies that $d_j^1 < \min_i\{d_i^1\}$ which is a contradiction. Also note from Theorem S2 that $\Psi(n) \longrightarrow 0$ and that $\Phi(\cdot) \in (0,1)$, $\phi(\cdot) \in (0,(2\pi)^{-1/2}]$. As a result the following two scenarios have joint probability 1:

- $d^- - \mu(n) = 0$ in the limit: As $\Phi(\cdot)$, $\phi(\cdot)$ are bounded and $\Psi(n) = 0$ in the limit, we also have $EI(n) = 0$ in the limit.
- $\infty < d^- - \mu(n) < 0$ in the limit: As $\Psi(n) = 0$ in the limit, $\Phi\left(\frac{d^- - \mu(n)}{\sqrt{\Psi(n)}}\right) = 0$ in the limit. As $\phi(\cdot)$ is bounded we have that $EI(n) = 0$ in the limit.

which proves the corollary.

□

S7.4 Proof of Theorem S4

Theorem S4. *If $\ell(n)$, $\tilde{\boldsymbol{\sigma}}^{\mathbf{1}}$, and $m(n, \Theta)$ are almost surely bounded and $d_i^1 = \ell(n_i^1) + \mathfrak{e}$ where $\mathfrak{e} \sim N\left\{0, \frac{(\tilde{\sigma}_i^1)^2}{|j:n_j = n_i^1|}\right\}$ then under algorithm 2 with $\tau = 0$, the value $\mu(\tilde{n})$ converges almost surely to $\ell(\tilde{n})$ for every $\tilde{n} \in \{1, \ldots, N\}$.*

Proof. Our overall argument is to show that algorithm 2 leads to the multiplicity of $\tilde{n}$ in $\mathbf{n}$ tending to infinity, from which the result follows from Theorem S2.

To do this, we begin with the following two lemmas, the second of which describes the limiting behaviour of $EI(n)$ according to how often n occurs in $\mathbf{n}$: namely that if the multiplicity of n in $\mathbf{n}$ diverges, the value of $EI(n)$ converges to 0; otherwise, it remains positive. We introduce the index $EI_{\mathbf{n}}(n)$ to indicate the dependence of $EI(n)$ on $\mathbf{n}$ and assume that the function $\ell(n)$ is fixed. For a multiset $\mathbf{n}_i$, we denote $\text{mult}_{\mathbf{n_i}}(n)$ as the multiplicity of n in $\mathbf{n_i}$.

Lemma S5. *Suppose $m \times m$ matrix A is symmetric. Denote by I^1 the $m \times m$ matrix with $I^1_{ij} = 1_{i=j=1}$. Let x be a vector of length m and denote by A_x the matrix A with its top row replaced by x. Then for p in any interval containing 0 on which $A + pI^1$ is invertible we have*

$$\frac{\partial}{\partial p}\left(x^T(A + pI^1)^{-1}x\right) = -\frac{|A_x|^2}{|A + pI^1|^2}. \tag{54}$$

Proof. If $M(p)$ is invertible in a neighbourhood of p we have $\frac{\partial M^{-1}}{\partial p} = -M^{-1}\frac{\partial M}{\partial p}M^{-1}$, and if M is symmetric with dimensions $m \times m$ and first row M_1, then $MI^1M = M_1M_1^T$. Since $(A + pI)$ and A differ only in the top row, we have $\text{adj}(A + pI)_1 = \text{adj}(A)_1$, where $\text{adj}(\cdot)$ indicates the adjugate matrix and $\cdot_1$ the top row. We now have

$$\begin{aligned}
\frac{\partial}{\partial p}\left(x^T(A + pI^1)^{-1}x\right) &= -x^T(A + pI^1)^{-1}\frac{\partial(A + pI^1)}{\partial p}(A + pI^1)^{-1}x \\
&= -x^T(A + pI^1)^{-1}I^1(A + pI^1)^{-1}x \\
&= \frac{x^T\text{adj}(A + pI^1)I^1\text{adj}(A + pI^1)x}{|A + pI^1|^2} \\
&= -\frac{x^T\text{adj}(A + pI^1)_1\text{adj}(A + pI^1)_1^Tx}{|A + pI^1|^2} \\
&= -\frac{x^T\text{adj}(A)_1\text{adj}(A)_1^Tx}{|A + pI^1|^2} \\
&= -\frac{|A_x|^2}{|A + pI^1|^2}
\end{aligned}$$

as required. □

Lemma S6. *Let S_1 and S_2 be disjoint subsets of $[N] = \{1, \ldots, N\}$ with $S_1 \cup S_2 = [N]$. For a multiset $\mathbf{n}$ denote*

$$\begin{aligned} q_1(\mathbf{n}) &= \max_{n \in S_1} mult_{\mathbf{n}}(n) \\ q_2(\mathbf{n}) &= \min_{n \in S_2} mult_{\mathbf{n}}(n)\,. \end{aligned} \tag{55}$$

Suppose we have infinite sequences $\mathbf{n}$, $\mathbf{d}$. Let $\mathbf{n_i}$, $\mathbf{d_i}$ denote the (multiset) first i elements of each sequence, and let $\mathbf{n_i^1}$ be the unique values of n in $\mathbf{n_i}$ and $\mathbf{d_i^1}$, $\tilde{\boldsymbol{\sigma}}_{\boldsymbol{i}}^{\mathbf{1}}$ be the associated mean and sample standard deviation, with $\tilde{\boldsymbol{\sigma}}_{\boldsymbol{i}}^{\mathbf{1}}$ upper bounded. Suppose that $q_1(\mathbf{n_i}) \leq m_1$ for all i and $q_2(\mathbf{n_i}) \to \infty$, and the set $\{k_2(n, \Theta_i) = k_2\left(n, \Theta(\mathbf{n_i}, \mathbf{d_i}, \tilde{\boldsymbol{\sigma}}_{\boldsymbol{i}})\right) : n \in \{1, \ldots, N\}, i \in \mathbb{N}\}$ is almost surely asymptotically bounded. Then for sufficiently large σ_u:

$$\limsup_{i \to \infty} EI_{\mathbf{n}_i}(n) = \begin{cases} e_n > 0 & \text{if } n \in S_1 \\ 0 & \text{if } n \in S_2 \end{cases} \tag{56}$$

almost surely.

Proof. We will in fact show that even $\liminf EI_{\mathbf{n_i}}(n) > 0$ for $n \in S_1$, but lim sup will suffice for our purposes. We note that

$$EI_{\mathbf{n}_i}(n) > 0 \Leftrightarrow \sqrt{\Psi_{\mathbf{n}_i}(n)}\phi\left(\frac{d^-_{\mathbf{n}_i} - \mu_{\mathbf{n}_i}(n)}{\sqrt{\Psi_{\mathbf{n}_i}(n)}}\right) > (\mu_{\mathbf{n}_i}(n) - d^-_{\mathbf{n}_i})\Phi\left(\frac{d^-_{\mathbf{n}_i} - \mu_{\mathbf{n}_i}(n)}{\sqrt{\Psi_{\mathbf{n}_i}(n)}}\right). \tag{57}$$

We will show that for all n, we have

$$P\left(-\infty < \liminf_{i \to \infty} \left(d^-_{\mathbf{n_i}} - \mu_{\mathbf{n_i}}(n)\right)\right) = 1\,. \tag{58}$$

By the argument in Theorem S2 and corollary S3 we have for $n \in S_2$ that $\lim_{i\to\infty} \Psi_{\mathbf{n_i}}(n) = 0$, from which both sides of 57 converge to 0. For $n \in S_1$ we will show $\lim_{i\to\infty} \Psi_{\mathbf{n_i}}(n) > 0$, in which case we may define

$$z_{\mathbf{n}_i}(n) = \frac{\mu_{\mathbf{n}_i}(n) - d^-_{\mathbf{n}_i}}{\sqrt{\Psi_{\mathbf{n}_i}(n)}}\,,$$

from which inequality 57 reduces to

$$\phi(z_{\mathbf{n}_i}(n)) > z_{\mathbf{n}_i}(n)\Phi(-z_{\mathbf{n}_i}(n))\,,$$

which holds for all $-\infty \leq z_{\mathbf{n}_i}(n) < \infty$. Since $z_{\mathbf{n}_i}(n)$ is asymptotically bounded between positive values, the result follows.

Beginning with $d^-_{\mathbf{n_i}}$, we note that $d^-_{\mathbf{n_i}}$ is the minimum of

1. Values of $\mathbf{d_i^1}$ corresponding to values of $\mathbf{n_i^1}$ in S_1; and
2. Values of $\mathbf{d_i^1}$ corresponding to values of $\mathbf{n_i^1}$ in S_2

For sufficiently large s, the sequence $\{n_j = (\mathbf{n})_j : j > s\}$ never contains any $n \in S_1$ again; hence, the minimum of item 1 is determined after finitely many i and its limit is finite. Since $\tilde{\boldsymbol{\sigma}}_{\boldsymbol{i}}^{\mathbf{1}}$ is upper-bounded, all values of $\mathbf{d_i^1}$ in item 2 converge to finite values in $\{\ell(n) : n \in S_2\}$ almost surely. Hence $d^-_{\mathbf{n_i}}$ converges almost surely to a finite value.

Since $\limsup_{i\to\infty}$ and $\liminf_{i\to\infty}$ of $m\left(n;\Theta\left(\mathbf{n_i},\mathbf{d_i},\tilde{\boldsymbol{\sigma}}_{\boldsymbol{i}}^{\mathbf{1}}\right)\right)$ are almost surely finite, all terms in $\mu(n)$ are asymptotically finite, from which equation 58 follows.

It remains to consider $\Psi_{\mathbf{n}_i}(n)$ for $n \in S_1$. Firstly take $n \in \mathbf{n^1}$ and suppose W.L.O.G that $\boldsymbol{n}_{\boldsymbol{i}\,1}^{\mathbf{1}} = n$. Since $n \in S_1$ we have $\lim_{i\to\infty} \mathrm{mult}_{\mathbf{n_i}}(n) > 0$ so $\lim_{i\to\infty}(\boldsymbol{\sigma}_{\boldsymbol{i}}^{\mathbf{1}})_1$ exists and is positive. Denoting $\boldsymbol{\sigma'}$ as $\boldsymbol{\sigma}_{\boldsymbol{i}}^{\mathbf{1}}$ with 0 substituted for the first element, we have

$$\begin{aligned}\frac{\partial}{\partial(\boldsymbol{\sigma}_{\boldsymbol{i}}^{\mathbf{1}})_1^2}\Psi_{\mathbf{n_i}}(n) &= \frac{\partial}{\partial(\boldsymbol{\sigma}_{\boldsymbol{i}}^{\mathbf{1}})_1^2}\left(k(n,n) - k(n,\mathbf{n_i^1})[k(\mathbf{n_i^1},\mathbf{n_i^1}) + \mathrm{diag}((\boldsymbol{\sigma}_{\boldsymbol{i}}^{\mathbf{1}})^2)]^{-1}k(\mathbf{n_i^1},n)\right)\\ &= \frac{\left|k(\mathbf{n_i^1},\mathbf{n_i^1}) + \mathrm{diag}((\boldsymbol{\sigma'})^2)\right|^2}{\left|k(\mathbf{n_i^1},\mathbf{n_i^1}) + \mathrm{diag}((\boldsymbol{\sigma}_{\boldsymbol{i}}^{\mathbf{1}})^2)\right|^2}\\ &> 0\end{aligned}$$

by Lemma S5; hence $\Psi_{\mathbf{n_i}}(n)$, considered as a function of $(\boldsymbol{\sigma}_{\boldsymbol{i}}^{\mathbf{1}})_1^2$, is increasing. Given that $\lim_{i\to\infty}(\boldsymbol{\sigma}_{\boldsymbol{i}}^{\mathbf{1}})_j$ is 0 for $(\mathbf{n_i^1})_j \in S_2$ and is positive for $(\mathbf{n_i^1})_j \in S_1$, we conclude that $\lim_{i\to\infty}\Psi_{\mathbf{n_i}}(n)$ is positive when $n \in S_1$ and $n \in \mathbf{n^1}$.

If $n \notin \mathbf{n^1}$, so n never occurs in any $\mathbf{n_i}$, then we firstly note that since $k(n,n) < k(n,m)$ for any $m \neq n$, we have:

$$k(n,n) - k(n,\mathbf{n_i^1})[k(\mathbf{n_i^1},\mathbf{n_i^1}))]^{-1}k(\mathbf{n_i^1},n) > 0\,.$$

This omits the term $\mathrm{diag}((\boldsymbol{\sigma}_{\boldsymbol{i}}^{\mathbf{1}})^2$ from the expression for $\Psi_{\mathbf{n_i}}(n)$. However, if we denote k'_j the matrix $k(\mathbf{n_i^1},\mathbf{n_i^1}) + \mathrm{diag}((\boldsymbol{\sigma}_{\boldsymbol{i}}^{\mathbf{1}})^2)$ with the jth row replaced by $k(n,\mathbf{n_i^1})$, we have from Lemma S5:

$$\frac{\partial}{\partial(\boldsymbol{\sigma}_{\boldsymbol{i}}^{\mathbf{1}})_j^2}\Psi_{\mathbf{n_i}}(n) = \frac{\left|k'_j\right|^2}{\left|k(\mathbf{n_i^1},\mathbf{n_i^1}) + \mathrm{diag}((\boldsymbol{\sigma}_{\boldsymbol{i}}^{\mathbf{1}})^2)\right|^2} > 0\,,$$

for any element $(\boldsymbol{\sigma}_{\boldsymbol{i}}^{\mathbf{1}})_j^2$ of $(\boldsymbol{\sigma}_{\boldsymbol{i}}^{\mathbf{1}})^2$; hence $\Psi_{\mathbf{n_i}}(n)$ is increasing in any such element and its positivity follows. This completes the proof of the lemma.

□

Now suppose that some $n \in \{1,\ldots,N\}$ occurs only finitely often in $\mathbf{n^1}$. Then there must be some largest set S_1 of such n, with complement $S_2 = \{1,\ldots,N\} \setminus S_1$. Since every element in S_1 occurs in $\mathbf{n^1}$ with finite multiplicity there must be some j such that no $n \in S_1$ occurs amongst the values $\{(\mathbf{n^1})_{j+1},(\mathbf{n^1})_{j+2},\ldots\}$. But from Lemma S6, there will almost surely eventually be some $J > j$ for which some value in $\{EI_{\mathbf{n_J}}(n) : n \in S_1\}$ exceeds all values in $\{EI_{\mathbf{n_J}}(n) : n \in S_1\}$, and hence $(\mathbf{n^1})_{J+1} \in S_1$ (as long as τ is sufficiently small), contradicting the choice of j. So the event that an $n \in \{1,\ldots,N\}$ occurs in $\mathbf{n^1}$ with finite multiplicity has probability 0. This completes the proof.

□

S7.5 Repetitive Expected Improvement

Typically expected improvement algorithms specify τ as a definitive stopping criterion when evaluations of the true function are noiseless Brochu et al. (2010). However, in the presence of noise the termination of algorithm 2 may result in the selection of a hold-out set size whose cost evaluation has high sample variance. As shown by Theorem S4 when $\tau = 0$ this premature termination does not occur and we select the optimal hold-out set with certainty, but this is not practical as this would lead to an algorithm which does not terminate in finite time.

In order to mitigate this issue, we derive a further stopping criterion after the expected improvement algorithm has terminated. Namely, we set a threshold $\mathfrak{s}$ such that for n_i^1,

$$d_i^1 - 3\sigma_i^1 > d^- \cup \sigma_i^1 < \mathfrak{s}\,.$$

This ensures we have confidence in either the value of d_i^1 or confidence that further evaluation of d_i^1 will not result in $d^- = d_i^1$. If any $d_i^1 \in \mathbf{d}^1$ do not meet this criteria, we evaluate these points again and restart the expected improvement algorithm. This process is detailed in algorithm 3.

Algorithm 3: Repetitive emulation OHS estimation; minimum cost improvement τ

```
Run algorithm 2 ;
Let 𝔫 = {n_i^1 ∈ n^1 : d_i^1 − 3σ_i^1 > d^− ∪ σ_i^1 < 𝔰} ;
if 𝔫 ≠ ∅ then
    for ñ ∈ 𝔫 do
        Estimate d(ñ) ≈ k_2(ñ) ;
        n ← (n ∪ ñ); d ← (d ∪ d{ñ)} ;
        Coalesce n, d into n^1, d^1 and obtain σ^1 ;
    end
    Re-estimate functions μ(n), Ψ(n), EI(n), with Θ = Θ(n^1, d^1, σ^1) ;
    Return to step 1 ;
end
return n_*^final = arg min_{n_i ∈ n^1} {d_i^1}
```

S7.6 Extensions

Various extensions of the emulator may improve our surrogate of the loss function, for example specifying priors on the parameters $\theta, \sigma_u^2, \zeta$ and using the likelihood provided by the Gaussian process to marginalize out these parameters. An explicit approach is given in Andrianakis and Challenor (2011), but under linearity assumptions which do not hold in our case, so analytic tractability would be lost. If we were able to cheaply estimate the derivative of the cost function at design points, this could be incorporated into our emulator (Killeya, 2004), enabling greater posterior accuracy around these points. Direct estimation of gradients from only estimates of $\ell(n)$ usually requires double the number of evaluations as estimation of $\ell(n)$ values, and so has the potential to become a more costly procedure than the method presented in section 5.3.

S8 Simulations

S8.1 Simulation of holdout, naive updating, and no-update strategies

We simulated a population of 2×10^5 samples at 50 timepoints, with ten timepoints per epoch (time between updates). We considered a risk score on 22 'visible' covariates similar to those of the ASPRE score Rolnik et al. (2017a) with true risk also depending on a 'latent' covariate not included in the risk score. We designated the true risk function f_t as a logistic model with coefficients varying continuously as a Gaussian process of t. We also used a logistic regression model for fitting all risk scores. At each time point, we computed risk scores using each method, and made interventions on the 10% of samples with highest predicted risk by reducing values of visible and latent covariates. We defined total cost as the sum of post-intervention risk across all samples. Hold-out sets were used in the final time-point of each epoch.

We included an 'alternative' updating strategy in which we recorded a binary indicator of whether an individual sample underwent a risk-score guided intervention (which, as per Supplement S3.2, is not always possible). When making a treatment decision for an individual, we set the value of this treatment indicator as a constant value.

S8.2 Optimal holdout set size arising from a simulated example

In this section, we analyse the dynamics of a roughly realistic, binary outcome system, subject to predictions from different families of risk models. Our main aim is to demonstrate the natural emergence of an optimal holdout set size from a reasonable setting.

We generated datasets with a population size $N = 5000$ with seven standard normally distributed covariates and outcomes Y under a ground-truth logistic model, either with interaction terms (i.e., non-linear) or without (linear). We considered risk scores ρ derived from either logistic regression models (not including interaction terms) or random forests. We designated cost functions C_1, C_2 to have value 0 for true-negatives, 0.5 for false- or true- positives, and 1 for false-negatives.

Supplementary Figure S6 shows simulation results using either linear or logistic prediction models and linear or non-linear underlying models for $Y \mid X$. We can observe that an optimal holdout set size can arise naturally from standard predictive models, since empirical k_2 curves for both a random forest and logistic regression satisfy assumptions B2 and B4 in the main manuscript. The optimal holdout set size occurs at a value n smaller than that at which $k_2(n)$ is nearly 'flat', indicating that unnecessarily large training sets are suboptimal. However, since $\ell(n)$ rises only linearly as n increases, it is generally less costly to slightly overestimate rather than underestimate the optimal holdout set size. Finally, the rightmost panels illustrate that the optimal holdout set size is not necessarily smaller for a more accurate model: the random forest model (non-lin ρ) in the non-linear underlying case (right panels) leads to uniformly lower expected costs $k_2(n)$ at all potential holdout set sizes, although the optimal holdout set size is larger.

S9 Optimal holdout size in ASPRE

S9.1 Implementation

We implemented the complete ASPRE model as described in Rolnik et al. (2017b). We simulated a population of individuals with a similar distribution of ASPRE model covariates. We computed the ASPRE scores for our simulated individuals, and found a linear transformation of these scores such that, should the scores exactly specify the probability of PRE, the expected population prevalence and sensitivity of the score would match those reported in Rolnik et al. (2017a): prevalence π_{PRE}, and sensitivity amongst 10% highest scores: 12.3%. We then simulated PRE incidence according to these transformed scores.

We found that a generalised linear model with logistic link performed almost as well as the ASPRE score on our simulated data, so we used this model type to estimate the learning curve in the interests of simplicity.

To choose values $\mathbf{n}$ and $\mathbf{k_2}/\mathbf{d}$, we initially chose a set $\mathbf{n}$ of 20 random values from $[500, 30000]$. For each size n in $\mathbf{n}$, we took a random sample of our data of size n, fitted a logistic model to that sample, and estimated corresponding expected costs per individual $\mathbf{k_2}$ as above. We fitted values $\theta = \theta(\mathbf{n}, \mathbf{k_2}) = (a, b, c)$ parametrising k_2 as the maximum-likelihood estimator of θ under the model

$$(\mathbf{k_2})_i \sim N\left(k_2((\mathbf{n})_i, \theta), \sigma^2\right) \sim N(a(\mathbf{n})_i^{-b} + c, \sigma^2)$$

for a fixed values σ, noting that the estimate of θ is independent of σ. For the parametric algorithm, we then set all values of $\boldsymbol{\sigma}$ to the same value, chosen empirically as the sample variance of

$$\mathbf{k_2} - k_2(\mathbf{n}, \theta(\mathbf{n}, \mathbf{k_2})) \tag{59}$$

. For the emulation algorithm, we set values $\mathbf{d}$ as

$$\mathbf{d}_i = k_1(\mathbf{n})_i + (\mathbf{k_2})_i(N - (\mathbf{n})_i)\,,$$

transforming values $\boldsymbol{\sigma}$ correspondingly for use in the emulation algorithm. We then sequentially chose 100 additional values $\mathbf{n}$ using both algorithm 1 and 2, setting $\boldsymbol{\sigma}$ as the same value found in S9.1. After choosing the 120 values of $\mathbf{n}$ using algorithm 1, we re-estimated $\mathbf{k_2}/\mathbf{d}$ for each of these values before estimating the OHS and confidence interval to avoid any potential regression-to-the mean effects from choosing next-values-of-n so as to minimise estimated confidence interval width.

Our complete pipeline is available at `https://github.com/jamesliley/OptHoldoutSize_pipelines`, and a comprehensive vignette is included in our R package `OptHoldoutSize` on CRAN and at `https://github.com/jamesliley/OptHoldoutSize`.

S10 Supplementary figures

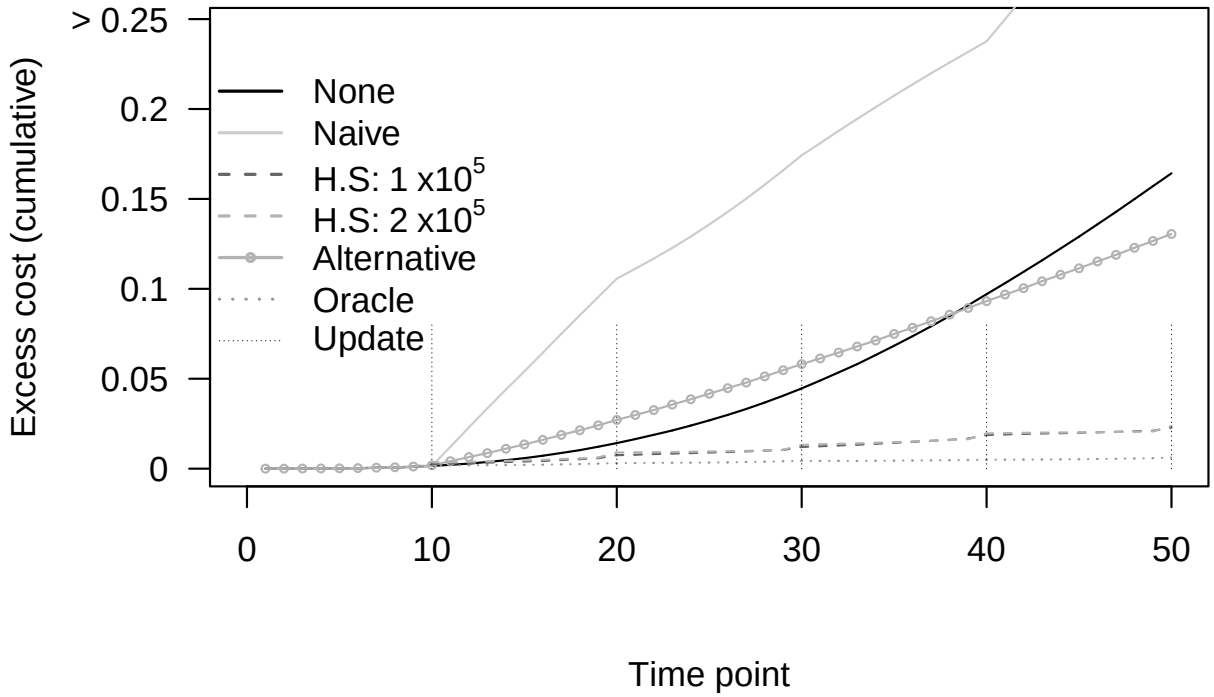


Figure S1: Cumulative costs of updating strategies, defined analogously to figure 2

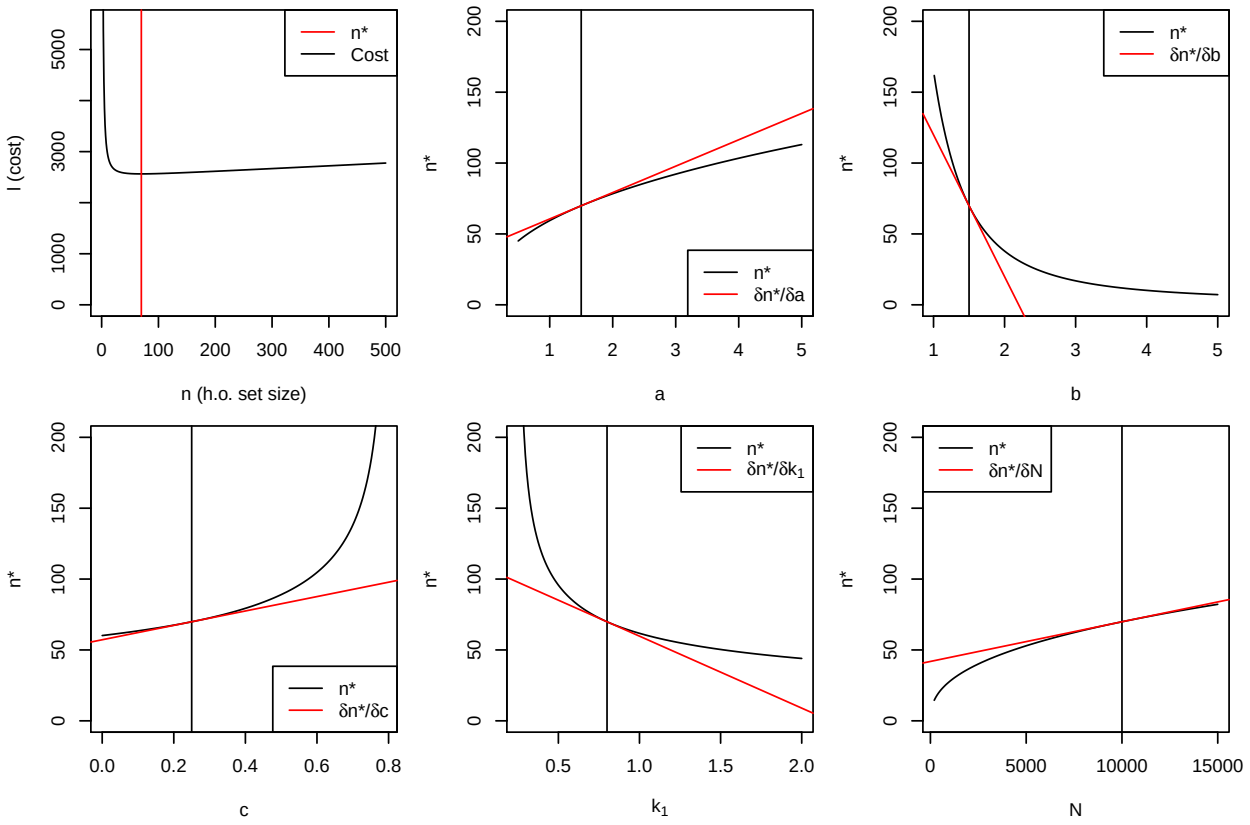


Figure S2: Dependence of optimal holdout set size on parameters of estimated learning curve $(a, b, c$, with $k_2(n; a, b, c) = an^{-b} + c)$, cost in intervention set k_1, and total number of samples N. Figures show change in optimal holdout set size n_* while varying one parameter and holding others constant at $(a, b, c) = \left(\frac{3}{2}, \frac{3}{2}, \frac{1}{4}\right)$, $k_1 = \frac{4}{5}$, $N = 10^4$.

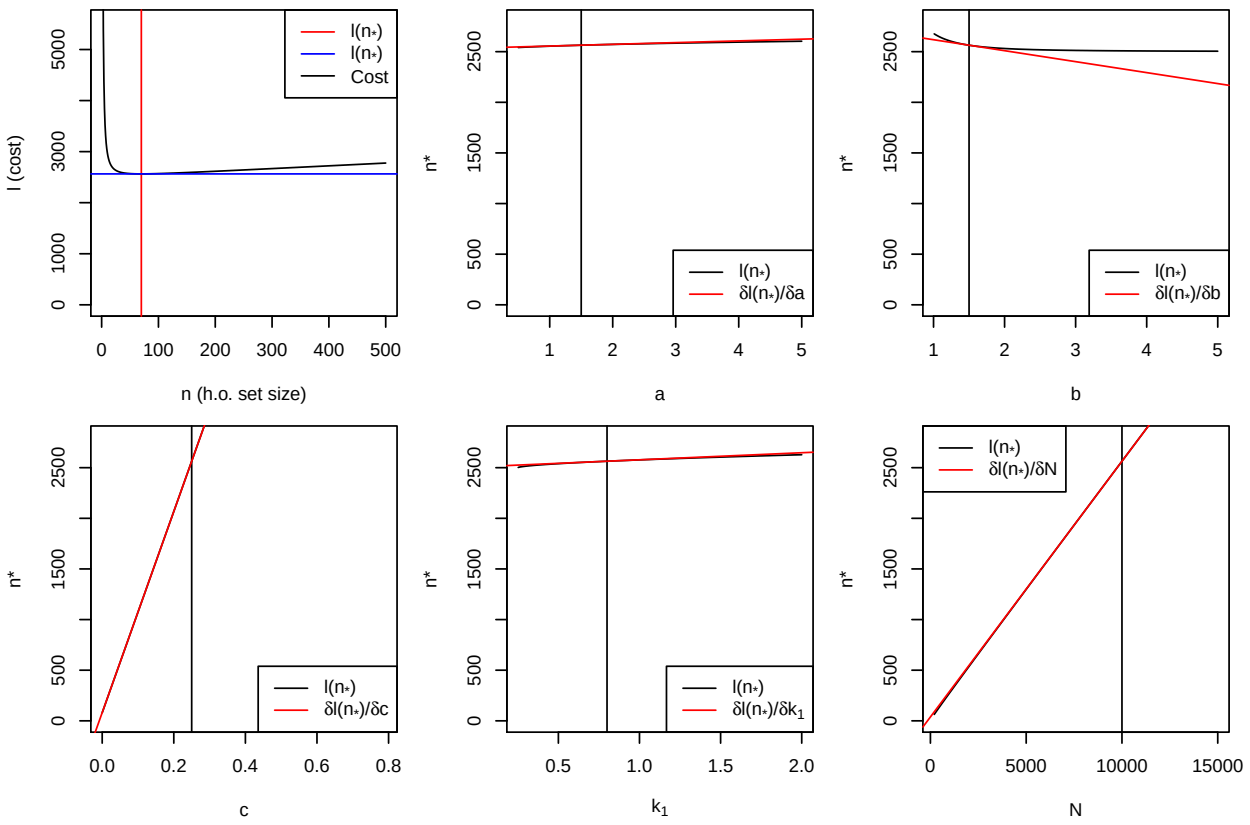
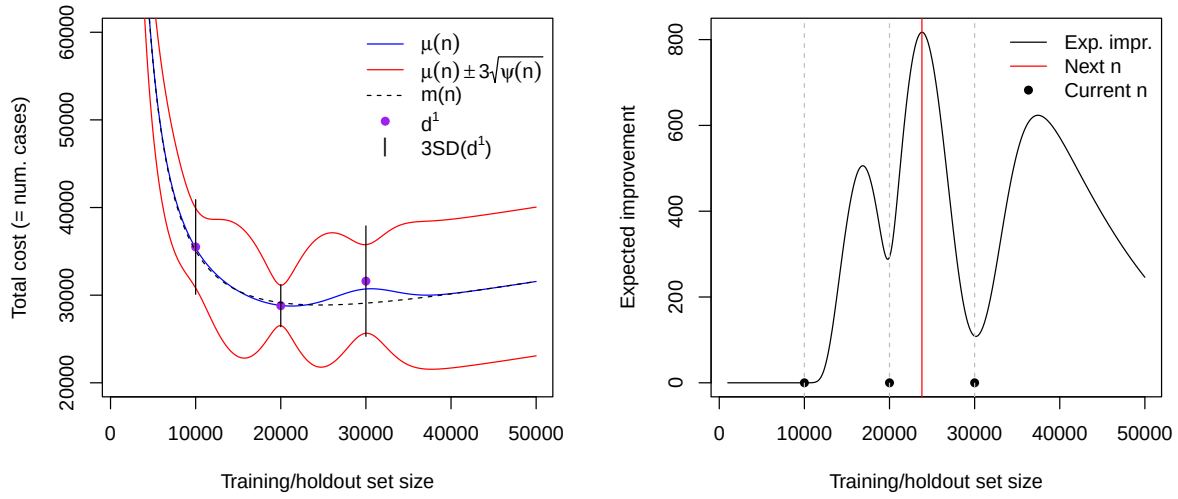
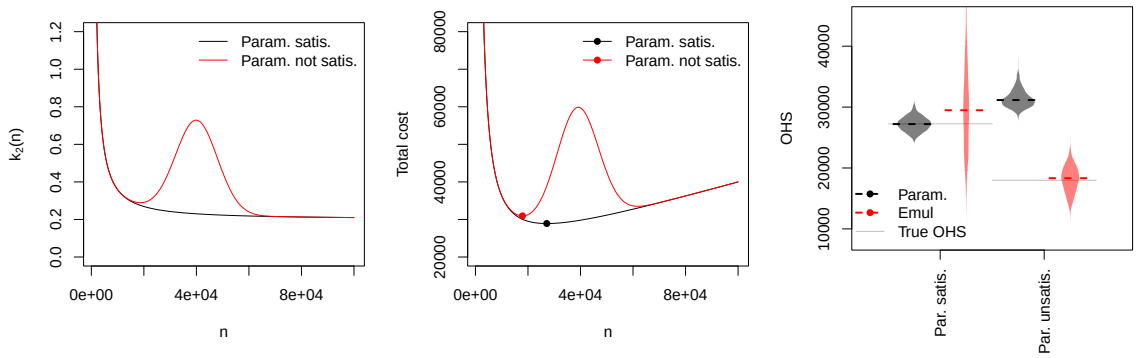


Figure S3: Dependence of minimum total cost on parameters of estimated learning curve $(a, b, c$, with $k_2(n; a, b, c) = an^{-b} + c)$, cost in intervention set k_1, and total number of samples N. Figures show change in minimal cost $\ell(n_*)$ while varying one parameter and holding others constant at $(a, b, c) = \left(\frac{3}{2}, \frac{3}{2}, \frac{1}{4}\right)$, $k_1 = \frac{4}{5}$, $N = 10^4$.



(a) (b)

Figure S4: Left panel shows emulator constructed using three $k_2()$ values (see pipelines). Function $m(n, \Theta)$ is constructed using θ derived from these three $k_2()$ estimates. Note reduced pointwise posterior variance at sample points. Rightmost panel shows expected improvement plot for the emulator constructed in panelS4a. Note local minima at existing sample points.



(a) Versions of k_2 (b) Costs associated with k_2 (c) Distr/meds of est. optimal holdout set size

Figure S5: Parametric and emulation algorithms with parametric assumptions satisfied or unsatisfied. OHS: optimal holdout set size

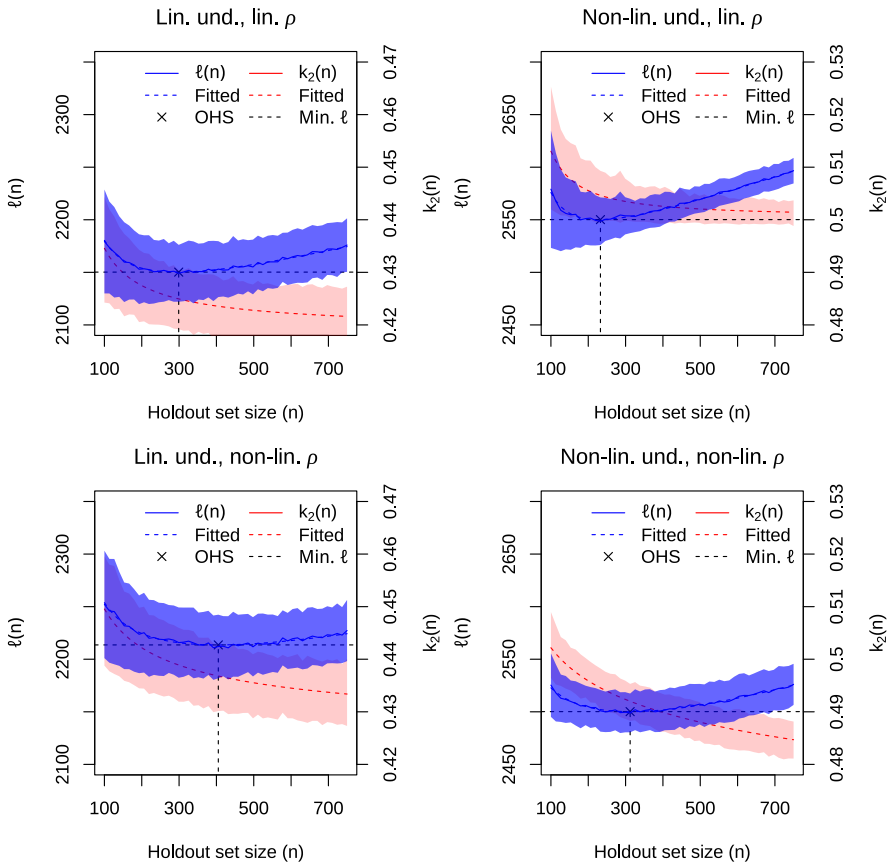


Figure S6: Examples of cost functions as per Theorem 4.1 arising naturally from a basic risk score, with varying underlying model (und.), risk score type (ρ) and one point-wise standard deviation (shaded regions). The contributions of terms $k_1 n$ to $\ell(n)$ depend only on the underlying model and are the same in each column. OHS: optimal holdout set size

References

Y. Andrianakis and P. Challenor. Parameter estimation for Gaussian process emulators, 2011.

R. G. Bower, M. Goldstein, and I. Vernon. Galaxy formation: a Bayesian uncertainty analysis. *Bayesian analysis*, 5(4):619–669, 2010.

E. Brochu, V. M. Cora, and N. De Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

A. D. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(10), 2011.

M. R. Killeya. *Thinking inside the box: using derivatives to improve Bayesian black box emulation of computer simulators with application to compart mental models*. PhD thesis, Durham University, Durham, UK, 2004.

J. Liley, G. Bohner, S. R. Emerson, B. A. Mateen, K. Borland, D. Carr, S. Heald, S. D. Oduro, J. Ireland, K. Moffat, R. Porteous, S. Riddell, S. Rogers, N. Cunningham, C. Holmes, K. Payne, S. J. Vollmer, C. A. Vallejos, and L. J. M. Aslett. Development and assessment of a machine learning tool for predicting emergency admission in Scotland. *npj Digital Medicine*, page (to appear), 2024.

M. Locatelli. Bayesian algorithms for one-dimensional global optimization. *Journal of Global Optimization*, 10(1):57–76, 1997.

D. L. Rolnik, D. Wright, L. Poon, A. Syngelaki, N. O'Gorman, C. de Paco Matallana, R. Akolekar, S. Cicero, D. Janga, M. Singh, et al. ASPRE trial: performance of screening for preterm pre-eclampsia. *Ultrasound in obstetrics & gynecology*, 50(4):492–495, 2017a.

D. L. Rolnik, D. Wright, L. C. Poon, N. O'Gorman, A. Syngelaki, C. de Paco Matallana, R. Akolekar, S. Cicero, D. Janga, M. Singh, et al. Aspirin versus placebo in pregnancies at high risk for preterm preeclampsia. *New England Journal of Medicine*, 377(7):613–622, 2017b.

I. O. Ryzhov. On the convergence rates of expected improvement methods. *Operations Research*, 64(6):1515–1528, 2016.

E. Vazquez and J. Bect. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and inference*, 140(11):3088–3095, 2010.

I. Vernon, J. Liu, M. Goldstein, J. Rowe, J. Topping, and K. Lindsey. Bayesian uncertainty analysis for complex systems biology models: emulation, global parameter searches and evaluation of gene functions. *BMC systems biology*, 12(1):1–29, 2018.

T. Viering and M. Loog. The shape of learning curves: a review. *arXiv preprint arXiv:2103.10948*, 2021.