Optimal sizing of a holdout set for safe predictive model updating

Sami Haidar-Wehbe¹,⊥, Samuel R. Emerson¹,⊥, Louis J. M. Aslett¹,², James Liley¹

¹: Department of Mathematical Sciences, Durham University, UK
²: Alan Turing Institute, London, UK
⊥: Equal contribution

Correspondence to {louis.aslett, james.liley}@durham.ac.uk

July 11, 2022

Abstract

Predictive risk scores are increasingly used to guide clinical or other interventions in complex settings, particularly healthcare. Directly updating a risk score used to guide interventions leads to biased risk estimates. We propose updating using a ‘holdout set’ – a subset of the population that does not receive risk-score-guided interventions – to prevent this. Since samples in the holdout set do not benefit from risk predictions, its size must trade off performance of the updated risk score whilst minimising the number of held out samples. We prove that this approach outperforms simple alternatives, and by defining a general loss function describe conditions under which an optimal holdout size (OHS) can be readily identified. We introduce parametric and semi-parametric algorithms for OHS estimation and demonstrate their use on a recent risk score for pre-eclampsia. Based on these results, we argue that a holdout set is a safe, viable and easily implemented means to safely update predictive risk scores.

Holdout set; Machine learning in healthcare; Machine learning safety; Model update; Predictive model; Treatment decision.

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 \( Y \) typically represents 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 \( Y \). For example, the QRISK3 score predicts thromboembolic risk given predictors including age and hypertension (Hippisley-Cox et al., 2017), and a high score may prompt prescription of antihypertensives.
Risk scores are typically developed by regressing observations of \( Y \) on \( X \). Should the distribution of \((X,Y)\) 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.

Updating of the risk score will involve obtaining new observations of \((X,Y)\). Crucially the distribution of \((X,Y)\) may also change 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 difficult or impossible to infer or measure, and indeed the fact that intervention took place may be unrecorded. In the QRISK3 example above, individuals prescribed antihypertensives in response to higher QRISK3 scores should have lower thromboembolic risk than they would have if QRISK3 was not used. Should a new risk score be fitted to observed \((X,Y)\), the effect of hypertension on risk would be underestimated, and overall risk estimates upwards-biased. This bias is worsened by heavier intervention resulting in risk scores becoming ‘victims of their own success’ (Lenert et al., 2019). This framework of directly updating a risk score on an ‘intervened’ 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., 2021b) when it is not.

In a previous paper 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 (Liley et al., 2021b). In this work, we formally develop this proposal for practical use in real-world predictive risk score updating and prove its suitability. We address a vital tension in the choice of an optimal 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.

Our work is methodological, and pre-empts a class of problems rather than analysing an existing risk score. We show that the updating paradox above is inevitable for risk scores on complex systems intended to guide interventions. The holdout set approach provides a simple method to facilitate ongoing use of such risk scores in the presence of both drift and intervention effects.

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 USA 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.

Our paper is structured in the following way. In Section 2, we review relevant literature and more precisely define the problem. We then give a motivating example. In Section 4, we demonstrate dominance of the holdout-set approach over simpler approaches, and construct an optimisation problem to find an OHS. In Section 5, we describe two algorithms for OHS
estimation, using explicit parametrisation 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.

2 Review of related work

Widespread collection of electronic health records has spurred development of new diagnostic and prognostic risk scores (Cook and Collins, 2015; Liley et al., 2021a), 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 USA 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 et al., 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).

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. 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 for successive predictions to converge to a stable setting where they ‘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 et al. (2021). These approaches seek to minimise a ‘performative’ loss to the population in the presence of an arbitrary risk score. Our alternative 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 ‘stability’ is not necessarily desirable in terms of 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 mild intervention after which true risk remained at 30%, regardless of treatment cost.
We found no 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.

OHS estimation requires quantification of expected material costs when using risk scores trained to data of various sizes. Such costs will typically depend on the error of risk predictions. The relation of predictive error to training set size is known as the ‘learning curve’, which can sometimes be accurately parametrised (Amari, 1993). A recent review paper suggests a power-law is accurate for simple models (Viering and Loog, 2021).

3 Motivating example

In this 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. 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., 2017b), 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., 2017a). We will not differentiate early- and late-stage PRE.

Due to changing population demographics, the influence of risk factors is likely to change over time, and the ASPRE score will need to be periodically updated. As discussed, a naive re-fitting of a risk score on the basis of (X : maternal assessment in early pregnancy and (Y : eventual PRE incidence could lead to dangerous underestimation of PRE risk, due to individuals previously assessed as high risk being treated in response to the assessment.

This could be avoided by maintaining a held-out 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., 2017a). However, an inappropriately small hold-out set would lead to an inaccurate updated model, reducing the benefit of future use of the score.

Our first question is whether a hold-out set is worth the cost, as opposed to simply continuing to use the existing score or updating naively. We will show that, as long as drift and intervention effects occur, the cost is generally justified. A second question is, if implemented, how large should a holdout set be? We will address this question in
4 Theory

4.1 General setup

Our general strategy for safe model updating using a hold-out set is illustrated as a causal graph in Figure 1. Each of the three columns is called an ‘epoch’ (0,1,2 in subscripts), representing a period of time in which a risk score is deployed and data gathered to update the risk score. Epochs correspond to consecutive intervals of continuous time, where epoch 0 corresponds to interval \([0, 1)\), epoch 1 to \([1, 2)\) and so on. Ellipses containing \(X\) or \(Y\) are covariates and outcomes (respectively) of populations of samples. We use the shorthand \(\{X^h\} \) and \(\{X^i\} \) (holdout and intervention respectively) to mean sets of samples from a random process \(X_t\), with \(t \in [e, e+1)\) and \(X_t \sim \mu_t\), where \(\mu_t\) may change with \(t\) but has constant support (e.g. \(R^p\)), and \(\{Y^h\} \) and \(\{Y^i\} \) to be corresponding set of observations of random processes \(Y^h_t\), \(Y^i_t\) representing outcomes. We leave the distributions of \(Y^h_t|X_t\) and \(Y^i_t|X_t\) unspecified for the moment. We presume that the times \(t\) corresponding to \(\{X^i\} \cup \{X^h\}\) are uniformly distributed on \([e, e+1)\) but will not presume this for \(\{X^i\}\) or \(\{X^h\}\) individually.

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 \(\rho_0\) (leftmost epoch). Once \(\rho_0\) is in use in the intervention set in epoch 1 (ellipses \(\{X^i\}, \{Y^i\}\)), a second causal pathway through \(\rho_0\) is established from \(\{X^i\}\) to \(\{Y^i\}\), but there remains only one causal pathway from \(\{X^h\}\) to \(\{Y^h\}\) in the holdout set (middle epoch). The updating process can be continued rightwards (\(\rho_1, \rho_2, \ldots\)).

To demonstrate why we opt to use a holdout set approach, we will define two functions dependent on \(t\). We define \(f_t(x)\) as the probability of event \(Y_t = 1\) given \(X_t = x\) if no risk score is in place, and \(g_t(x; \rho)\) as the probability of event \(Y_t = 1\) given \(X_t = x\) when a risk score is in place.
score \( \rho \) is used. In the holdout set schema in figure 1, we have \( P(Y_t^h = 1) = f_t(X_t^h) \) and \( P(Y_t^i = 1) = g_t(X_t^i, \rho_{e-1}) \) for \( t \in [e, e+1) \).

We generally wish to estimate \( f_t \) rather than \( g_t \), since \( f_t(x) \) gives a risk of the event in question under standard practice. An agent may opt to intervene in addition to standard practice if \( f_t(x) \) is high. The presence of drift is tantamount to an assumption that \( f_t \) changes with \( t \), and indeed we will assume that it changes Lipschitz-continuously. Change in \( f_t \) with \( t \) gives a motivation for updating a risk score at all, and continuous change underlies the assumption that a risk score will still be useful for a period of time after it is fitted. Drift will occur in many realistic settings (Tsybakov, 2004). The presence of intervention effects asserts that \( f_t \) and \( g_t \) differ; or rather, that some action may be taken in response to the risk score. Since there is little reason to use a risk score otherwise, we may assume intervention effects are typically present.

Suppose we have a risk score \( \rho_0 \) fitted at \( t \approx 0 \) which approximates \( f_0 \). We consider several options for what to do at \( t = 1 \): firstly, continue using our score \( \rho_0 \) (no update); secondly, refit a risk score \( \rho_1^i \) to observations \( X_1, Y_1 \) with \( X_1 \sim \mu_1 \) and \( Y_1|X_1 \sim g_1(X_1; \rho_0) \) and use the new risk score at \( t = 1 \) (naive update) or thirdly, between \( t = 0 \) and \( t = 1 \), leave out a ‘holdout’ set of samples and refit a new risk score \( \rho_1^h \) to these samples (holdout-update). If using the no-update option, performance would be poor as \( \rho_0 \approx f_0 \approx f_1 \) (due to drift). If using the naive-update option, performance would be poor as \( \rho_1^i \approx g_1(\cdot; \rho_0) \approx f_1 \) (due to drift). Refitting a risk score at all, and continuous change in \( \rho \) secondly, refit a risk score. If using the no-update option, performance would be poor as \( \rho_0 \approx f_0 \approx f_1 \) (due to drift). If using the naive-update option, performance would be poor as \( \rho_1^i \approx g_1(\cdot; \rho_0) \approx f_1 \) (due to drift). If using the holdout-update option, performance would be poor as \( \rho_1^h \approx g_1(\cdot; \rho_0) \approx f_1 \) (due to drift). Because of this, we can generally choose a finite holdout set size \(|\{X^h_t\}|\) above which the error in estimating \( f_2 \) will be less than even the asymptotic error of the no-update or naive-update options. We state and prove this observation formally in Supplement S2. We will briefly demonstrate this by simulation in section 6.1.

### 4.2 Cost specification

We consider the aggregate sets of samples \( \{X^h_t\}, \{Y^h_t\}\cup\{X^i_{e+1}\}, \{Y^i_{e+1}\}\), and denote the total number of samples as \( N \). We denote the \( n \) samples in the ‘holdout’ set \( \{X^h_t\}, \{Y^h_t\}\) as \( D_n \) (where ‘\( D \)’ indicates ‘data’). Since we are choosing \( \{X^h_t\}, \{Y^h_t\}\) to be as close in time to \( \{X^i_{e+1}\}, \{Y^i_{e+1}\}\) as possible, we will presume that we have

\[
X^h_e \sim \mu_{e+1} \\
X^i_{e+1} \sim \mu_{e+1}
\]

A risk score is fitted to \( D_n \) which approximates \( f_{e+1}(x) \) and is used in the intervention set \( \{X^i_{e+1}\}, \{Y^i_{e+1}\}\). We will presume that samples in \( D_n \) are pairwise independent, as are samples in \( \{X^i_{e+1}\}, \{Y^i_{e+1}\}\), 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
respectively, covering both the cost of potentially managing the event $Y = 1$ and any costs of intervention. Function $C_2(X; D_n)$ 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:

$$k_1 = \mathbb{E}_{X \sim \mu e+1, C_1} \{C_1(X)\} \quad k_2(n) = \mathbb{E}_{X \sim \mu e+1, C_2, D_n} \{C_2(X; D_n)\}$$

(1)

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 cost $\ell$ across the sample population as

$$\ell(n) = k_1 n + k_2(n)(N - n)$$

(2)

The meaning of the $\ell$ is contextual; for instance, in QRISK3 it may mean total number of deaths for a fixed healthcare budget.

### 4.3 Sufficient conditions for 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 1.** $k_1$ does not depend on $n$: in a medical context, treatment plans and outcomes for patients without risk scores do not depend on the number of such patients.

**Assumption 2.** $k_2(n)$ is monotonically decreasing in $n$: the more data available to train the risk score, the greater its clinical utility.

**Assumption 3.** There exists $M \in (0, N)$ such that $n \geq M \iff k_2(n) \leq k_1$: 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 4.** $E \{k_2(i + 1) - k_2(i)\} > E \{k_2(j + 1) - k_2(j)\}$ 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 2, $k_2''(n) > 0$, given assumption 4. This leads to the result that there exists an optimal size for the holdout set minimizing the expected total cost. The proof is given in Supplement S3.

**Theorem 1.** Suppose assumptions 1-4 hold. Then there exists an OHS $N_* \in \{1, \ldots, N-1\}$ with $N \in \mathbb{N}$, such that: $\ell(i) \geq \ell(j)$ for $0 < i < j < N_*$ and $\ell(i) \leq \ell(j)$ for $N_* < i < j < N$.
As an immediate consequence, we note that the OHS always exceeds the minimal training sample size required to match baseline treatment.

**Corollary 1.** *The value of $N_*$ always exceeds the value of $M$ in assumption 3, since if $N' < M$ we have $\ell(N') = k_1N' + k_2(N')(N - N') > k_1N' + k_1(N - N') = k_1N \geq \ell(N_*)$*

Consequently assumption 4 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 $k_1$.

### 4.4 Robustness to assumptions

The applicability of assumptions 1-4 in real world settings requires careful consideration. We address violations of assumptions 2 and 4 in section 5.3 and assumptions 1 and 4 here.

Assumption 1 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. 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 (Supplement S2).

If ethically appropriate, assumption 1 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 3 fails due to $k_2(0) \leq k_1$, we may show a weaker result (the presence of a non-trivial but potentially non-unique minimum loss) by replacing assumptions 2, 3 and 4 with:

**Assumption 5.** *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 2.** *Suppose assumptions 1 and 5 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 $i \in \{0, N\}$*

In a setting in which $k_1 > k_2(0)$ but one or more of assumptions 2-4 do not hold, we have
Theorem 3. Suppose assumption 1 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 \( i \in \{0, N\} \).

Both results are proved in Supplement S3.

In the setting where \( k_1 = k_2(0) \) and assumption 1 does not hold, it may be reasonable in some settings 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 in assumption 1.

5 Estimation of OHS

5.1 Estimation of \( k_2(n) \)

We are aiming to find a holdout set size which minimises costs during epoch \( e \geq 1 \) \((t \in [e, e+1])\) (noting that the holdout set will be used late in the epoch when \( t \approx e+1 \)) and must do this during epoch \( e-1 \) (when \( t < e \)). 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 \( X \sim \mu_{e-\epsilon} \approx \mu_e, Y \mid X \sim f_{e-\epsilon}(X) \approx f_e(X), \) with \( \epsilon \) small

where 2 is 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 drift may occur in \( \ell(n) \).

At time \( t = e-\epsilon \), we now need to estimate constants \( N, k_1, \) and the function \( k_2 \). 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); 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. The function \( k_2(\cdot) \) (equation 1) is more difficult to estimate, as it concerns quantifying costs of hypothetical risk scores. We may tractably estimate \( k_2(\cdot) \) by assuming that

\[
\mathbb{E}_{X \sim \mu_e, C_2} \{C_2(X; D_n)\} = \mathcal{L}\{\text{err}(\rho_{D_n})\}
\]

where \( \rho_{D_n} \) is a risk score fitted to \( D_n \), \( \text{err}(\cdot) \) is a measure of error, and \( \mathcal{L} \) is some nondecreasing function. We claim that in general circumstances we may take \( \mathcal{L}(\cdot) \) to be linear and \( \text{err}(\cdot) \) to be expected mean-squared error (MSE) or a similar general loss over \( X \sim \mu_e \). We derive this heuristically in Supplement S4 and derive expressions for \( k_2(n) \) directly in a specific case in section 6.3. Once \( \mathcal{L} \) is known, this allows \( k_2(n) \) to be estimated readily by establishing the ‘learning curve’ of the risk score using item 2.
Some direct estimates of $k_2(n)$ are necessary to determine $\mathcal{L}$. A direct 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 is potentially ethically tenuous and expensive. Simpler options include estimating the function $\mathcal{L}$ 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 $\mathcal{L}$, 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 estimated $\mathcal{L}$.

5.2 Parametric estimation of OHS

A natural algorithm for estimating the OHS is immediately suggested by Theorem 1: assume $k_2$ is known up to parameters $\theta$, and estimate $N$, $k_1$ and $\theta$ to estimate the OHS. Parameters $\theta$ 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 S5.

| Algorithm 1: Parametric OHS estimation overview |
|------------------------------------------------|
| 1. $n, k_2, \sigma^2 \leftarrow$ some initial values $n$ with $(k_2)_i \approx k_2(n_i)$, $(\sigma^2)_i = \text{var}(k_2(n_i))$; |
| 2. while $|n| < \text{total iterations}$ do |
| 3. Find best new value $\tilde{n}$ to add to $n$; |
| 4. Estimate $\hat{k}_2(\tilde{n}) \approx k_2(\tilde{n})$; |
| 5. $n \leftarrow (n \cup \tilde{n})$, $k_2 \leftarrow \{k_2 \cup \hat{k}_2(\tilde{n})\}$, $\sigma^2 \leftarrow \sigma^2 \cup \text{var}\{\hat{k}_2(\tilde{n})\}$; |
| 6. end |
| 7. return Re-estimate OHS $n_s^{\text{final}}$ from $n, k_2, \sigma$; |

Consistency of algorithm 3 depends on whether $n$ eventually contains enough elements of sufficient multiplicity to estimate $\Theta_0$ consistently. Sampling some positive proportion of values of $n$ randomly from $\{1, \ldots, N\}$ guarantees that the multiplicity of all $n \in n$ almost surely eventually exceeds any finite value, readily ensuring consistency. Finite-sample bias of $n_s^{\text{final}}$ depends on $\nabla_\Theta n_s$ and the variance of $\Theta$. See Supplementary Figure 6 for typical forms of $\nabla_\Theta n_s$. 

10
5.3 Semi-parametric (emulation) estimation of OHS

Parametrisation 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 et al., 2010). We quantify the uncertainty of the form for $k_2(n)$ through the construction of a Gaussian process emulator of $\ell$. 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 minimum of its posterior mean over $n$ to be our OHS estimate, efficiently choosing values of $n$ at which to estimate $\ell(n)$ using an ‘expected improvement’ function $EI(\cdot)$, which also provides a natural stopping criterion; if $EI(n) > \tau$ we roughly expect the minimum cost to decrease by at least $\tau$ from adding another estimate of $\ell(n)$ to our data. Our approximate procedure is given in algorithm 2. Detailed algorithm and proofs of consistency are in Supplement S6.

Algorithm 2: Emulation OHS estimation; minimum cost improvement $\tau$

1. $n, d, \sigma^2 \leftarrow$ some initial values $n$ with $(d)_i = d(n_i) \approx \ell(n), (\sigma^2)_i = \text{var}\{d(n_i)\}$;
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. $\tilde{n} \leftarrow \arg \max_{n \in \{1, \ldots, N\}} EI(n)$;
   5. Estimate $d(\tilde{n}) \approx k_2(\tilde{n})$;
   6. $n \leftarrow (n \cup \tilde{n}); d \leftarrow (d \cup d(\tilde{n})); \sigma^2 \leftarrow [\sigma^2 \cup \text{var}\{d(\tilde{n})\}]$;
   7. Re-estimate mean and variance of Gaussian process $\ell(n)$ and function $EI(n)$;
4. end
5. return $n_{final}^* = \arg \min_{n \in \{1, \ldots, N\}} \{\text{Mean of } \ell(n)\}$

6 Simulations

6.1 Simulation showing dominance of holdout set approach

We briefly demonstrate the theory described in section 4.1 using simulated data similar to our motivating example (details in Supplement S2.1). Figure 2 shows total cost of no-update (‘none’), naive-update (‘naive’) and holdout-update (‘H.S’) at two holdout sizes over time. As drift occurs in $f_t$, the total costs of the holdout-set approaches are lowest. Choice of 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 S7.
Figure 2: Loss over time of no-update, naive-update and holdout-update strategies.

6.2 Comparison of parametric and emulation algorithms

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

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

where: ‘p’/‘np’: ‘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 $1 \times 10^5$ and 0.4 respectively. For emulation, we use a kernel width $\zeta = 5000$ and variance $\sigma_u^2$ of $1 \times 10^7$.

The function $k_{2np}(n)$ exhibits ‘double-descent’ behaviour (Supplementary Figures 9a, 9b), which is possible for learning curves (Viering and Loog, 2021). This violates assumptions 2, 4.

We firstly show the distribution of estimates of OHS using our two algorithms when $k_2$ takes either form above. To fit $k_2$, we use $n$ as 200 values of $n$ randomly chosen from $\{1, \ldots, N\}$, with values $k_2$ independently sampled as $(k_2)_i \sim N(k_2(n_i), \sigma_i^2)$, where $\sigma \sim U(0.001, 0.02)$. Supplementary Figure 9c shows the distributions and medians of OHS estimates using the parametric and emulation algorithm in settings with parametric assumptions either satisfied or unsatisfied.

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 9b). Because the cost function is ‘flat’ around the minimum in the setting where parametric assumptions are satisfied (Supplementary Figure 9b), 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.

![Figure 3: Convergence rates with parametric (top) and emulation (bottom) algorithms, using either a random (black) or greedy (gray) method to select the next value of \( n \) to add to \( n \), with parametric assumptions satisfied (left) or unsatisfied (right). Simulations run for 200 datasets from underlying model. In larger panels, horizontal lines show true optimal holdout set (OHS) size; vertical lines indicate when \( \geq 2.5\% \) of runs return that OHS within a grid of size 1000. Smaller panels show root mean-square error between total costs from simulations and minimal total cost. Note variable axis scaling on left and right.](image)

We next examine the consequences of sampling the ‘next’ value of \( n \) greedily, using equation 31 as in algorithm 3 or \( EI \) as in algorithm 4, rather than randomly with \( \tilde{n} \) chosen uniformly in \( \{1, \ldots, N\} \), by comparing the rates of convergence of OHS estimates. Figure 3
show medians and a rough discrete kernel estimate of OHS estimates at various sizes of |n|, where n is generated by adding points \( \tilde{n} \) either randomly or systematically, after choosing the initial five values in n randomly.

Convergence is faster when ‘next points’ are picked systematically rather than randomly and when using parametric estimates (though these are biased and inconsistent for \( k_{2}^{\text{np}} \)). 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, that the non-parametric method shows bifurcation detecting both local minima whilst the parametric method converges to a mid-point which is far from optimal in terms of total costs.

### 6.3 Illustration in realistic setting

We now return to our motivating example. 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 are refitting ASPRE to use in a population of 5 million individuals, from which we have approximately 80,000 new pregnancies per year. Thus \( N \approx 400,000 \) (SE: 1500). See Supplement S8 for further details.

We presume a simple clinical action in which a fixed proportion \( \pi = 10\% \) of individuals at the highest assessed PRE risk are treated with aspirin. We assume that if untreated with aspirin, a proportion \( \pi_0 \) of individuals designated to be ‘low-risk’ (lowest 90%) will develop PRE, as will a proportion \( \pi_1 \) of individuals designated high-risk. Under current pre-ASPRE best-practice guidelines we have \( \pi_0 \approx 0.02 \) (SE 0.0009) and \( \pi_1 \approx 0.08 \) (SE 0.008) (see Supplement S8). Aspirin reduces PRE risk by approximately \( 1 - \alpha = 63\% \) (SE 0.09) (Rohnik et al., 2017b). 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

\[
k_1 = \pi_0(1 - \pi) + \pi_1\pi \alpha \approx 0.022
\]

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 can be used to estimate \( k_2(n) \) 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. Although \( k_2(n) \) is easy and fast to estimate here, in order to mimic a real example where such estimation is time consuming or costly we restrict ourselves to use only \( |n| = 120 \) values of n, determined using either algorithm 3 or 4. 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 10271 (90% CI 8103-12438), with minimum cost (expected cases over five years) of 8172. 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.

Figure 4: Estimation of cost functions (black lines), OHS (red dots) and error using parametric (middle) and emulation (left; blue lines $\mu(n) + 3\sqrt[3]{\Psi(n)}$) algorithms, and change in estimated OHS and error with number of sample points $|n|$ (right; red lines: parametric, blue/dashed lines: emulation). Note that the ‘best’ points (black dots) to optimize parametric estimation are spread-out to estimate $\theta$ well, but for emulation they are clustered for accurate locally approximation. Error measures for OHS in parametric and emulation algorithms (red/blue shaded respectively) have different meanings and are not comparable.

7 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 reasonable assumptions, 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 machine learning, which will become particularly important as risk scores come to be used to prompt intervention in real-world applications.

Other solutions to the model updating problem have been proposed, such as developing models that introduce missing causal connections (Alaa and van der Schaar, 2018; Sperrin et al., 2019) and using several predictive scores together (Liley, 2021). Such solutions tend to be difficult to implement: more comprehensive modelling generally requires some observation of the intervention, requiring more data over time, and parallel use of several risk scores is highly complex. With this in mind, holdout sets could prove valuable in updating strategies since, in principle, they can be applied in any setting. Availability of direct risk-score-naive data through the holdout set can also facilitates post-deployment maintenance and surveillance (Sperrin et al., 2019).

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. For the interested reader, see further discussion in Supplement S9.

Acknowledgements

SH’s contributions arose from an MSc dissertation for the MISCADA programme at Durham University. 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 is funded by the EPSRC doctoral training partnership at Durham University, grant reference EP/R513039/1. We thank Catalina Vallejos, Sebastian Vollmer and Bilal Mateen for helpful discussion.

Code availability

We have implemented all functionality described in this manuscript in an R package on CRAN (OptHoldoutSize) and at https://github.com/jamesliley/OptHoldoutSize. Pipelines to generate all results in this manuscript are available at https://github.com/jamesliley/OptHoldoutSize_pipelines.

References

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

Akolekar, Ranjit, Syngelaki, Argyro, Poon, Leona, Wright, David, and Nicolaides, Kypros H. Competing risks model in early screening for preeclampsia by biophysical and biochemical markers. Fetal diagnosis and therapy, 33(1):8–15 (2013).

Alaa, Ahmed M and van der Schaar, Mihaela. Autoprognosis: Automated clinical prognostic modeling via Bayesian optimization with structured kernel learning. arXiv preprint arXiv:1802.07207 (2018).

Amari, Shun-Ichi. A universal theorem on learning curves. Neural networks, 6(2):161–166 (1993).

Brochu, Eric, Cora, Vlad M, and De Freitas, Nando. 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).
Chalmers, John, Pullan, Mark, Fabri, Brian, McShane, James, Shaw, Matthew, Mediratta, Neeraj, and Poullis, Michael. Validation of EuroSCORE II in a modern cohort of patients undergoing cardiac surgery. *European Journal of Cardio-Thoracic Surgery*, 43(4):688–694 (2013).

Collins, Gary S, Reitsma, Johannes B, Altman, Douglas G, and Moons, Karel GM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Circulation*, 131(2):211–219 (2015).

Cook, J. A. and Collins, G. S. The rise of big clinical databases. *British Journal of Surgery*, 102(2):e93–e101 (2015). doi:10.1002/bjs.9723.

Drusvyatskiy, Dmitriy and Xiao, Lin. Stochastic optimization with decision-dependent distributions. *arXiv preprint arXiv:2011.11173* (2020).

Finlayson, Samuel G, Subbaswamy, Adarsh, Singh, Karandeep, Bowers, John, Kupke, Annabel, Zittrain, Jonathan, Kohane, Isaac S, and Saria, Suchi. The clinician and dataset shift in artificial intelligence. *The New England Journal of Medicine*, pages 283–286 (2020).

Hippisley-Cox, Julia, Coupland, Carol, and Brindle, Peter. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*, 357 (2017).

Izzo, Zachary, Zou, James, and Ying, Lexing. How to learn when data gradually reacts to your model. *arXiv preprint arXiv:2112.07042* (2021).

Kansagara, Devan, Englander, Honora, Salanitro, Amanda, Kagen, David, Theobald, Cecelia, Freeman, Michele, and Kripalani, Sunil. Risk prediction models for hospital readmission: a systematic review. *Jama*, 306(15):1688–1698 (2011).

Lenert, Matthew C, Matheny, Michael E, and Walsh, Colin G. Prognostic models will be victims of their own success, unless…. *Journal of the American Medical Informatics Association*, 26(12):1645–1650 (2019).

Li, Qiang and Wai, Hoi-To. State dependent performative prediction with stochastic approximation. *arXiv preprint arXiv:2110.00800* (2021).

Liley, James. Stacking interventions for equitable outcomes. *arXiv preprint arXiv:2110.04163* (2021).

Liley, James, Bohner, Gergo, Emerson, Samuel R, Mateen, Bilal A, Borland, Katie, Carr, David, Heald, Scott, Oduro, Samuel D, Ireland, Jill, Moffat, Keith, Porteous, Rachel, Riddell, Stephen, Rogers, Simon, Cunningham,
Nathan, Holmes, Chris, Payne, Katrina, Vollmer, Sebastian J, Vallejos, Catalina A, and Aslett, Louis JM. Development and assessment of a machine learning tool for predicting emergency admission in Scotland. *medRxiv* (2021a).

Liley, James, Emerson, Samuel R, Mateen, Bilal A, Vallejos, Catalina A, Aslett, Louis JM, and Vollmer, Sebastian J. Model updating after interventions paradoxically introduces bias. *AISTATS proceedings* (2021b).

Lu, Jie, Liu, Anjin, Dong, Fan, Gu, Feng, Gama, Joao, and Zhang, Guangquan. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363 (2018).

Mendler-Dünner, Celestine, Perdomo, Juan C, Zrnic, Tijana, and Hardt, Moritz. Stochastic optimization for performative prediction. *arXiv preprint arXiv:2006.06887* (2020).

Nashef, Samer AM, Roques, François, Sharples, Linda D, Nilsson, Johan, Smith, Christopher, Goldstone, Antony R, and Lockowandt, Ulf. EuroSCORE II. *European journal of cardio-thoracic surgery*, 41(4):734–745 (2012).

Perdomo, Juan, Zrnic, Tijana, Mendler-Dünner, Celestine, and Hardt, Moritz. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR (2020).

Rolnik, Daniel L, Wright, David, Poon, LCY, Syngelaki, Argyro, O’Gorman, Neil, de Paco Matallana, Catalina, Akolekar, Ranjit, Cicero, Simona, Janga, Deepa, Singh, Mandeep, et al. ASPRE trial: performance of screening for preterm pre-eclampsia. *Ultrasound in obstetrics & gynecology*, 50(4):492–495 (2017a).

Rolnik, Daniel L, Wright, David, Poon, Liona C, O’Gorman, Neil, Syngelaki, Argyro, de Paco Matallana, Catalina, Akolekar, Ranjit, Cicero, Simona, Janga, Deepa, Singh, Mandeep, et al. Aspirin versus placebo in pregnancies at high risk for preterm preeclampsia. *New England Journal of Medicine*, 377(7):613–622 (2017b).

Shahian, David M, Jacobs, Jeffrey P, Badhwar, Vinay, Kurlansky, Paul A, Furnary, Anthony P, Cleveland Jr, Joseph C, Lobdell, Kevin W, Vassileva, Christina, von Ballmoos, Moritz C Wyler, Thourani, Vinod H, et al. 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(5):1411–1418 (2018).

Sperrin, Matthew, Jenkins, David, Martin, Glen P, and Peek, Niels. Explicit causal reasoning is needed to prevent prognostic models being victims of their own
success. *Journal of the American Medical Informatics Association*, 26(12):1675–1676 (2019).

Stallard, Nigel, Miller, Frank, Day, Simon, Hee, Siew Wan, Madan, Jason, Zohar, Sarah, and Posch, Martin. Determination of the optimal sample size for a clinical trial accounting for the population size. *Biometrical Journal*, 59(4):609–625 (2017).

Topol, Eric J. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56 (2019).

Tsymbal, Alexey. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 106(2):58 (2004).

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

Viering, Tom and Loog, Marco. The shape of learning curves: a review. *arXiv preprint arXiv:2103.10948* (2021).

Wallace, Emma, Stuart, Ellen, Vaughan, Niall, Bennett, Kathleen, Fahey, Tom, and Smith, Susan M. Risk prediction models to predict emergency hospital admission in community-dwelling adults: a systematic review. *Medical care*, 52(8):751 (2014).
# Supplementary materials

## Table of Contents

| Section | Title | Page |
|---------|-------|------|
| S1      | General notation | 1    |
| S2      | Formal statement of dominance of holdout-set based updating | 2    |
| S2.1    | Simulation | 6    |
| S3      | Proof of Theorems 1, 2, and 3 | 7    |
| S4      | Estimation of $k_2(n)$ | 9    |
| S5      | Parametric OHS estimation | 12   |
| S5.1    | Explicit partial derivatives for $n^*$, $\ell$ with power-law parametrisation | 14   |
| S5.2    | Proof of Theorem S5 | 14   |
| S6      | Estimation of OHS by Bayesian Emulation | 17   |
| S6.1    | Emulation of cost function with nugget term | 19   |
| S6.2    | Proof of Theorem S6 | 21   |
| S6.3    | Proof of Theorem S7 | 21   |
| S6.4    | Proof of Theorem S8 | 22   |
| S7      | Optimal holdout set size arising from a simulated example | 26   |
| S8      | Estimation of parameters for optimal holdout size in ASPRE | 28   |
| S8.1    | Estimation of $N$ | 28   |
| S8.2    | Estimation of $k_1$ and $k_2$ | 28   |
| S8.3    | Implementation | 30   |
| S9      | Supplementary discussion | 32   |
| S10     | Supplementary figures | 33   |
S1 General notation

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.
S2 Formal statement of dominance of holdout-set based updating

As in section 4, we presume a random process \( X_t, t \in \mathbb{R}^+ \), where \( \{ X_t^h \} \cup \{ X_t \} \) is a set of instances of \( X_t \) for values of \( t \) in \([e,e+1)\). We presume \( X_t \) has distribution \( \mu_t \) where \( \mu_t \) has constant support \( \mathcal{X} \). We will assume \( \mu_t \) varies continuously, in a manner to be specified.

We consider two functions \( f_t : \mathcal{X} \to [0,1] \) and \( g_t(\cdot; \rho) : \mathcal{X} \to [0,1] \), which correspond to risk given a value of \( X_t \) without and with an existing risk score \( \rho \) in place respectively, as described in section 4. For our purposes in this section, \( \rho \) will always be \( \rho_0 \) (fitted at time \( t = 0 \)) so we will omit the dependence of \( g_t \) on \( \rho \). To avoid causal considerations, we will define two further Bernoulli random variables: \( Y_{f_t} \), with \( P(Y_{f_t} = 1 | X_{f_t} = x) = f_t(x) \) and \( Y_g \), with \( P(Y_g = 1 | X_{f_t} = x) = g_t(x) \).

We aim to estimate risk scores \( \rho \cdot : \mathcal{X} \to [0,1] \) which estimate \( f_t \) for \( t \) on some interval \([e,e+1)\). We will, for the moment, aim to estimate \( f_t \) for \( t \in [1,2) \). We presume that we already have:

1. A risk score \( \rho_0 \) which approximates \( f_0 \) to arbitrary accuracy
2. Arbitrarily many samples of \( X_t, Y^f_t \) with \( t \in [1 - \epsilon_1,1) \)
3. A total of \( m \) samples of \( X_t, Y^f_t \) with \( t \in [1 - \epsilon_1,1) \) (holdout set)

As in the main manuscript, considering the use of a risk score in time period \( t \in [1,2) \), we will call the use of \( \rho_0 \) a ‘no-update’ strategy, the use of a risk score \( \rho_{(n)} \) fitted to samples 2 a ‘naive-update’ strategy and the use of a risk score \( \rho_{(h)} \) fitted to samples 3 a ‘holdout-update’ strategy.

We denote the mean absolute error and mean square error of a risk score \( \rho \) at time \( t \) as

\[
\ell_t(\rho) = \int |\rho(x) - f_t(x)| d\mu_t
\]
\[
\ell^2_t(\rho) = \int (\rho(x) - f_t(x))^2 d\mu_t
\]

The following theorem establishes that under certain conditions both measures of error are lower for \( \rho_{(h)} \) than for \( \rho_0 \) or \( \rho_{(n)} \) when \( t \in [1,1+\delta) \) for some \( \delta > 0 \).

**Theorem S4.** Suppose that:

1. \( |f_0(x) - f_1(x)| \geq \gamma_1 \) for all \( x \in \mathcal{X}_1 \subseteq \mathcal{X} \) with \( P_{\mu_1}(\mathcal{X}_1) = \kappa_1 > 0 \) (non-negligible drift)
2. \( |g_1(\cdot) - f_1(x)| \geq \gamma_2 \) for all \( x \in \mathcal{X}_2 \subseteq \mathcal{X} \) and \( 0 \leq \epsilon < \epsilon_1 \) with \( P_{\mu_1}(\mathcal{X}_2) = \kappa_2 > 0 \) (non-negligible intervention)
3. \( f_t(x) \) is uniformly \( \alpha \)-Lipschitz continuous in \( t \) (continuous drift in \( Y^f_t|X \) )
4. \( \mu_t \) is \( \alpha_2 \)-Lipschitz continuous in \( t \) with respect to total variational distance distance (continuous drift in \( X \))

5. If a risk score \( \rho(x) \) is fitted to data \( D(n) \) consisting of \( n \) identically distributed samples from \( X_t, Y_t^f \) or \( X_t, Y_t^g \) then for any \( \epsilon > 0 \) we have \( \lim_{n \to \infty} P_{D(n)}(\sup_{x \in X} |f_t(x) - \rho(x)| < \epsilon) = 1 \) (risk score consistency)

Suppose \( \rho_1, \rho_2, \rho_3 \) are fitted as above. Then given any \( \epsilon_1, \epsilon_2 > 0 \) and some fixed \( \delta \) satisfying

\[
\delta < \min\left(\frac{\gamma_1 \kappa_1}{2\alpha + \gamma_1 \alpha_2}, \frac{\gamma_2 \kappa_2}{2\alpha + \gamma_2 \alpha_2}\right)
\]

we can find an \( m_1 \) such that for \( m \geq m_1 \) we have

\[
P_{D(m)}(\ell_t(\rho(h)) < \min(\ell_t(\rho_0), \ell_t(\rho(n)))) > 1 - \epsilon_2
\]

for \( 1 \leq t \leq 1 + \delta \). If \( \delta \) satisfies:

\[
\delta < \min\left(\sqrt{\frac{(2\alpha^2 + \alpha_2 \gamma_1^2)^2 + 4\alpha_2 \gamma_1^2 \kappa_1 - (2\alpha^2 + \alpha_2 \gamma_1^2)}{2\alpha^2}}, \sqrt{\frac{(2\alpha^2 + \alpha_2 \gamma_2^2)^2 + 4\alpha_2 \gamma_2^2 \kappa_2 - (2\alpha^2 + \alpha_2 \gamma_2^2)}{2\alpha^2}}\right)
\]

then we may find an \( m_2 \) such that for \( m \geq m_2 \)

\[
P_{D(m)}(\ell_t^2(\rho(h)) < \min(\ell_t^2(\rho_0), \ell_t^2(\rho(n)))) > 1 - \epsilon_2
\]

for \( 1 \leq t \leq 1 + \delta \).

Proof. From \( \alpha \)-uniform Lipschitz continuity of \( f_t \) (assumption 3) we have, for all \( x, \delta \):

\[
f_t(x) - \alpha \delta \leq f_{t+\delta}(x) \leq f_t(x) + \alpha \delta
\]

and from assumption 4 we have

\[
\left| \int_{\mathcal{X}_0} (d\mu_t - d\mu_{t+\delta}) \right| \leq \alpha_2 \delta
\]

From assumption 5 we have

\[
P\left( \sup_{x \in \mathcal{X}} |\rho(h)(x) - f_1(x)| < \epsilon \right) > 1 - q(\epsilon, m)
\]

where \( q(\epsilon, m) \) is a function which converges to 0 as \( m \to \infty \) for any \( \epsilon > 0 \).
Now for $\delta < \frac{\gamma_1 \kappa_1}{2\alpha + \gamma_1 \alpha_2}$ we have

$$
\ell_{1+\delta}(\rho_0) = \int |f_0(x) - f_{1+\delta}(x)| d\mu_{1+\delta}
\geq \int |f_0(x) - f_1(x)| d\mu_{1+\delta} - \int |f_{1+\delta}(x) - f_1(x)| d\mu_{1+\delta}
\geq \int \inf_{x \in X_1} |f_0(x) - f_1(x)| d\mu_{1+\delta} - \int \sup_{x \in X} |f_{1+\delta}(x) - f_1(x)| d\mu_{1+\delta}
\geq \int \gamma_1 d\mu_{1+\delta} - \alpha \delta
\geq \gamma_1 \left( \int d\mu_t - \alpha_2 \delta \right) - \alpha \delta
\geq \gamma_1 (\kappa_1 - \alpha_2 \delta) - \alpha \delta
> \alpha \delta\tag{12}
$$

For $\delta$ satisfying inequality 7 we have

$$
\ell_{2+\delta}(\rho_0) = \int (f_0(x) - f_{1+\delta}(x))^2 d\mu_{1+\delta}
= \int (f_0(x) - f_1(x))^2 d\mu_{1+\delta} + \int (f_{1+\delta}(x) - f_1(x))^2 d\mu_{1+\delta}
- \int (f_{1+\delta}(x) - f_1(x))(f_0(x) - f_1(x)) d\mu_{1+\delta}
\geq \int \inf_{x \in X_0} |f_0(x) - f_1(x)|^2 d\mu_{1+\delta} + \int \inf_{x \in X} |f_{1+\delta}(x) - f_1(x)|^2 d\mu_{1+\delta}
- 2 \left( \int (f_{1+\delta}(x) - f_1(x))^2 d\mu_{1+\delta} \int (f_0(x) - f_1(x))^2 d\mu_{1+\delta} \right)^{1/2}
\geq \gamma_1^2 \int d\mu_{1+\delta} + 0
- 2 \left( \int \sup_{x \in X} |f_{1+\delta}(x) - f_1(x)|^2 d\mu_{1+\delta} \int \sup_{x \in X} |f_0(x) - f_1(x)|^2 d\mu_{1+\delta} \right)^{1/2}
\geq \gamma_1^2 \left( \int d\mu_t - \alpha_2 \delta \right) - 2\sqrt{\alpha^2 \delta^2 - \alpha^2}
= \gamma_1^2 (\kappa_1 - \alpha_2 \delta) - 2\alpha^2 \delta
> \alpha^2 \delta^2\tag{13}
$$

Similarly for $\delta < \frac{\gamma_2 \kappa_2}{2\alpha + \gamma_1 \alpha_2}$ and $0 < \epsilon_1 \leq \epsilon$ we have

$$
\ell_{2+\delta}(\rho_2) \geq \gamma_2 (\kappa_2 - \alpha_2 \delta) - \alpha \delta > \alpha \delta\tag{14}
$$
and for $\delta$ satisfying inequality 7 we have
\[
\ell_{1+\delta}^2(\rho_2) \geq \gamma_2^2(\kappa_2 - \alpha_2\delta) - 2\alpha^2\delta > \alpha^2\delta^2
\] (15)

Looking at $\rho(h)$ we have
\[
\ell_{1+\delta}^1(\rho(h)) = \int |\rho(h)(x) - f_{1+\delta}(x)|d\mu_{t+\delta}
\]
\[
\leq \int |\rho(h)(x) - f_1(x)|d\mu_{t+\delta} + \int |f_{1+\delta}(x) - f_1(x)|d\mu_{t+\delta}
\]
\[
\leq \int \sup_{x \in \mathcal{X}} |\rho(h)(x) - f_1(x)|d\mu_{t+\delta} + \int \sup_{x \in \mathcal{X}} |f_{1+\delta}(x) - f_1(x)|d\mu_{t+\delta}
\]
\[
= \sup_{x \in \mathcal{X}} |\rho(h)(x) - f_1(x)| + \alpha\delta
\] (16)

and
\[
\ell_{1+\delta}^2(\rho(h)) = \int (\rho(h)(x) - f_{1+\delta}(x))^2d\mu_{t+\delta}
\]
\[
\leq \int \sup_{x \in \mathcal{X}} |\rho(h)(x) - f_{1+\delta}(x)|^2d\mu_{t+\delta}
\]
\[
\leq \int \left( \sup_{x \in \mathcal{X}} |\rho(h)(x) - f_1(x)| + \sup_{x \in \mathcal{X}} |f_{1+\delta}(x) - f_1(x)| \right)^2d\mu_{t+\delta}
\]
\[
\leq \left( \sup_{x \in \mathcal{X}} |\rho(h)(x) - f_1(x)| + \alpha\delta \right)^2
\] (17)

Given the a choice of $\delta$ satisfying inequality 5, we have
\[
\epsilon_3 \overset{\text{def}}{=} \min (\gamma_1 (\kappa_1 - \alpha_2\delta), \gamma_2 (\kappa_2 - \alpha_2\delta)) - 2\alpha\delta > 0
\] (18)

so we may select $m_1$ such that for $m \geq m_1$ we have
\[
P \left( \sup_{x \in \mathcal{X}} |\rho(h)(x) - f_1(x)| < \epsilon_3\delta \right) > 1 - \epsilon_2
\]
in which case, given inequalities 12, 14, and 16, expression 6 holds.

Likewise, given a choice of $\delta$ satisfying inequality 7, we have
\[
\epsilon_4 \overset{\text{def}}{=} \sqrt{\min (\gamma_1^2(\kappa_1 - \alpha_2\delta) - 2\alpha^2\delta, \gamma_2^2(\kappa_2 - \alpha_2\delta) - 2\alpha^2\delta)} - \alpha\delta > 0
\] (19)

so we may select $m_2$ such that for $m \geq m_2$ we have
\[
P \left( \sup_{x \in \mathcal{X}} |\rho(h)(x) - f_1(x)| < \epsilon_4 \right) > 1 - \epsilon_2
\]
in which case, given inequalities 13, 15, and 17, expression 8 holds

\[\square\]
Remark 1 (Robustness of assumptions). The assumptions of theorem S4 all fundamentally underlie usability of risk scores in general, with the exception of assumption 1, which is a well-established effect for many risk scores Tsymbal (2004); Zliobaite (2010).

Assumptions 3,4 state that ‘drift’ must be continuous with time. This is essentially because continuous drift implies that a risk score that is consistent at time t will still be useful at a time t+ε. Without such an assumption, risk scores are essentially unusable. We note that the conditional distribution of intervened outcomes Y^g|X need not be continuous. Obvious extensions to the theorem will give similar results for almost-continuous drift.

Assumption 2 essentially states that use of the risk score leads to some change in the risk of an outcome given covariates. Except in special circumstances, this is typically the very reason for a risk score to be fitted and used, so this assumption is generally justified.

Assumption 1 essentially states that a non-negligible change occurs over time in the distribution of the non-intervened outcome conditional on covariates; that is, should no risk score be in place, the risk to an individual sample with fixed covariates will vary non-negligibly over time for a positive proportion of samples. This assumption underlies the need to update risk scores; if essentially no drift occurs, a risk score need not be updated at all.

Assumption 2 may be weakened to |g_1(x) - f_1(x)| > γ_2 if uniform Lipschitz continuity of g_t in t can be assumed.

The processes µ_t and f_t, which govern drift processes, are not considered to be random, but the theorem trivially holds if either has random perturbations for which all elements of their respective domain remain constrained by the theorem assumptions.

S2.1 Simulation

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. 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.
S3 Proof of Theorems 1, 2, and 3

Theorem 1. Suppose assumptions 1-4 in the main manuscript hold. Then there exists a $N_s \in (0, N)$ with $N \in \mathbb{N}$, which we call the optimal holdout set size, such that:

\[
\ell(i) \geq \ell(j) \text{ for } 0 < i < j < N_s \\
\ell(i) \leq \ell(j) \text{ for } N_s < i < j < N
\]

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

\[
\frac{\partial}{\partial n} k_2(n) < 0 \quad (20)
\]

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

\[
\ell'(n) = \frac{\partial}{\partial n} \left( k_1 n + k_2(n)(N - n) \right) = k_1 + k_2'(n)(N - n) - k_2(n) \quad (21)
\]

\[
= (k_1 - k_2(n)) + k_2'(n)(N - n) \quad (22)
\]

By assumption 3 in the main manuscript, $k_1 < k_2(0)$, and, from equation 20, $k_2'(0) < 0$, so both terms in equation 23 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 3 in the main manuscript. We thus have $\ell'(N) > 0$. By assumption, $\ell$ is smooth, so by Bolzano’s Theorem, there must exist at least one point $n_s$ for which $\ell'(n_s) = 0$, which is an extremum of $\ell$.

We now prove that this extremum is unique and a minimum. First, by assumption 4 in the main manuscript, we may impose that

\[
\frac{\partial^2}{\partial n^2} k_2(n) > 0 \quad (24)
\]

Taking the second derivative of $\ell$

\[
\frac{\partial^2}{\partial n^2} \ell(n) = \frac{\partial^2}{\partial n^2} \left( k_1 n + k_2(n)(N - n) \right) = k_2''(n)(N - n) - 2k_2'(n) \quad (25)
\]

and using equations 20 and 24, we see that $\ell''(n)$ is strictly positive, and, as a consequence, $\ell'(n)$ is monotonically increasing. Therefore, the extremum of $\ell(n)$ at $n_s$ we found earlier is unique and, as $\ell''(n) > 0$, it is a minimum.

If $n_s \in 1..(N - 1)$, let $N_s = n_s$. If $n_s \notin \mathbb{N}$, let $N_s$ be the closest natural number to either side of $n_s$. From assumption 3 in the main manuscript, $N_s$ cannot be 0 or $N$. In both scenarios, this completes the proof. \qed
For the proofs of theorems 2 and 3, we will use the following lemma:

**Lemma S1.** Suppose assumption 1 holds and there exists $0 < M < N$ such that $k_2(M) < k_1$. Then $\ell(M) < \ell(N)$.

**Proof.**

\[
k_2(M) < k_1 \implies (N - M)k_2(M) < (N - M)k_1 \\
\implies (N - M)k_2(M) + M k_1 < N k_1 \\
\implies \ell(M) < \ell(N)
\]

We now have

**Theorem 2.** Suppose assumptions 1 and 5 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 5.

It immediately follows from assumption 5 that $k_2(M) < k_1$, so by Lemma S1 $\ell(M) < \ell(N)$. If $k_1 = k_2(0)$ then we are done as $\ell(N) = \ell(0)$. If $k_1 > k_2(0)$ then from assumption 5 we have

\[
k_1 - k_2(0) < \frac{N - M}{N}(k_1 - k_2(M)) \implies N(k_1 - k_2(0)) < (N - M)(k_1 - k_2(M)) \\
\implies N k_1 - N k_2(0) < N k_1 - (N - M)k_2(M) - M k_1 \\
\implies \ell(N) - \ell(0) < \ell(N) - \ell(M) \\
\implies \ell(M) < \ell(0)
\]

as needed.

**Theorem 3.** Suppose assumption 1 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

\[
k_1 < k_2(0) \implies N k_1 < N k_2(0) \implies \ell(N) < \ell(0)
\]

so all that is needed is $\ell(M) < \ell(N)$, which is given by Lemma S1.
S4 Estimation of $k_2(n)$

We argue in this section that $k_2(n)$ can generally be 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 S2 as the probability of $Y = 1$ given $X$ at time $t$ when no risk score is used). We will write $c_2(x, \rho) = E_{C_2}\{C_2(x; d_n)\}$, where $E_{C_2}$ indicates expectation over randomness in actual cost for a given sample with a given risk score, and $\rho$ 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 $\rho$, $f_t$, then a corresponding form for $k_2$ may be immediate. For instance, if one of

\[
c_2(x, \rho) = c^0 + c^1|\rho - f_t(x)| \\
c_2(x, \rho) = c^0 + c^1(\rho - f_t(x))^2 \\
c_2(x, \rho) = c^0 + c^1 f_t(x)(\rho - f_t(x))^2
\]

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

\[
E_{D_n}\{E_X[|\rho(X) - f_t(X)|]\} \\
E_{D_n}\{E_X[(\rho(X) - f_t(X))^2]\} \\
E_{D_n}\{E_X[f_t(X)(\rho(X) - f_t(X))^2]\}
\]

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 $\rho$, 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 $\rho$, and write

\[
c_2(x, \rho) = c_2(x, f_t(x)) + \frac{1}{2} \frac{\partial^2 c_2}{\partial \rho^2} (x, f_t(x)) (\rho - f_t(x))^2 + O((\rho - f_t(x))^3)
\]
noting that \( \frac{\partial^2 c_2}{\partial \rho^2} (x, f_t(x)) = 0 \) by assumption.

The value \( \frac{\partial^2 c_2}{\partial \rho^2} (x, f_t(x)) \) represents the curvature with respect to \( \rho \) 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

\[
E_X [c_2(X; \rho)] = E_X \left[ c_2(x, f_t(x)) + \frac{\partial^2 c_2}{\partial \rho^2} (x, f_t(x)) (\rho(X) - f_t(X))^2 + O \left( \sup_x (\rho(x) - f_t(x))^3 \right) \right] \\
\approx E_X [c_2(x, f_t(x))] + E_X \left[ \frac{\partial^2 c_2}{\partial \rho^2} (x, f_t(x)) \right] E_X [(\rho(X) - f_t(X))^2] + O \left( \sup_x (\rho(x) - f_t(x))^3 \right)
\]

def \( c_0 + c_2 \text{MSE}(\rho) + O \left( \sup_x (\rho(x) - f_t(x))^3 \right) \)

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

\[
k_2(n) = E_{D_n} \{ E_X [c_2(X; \rho)] \}
\approx c_0 + c_2 E_{D_n} \{ \text{MSE}(\rho) \}
\]

so \( k_2(n) \) is approximately linear in \( E_{D_n} \{ \text{MSE}(\rho) \} \).

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 \( \rho_0 \). 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 \( \gamma_i \), and that an event \( Y = 1 \) has a fixed cost \( \gamma_y \). Then

\[
c_2(x, \rho) = E_{C_2} \{ (\text{Cost of an event}) + (\text{Cost of intervening}) \}
\]

\[
= \begin{cases} 
\gamma_y f_t(x) + 0 & \text{if } \rho < \rho_0 \\
\gamma_y \alpha f_t(x) + \gamma_i & \text{if } \rho \geq \rho_0
\end{cases}
\]

(29)
disregarding potential baseline costs common to all samples. We may now apply the heuristic above more directly, by presuming that the threshold $\rho_0$ 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:

$$c_2(x, \rho) - c_2(x, f_t(x)) = \begin{cases} 
\gamma_y f_t(x) & \text{if } \rho < \rho_0, f_t(x) < \rho_0 \\
\gamma_y \alpha f_t(x) + \gamma_i - \gamma_y f_t(x) & \text{if } \rho \geq \rho_0, f_t(x) < \rho_0 \\
\gamma_y f_t(x) - (\gamma_y \alpha f_t(x) + \gamma_i) & \text{if } \rho < \rho_0, f_t(x) \geq \rho_0 \\
\gamma_y \alpha f_t(x) + \gamma_i & \text{if } \rho \geq \rho_0, f_t(x) \geq \rho_0
\end{cases}$$

$$= 1_{(\rho < \rho_0) \ XOR (f_t(x) < \rho_0)} |\gamma_y (\alpha - 1) f_t(x) + \gamma_i|$$

and, denoting $\delta_{\rho}(x) = (\rho(x) < \rho_0) \ XOR (f_t(x) < \rho_0)$, we have

$$k_2(n) = k_2(n) - \lim_{n \to \infty} k_2(n) + \lim_{n \to \infty} k_2(n)$$
$$= \lim_{n \to \infty} k_2(n) + \mathbb{E}_{D_n} \{ \mathbb{E}_X [c_2(X, \rho) - c_2(X, f_t)] \}$$
$$= c^0 + \mathbb{E}_{D_n} \{ \mathbb{E}_X [1_{\delta_{\rho}(X)} c^1 f_t(x) + c^2] \}$$

(30)

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.
S5  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 \( k'_2, k''_2, \ell' \) 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 \( \Theta \) is an unbiased estimate of \( \Theta_0 \), so \( \Theta_0 \) 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 \( \Theta_0 \).

**Theorem S5.** Assume that \( k'_2(n; \theta), k''_2(n; \theta) \) and \( \nabla_\theta k_2(n; \theta) \) are continuous in \( n \) and \( \theta \) in some neighbourhood of \((n_0, \Theta_0)\), and that \( \Theta_0 \) parametrizes a setting satisfying assumptions 1-4. Suppose that \( \Theta \) behaves as a mean of \( m \) appropriately-distributed samples in satisfying \( \sqrt{m}(\Theta - \Theta_0) \rightarrow N(0, \Sigma) \) in distribution where \( \Theta_0 \) does not depend on \( m \), that an estimate \( \hat{\Sigma} \) of \( \Sigma \) is available which is independent of \( \Theta \) and satisfies \( ||\hat{\Sigma} - \Sigma||_2 \rightarrow 0 \) in distribution, and that \( n_0 \) is finite and unique as above. Then denoting

\[
\beta_\Theta = \frac{\partial^2 \ell}{\partial n \partial \Theta_i} / \frac{\partial^2 \ell}{\partial n^2}, \quad \gamma_\Theta = \frac{\partial \ell}{\partial \Theta_l}
\]

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

\[
\sqrt{m}(n_* - n_0) \rightarrow N\left(0, \beta_{\Theta_0}^l \Sigma \beta_{\Theta_0}\right), \quad \sqrt{m} \{\ell(n_*) - \ell(n_0)\} \rightarrow N\left(0, \gamma_{\Theta_0}^l \Sigma \gamma_{\Theta_0}\right)
\]

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

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

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

The proof is given in Supplement S5.2 below. A consequence is that for sufficiently accurately estimated costs, the OHS will be a non-trivial size:

**Corollary S2.** Under assumptions of Theorems 1, S5, \( P\{1 < n_*(\Theta) < N\} \rightarrow 1 \) as \( m \rightarrow \infty \).

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 S5.1. Examples of variation in $n_*$ and $\ell(n_*)$ with a power-law form for $k_2$, are shown in Supplementary Figures 6, 7.

Note that confidence intervals must be interpreted with care: if the sampling distributions for $k_1$ and $\theta$ admit the possibility that assumptions of Theorem 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 $\Theta$ is estimated from a multiset $n$ of values in $\{1, \ldots, N\}$ and estimates $\text{d}$ of $k_2(n)$ for each $n \in n$ with known finite sampling variances $\sigma^2$. For certain multisets $n$, estimates of $\Theta$ will not converge; for instance, if $n$ contains only a single value repeated. The value $m$ in Theorem S5 should be interpreted as an 'effective' population size, such that $\sqrt{m} \{\Theta(n) - \Theta_0\} \to N(0, \Sigma)$ 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 $n$ of points at which estimates $k_2$ of $k_2(n)$ have been made already. We denote by $\Theta(n, k_2, \sigma)$, $\Sigma(n, k_2, \sigma)$ and $I_\alpha(n, k_2, \sigma)$ respectively the estimates of $\Theta_0$, $\lim_{m \to \infty} \text{var} \{\sqrt{m} \{\Theta(n, k_2, \sigma) - \Theta_0\}\}$ and the width of the confidence interval $I_\alpha \{\Theta(n, k_2, \sigma), \Sigma(n, k_2, \sigma)\}$. Suppose we have the option of estimating $d(n)$ for one value of $n \in \{1, \ldots, N\}$ with known variance $\text{var} \{d(n)\} = \sigma^2$. We select $\tilde{n}$ as:

$$\tilde{n} = \arg\min_n \mathbb{E}[d(n) \sim N(k_2(n), \sigma)] \left[ I_\alpha \left\{ n \cup n, k_2 \cup \hat{k}_2(n), \sigma \cup \sigma \right\} \right]$$

(31)

that is, `select the $\tilde{n}$ which will minimize the expected OHS confidence interval width if added to our set $n$, with expectation computed with respect to our current parameter estimates’ If no minimum exists, $\tilde{n}$ is selected uniformly from $1, \ldots, N$. Algorithm 3, an expanded version of algorithm 1 in the main manuscript, shows our full parametric estimation procedure.

**Algorithm 3:** Parametric OHS estimation; with $n_{\text{add}}$ estimates of $k_2(\cdot)$

1. $n, k_2, \sigma^2 \leftarrow$ some initial values $n$ with $(k_2)_i = \hat{k}_2(n_i) \approx k_2(n_i), (\sigma^2)_i = \text{var}(\hat{k}_2(n_i))$;
2. while $|n| < n_{\text{add}}$ do
   3. Find best new value $\tilde{n}$ to add to $n$ as per formula 31;
   4. Estimate $\hat{k}_2(\tilde{n}) \approx k_2(\tilde{n})$;
   5. $n \leftarrow (n \cup \tilde{n}), k_2 \leftarrow \{k_2 \cup \hat{k}_2(\tilde{n})\}, \sigma^2 \leftarrow \sigma^2 \cup \text{var} \{\hat{k}_2(\tilde{n})\}$;
5. end
6. Re-estimate OHS $n_{*\text{final}} = n_{*\text{final}} \{\Theta(n, k_2, \sigma)\}$;
7. return $n_{*\text{final}}$
S5.1 Explicit partial derivatives for $n^*$, $\ell$ with power-law parametrisation

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

$$k_2(n; \theta) = an^{-b} + c$$  \hspace{1cm} (32)

then we have

$$\frac{\partial n^*}{\partial a} = \frac{1}{a} \left( bNn^* - (b - 1)n^*_n \right)$$

$$\frac{\partial n^*}{\partial b} = \frac{Nn^*(b \log(n^*_n) - 1) - n^*_n^2 ((b - 1) \log(n^*_n) - 1)}{b(b + 1)N - b(b - 1)n^*_n}$$

$$\frac{\partial n^*}{\partial c} = \frac{1}{a} \left( \frac{n^*_n^{b+2}}{b(b + 1)N - b(b - 1)n^*_n} \right)$$

$$\frac{\partial n^*}{\partial k_1} = \frac{bn^*_n}{b(b + 1)N - b(b - 1)n^*_n}$$

and, more simply

$$\frac{\partial}{\partial a} \ell(n^*_n; \theta) = (N - n^*_n)n^*_n^{-b}$$

$$\frac{\partial}{\partial b} \ell(n^*_n; \theta) = - \log(n^*_n)(N - n^*_n)an^*_n^{-b}$$

$$\frac{\partial}{\partial c} \ell(n^*_n; \theta) = N - n^*_n$$

$$\frac{\partial}{\partial k_1} \ell(n^*_n; \theta) = n^*_n$$

$$\frac{\partial}{\partial N} \ell(n^*_n; \theta) = an^*_n^{-b} + c$$

S5.2 Proof of Theorem S5

**Theorem S5.** Assume that $k_2'(n; \theta)$, $k_2''(n; \theta)$ and $\nabla_\theta k_2(n; \theta)$ are continuous in $n$ and $\theta$ in some neighbourhood of $(n_0, \Theta_0)$, and that $\Theta_0$ parametrizes a setting satisfying assumptions 1-4. Suppose that $\Theta$ behaves as a mean of of $m$ appropriately-distributed samples in satisfying $\sqrt{m}(\Theta - \Theta_0) \to N(0, \Sigma)$ in distribution where $\Theta_0$ does not depend on $m$, that an estimate $\hat{\Sigma}$ of $\Sigma$ is available which is independent of $\Theta$ and satisfies $||\hat{\Sigma} - \Sigma||_2 \to 0$ in distribution, and that $n_0$ is finite and unique as above. Then denoting

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

14
we may uniquely define \( n_0 = \{ n : \ell'(n; \Theta_0) = 0 \} \) and we have
\[
\sqrt{m}(n_\ast - n_0) \rightarrow N(0, \beta_\Theta^2 \Sigma \beta_\Theta) , \quad \sqrt{m} \{ \ell(n_\ast) - \ell(n_0) \} \rightarrow N(0, \gamma_\Theta^2 \Sigma \gamma_\Theta) \tag{33}
\]
in distribution, and denoting \( z_\alpha = \Phi^{-1}(1 - \alpha/2) \), the confidence intervals
\[
I_{\alpha}(\Theta, \hat{\Sigma}) = \left[ n_\ast(\Theta) \pm z_\alpha \sqrt{\frac{\beta_\Theta^2 \Sigma \beta_\Theta}{m}} \right] , \quad J_{\alpha}(\Theta, \hat{\Sigma}) = \left[ \ell(n_\ast) \pm z_\alpha \sqrt{\frac{\gamma_\Theta^2 \Sigma \gamma_\Theta}{m}} \right]
\]
satisfy \( P \{ n_0 \in I_{\alpha}(\Theta, \hat{\Sigma}) \} \rightarrow 1 - \alpha \) and \( P \{ \ell(n_0) \in J_{\alpha}(\Theta, \hat{\Sigma}) \} \rightarrow 1 - \alpha \) as \( m \rightarrow \infty \).

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

**Proof.** From \( \ell(n) = k_1 n + k_2(n; \theta)(N - n) \) and \( n_\ast = \{ n : \ell'(n; \Theta) = 0 \} \), where such \( n_\ast \) is unique, we have (as per section 5.2)
\[
(\nabla n_\ast)_i \frac{\partial n_\ast}{\partial \Theta_i} = \frac{\partial^2 \ell}{\partial n^2}(\beta_\Theta) , \quad (\nabla \ell(n_\ast))_i \frac{\partial \ell}{\partial \Theta_i} = (\gamma_\Theta)
\]
for all components \( \Theta_i \) of \( \Theta \). Thus partial derivatives of \( n_\ast \) exist as long as
\[
\frac{\partial^2 \ell}{\partial n^2} > 0 \tag{34}
\]
By assumption, \( \ell(\cdot; \Theta_0) \) has a minimum at \( n_0 \). Since
\[
\frac{\partial^2 \ell}{\partial n^2} = \frac{\partial^2}{\partial n^2} k_2(n; \theta) - 2 \frac{\partial}{\partial n} k_2(n; \theta) \tag{35}
\]
where both terms are continuous in a neighbourhood of \( n_0, \Theta_0 \) by assumption, the value of \( \frac{\partial^2 \ell}{\partial n^2} \) must be positive in some (possibly smaller) neighbourhood \( R_\delta \) of \( (n_0, \Theta_0) \) of width \( 2\delta \), and hence all partial derivatives of \( n_\ast \) and \( \ell(n_\ast) \) are defined (and indeed continuous) in \( R_\delta \). Within \( R_\delta \) we have
\[
n_\ast(\Theta) = n_\ast(\Theta_0) + (\nabla n_\ast|_{\Theta=\Theta_0}) \cdot (\Theta - \Theta_0) + O(||\Theta - \Theta_0||_2) \\
\ell(n_\ast) = \ell(n_\ast(\Theta_0), \Theta_0) + (\nabla \ell(n_\ast)|_{\Theta=\Theta_0}) \cdot (\Theta - \Theta_0) + O(||\Theta - \Theta_0||_2) \tag{36}
\]
\[ \ell(n_0) + \gamma^l_{\Theta_0} \cdot (\Theta - \Theta_0) + O(||\Theta - \Theta_0||_2) \] (37)

from which, given the assumption of asymptotic normality of \( \Theta \), assertions 33 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

\[
P \left( n_0 \geq n_*(\Theta) + z_\alpha \frac{\sqrt{\beta^t \hat{\Sigma} \beta_0}}{m} \right) = P \left( \frac{\sqrt{m}}{z_\alpha} (n_0 - n_*(\Theta)) \geq \sqrt{\beta^t \hat{\Sigma} \beta_0} \right)
\]

\[ = P \left( \frac{\sqrt{m}}{z_\alpha} (n_0 - n_*(\Theta)) \geq \left( \beta^{t}_{\Theta_0} \Sigma \beta_0 + \beta^{t}_{\Theta} (\hat{\Sigma} - \Sigma) \beta_0 + (\beta_\Theta - \beta_\Theta_0)^t (\Sigma (\beta_\Theta + \beta_\Theta_0))^{\frac{1}{2}} \right) \right)
\]

\[ \rightarrow P \left( \frac{\sqrt{m}}{z_\alpha} (n_0 - n_*(\Theta)) \geq \sqrt{\beta^t \Sigma_0 \beta_0} \right)
\]

\[ = \frac{\alpha}{2} \] (38)

since, by the assumption of convergence of \( \hat{\Sigma} \)

\[ \left| \beta^t_{\Theta} \left( \Sigma - \hat{\Sigma} \right) \beta_\Theta \right| \leq ||\beta_\Theta||_2 ||\Sigma - \hat{\Sigma}||_2 \]

\[ \rightarrow_p 0 \] (39)

and, since \( P(\Theta \in R_\delta) \rightarrow 1 \) by the asymptotic normality of \( \Theta \), we have from 36

\[ \left| (\beta_\Theta - \beta_\Theta_0)^t \Sigma (\beta_\Theta + \beta_\Theta_0) \right| = O(||\beta_\Theta - \beta_\Theta_0||_2) \]

\[ \rightarrow_p 0 \] (40)

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

\[ P(n_0 \in I_\alpha(\Theta, \hat{\Sigma})) \rightarrow 1 - \alpha \] (41)

as required. An identical argument holds for \( J_\alpha(\Theta, \hat{\Sigma}) \).
Our second algorithm for estimation of optimal holdout sizes uses Bayesian emula-
tion (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 ‘learn-
ing 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 \( \ell \) as an ‘emulator’ mod-
elled 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 \( n \) 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 \( n \) and their corresponding observed cost estimates \( d \) with
sampling variances \( \sigma^2 \), noting that \( \sigma \) has a slightly different meaning to that in section 5.2.
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

\[
\begin{align*}
  u(n) &\sim GP \{0, k(n, n')\} \\
  k(n, n') &= \sigma_u^2 \exp \left\{-\left(\frac{n - n'}{\zeta}\right)^2\right\}
\end{align*}
\]

where \( k \) is chosen to enforce smoothness in \( \ell(n) \), though other covariance functions hav-
ing varying degrees of smoothness could be used. The hyperparameters \( \theta, \sigma_u \) and \( \zeta \) are
problem-specific and must be specified; however, we will show that for sufficiently large |\( n \)|
mis-specification of \( \theta, \sigma_u \) and \( \theta \) is overcome.

We denote \( d_i \sim N \{\ell(n_i), \sigma_i^2\} \) where \( d_i = (d)_i \) is the ith element of \( d \), etc. Since
\( n \) may be a multiset, we take \( n^1 \) as the set of unique values in \( n \), with \( d^1, \sigma^1 \) defined
correspondingly with \( d^1_i \) as an inverse-variance weighted mean of \( \{d_j : n_j = n^1_i\} \) with
sample variance \( (\sigma^1_i)^2 \):

\[
d^1_i = \frac{\sum_{j:n_j = n^1_i} d_j \sigma_j^{-2}}{\sum_{j:n_j = n^1_i} \sigma_j^{-2}} \quad (\sigma^1_i)^2 = \frac{1}{\sum_{j:n_j = n^1_i} \sigma_j^{-2}}
\]

noting that \( \sigma^1_i \) may change with \( i \). Alternatively, we may account for the variation in \( d \)
through ‘inactive’ variables and opt to use a ‘nugget’ term; this approach is described in
detail in Supplement S6.1.
Now with input \( n \), with an unevaluated loss value, our emulator specifies that the joint distribution of \( \ell(n) \) and our observed output values \( d^1 \) is:

\[
\begin{bmatrix} \ell(n) \\ d^1 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} m(n, \Theta) \\ m(n^1, \Theta) \end{bmatrix}, \begin{bmatrix} k(n, n) & k(n, n^1) \\ k(n^1, n) & k(n^1, n^1) + \text{diag}\{(\sigma^1)^2\} \end{bmatrix} \right)
\]

where \( m(n^1, \Theta) = m(n^1_i, \Theta), k(n, n^1)_i = k(n^1_i, n) = k(n^1_i, n^1) = k(n^1_i, n^1)_i = k(n^1_i, n^1) \), \( \text{diag}\{(\sigma^1)^2\} = \sigma^1_{1i} \). By obtaining the conditional posterior distribution \( \pi_n = \pi\{\ell(n) \mid n, n^1, d^1, \sigma^1\} \) and taking the expectation and variance we gain the Bayes linear update equations (Vernon et al., 2018):

\[
\begin{align*}
\mu(n) &= \mathbb{E}_{\pi_n}\{\ell(n)\} = m(n, \Theta) + k(n, n^1) \left[ k(n^1, n^1) + \text{diag}\{(\sigma^1)^2\} \right]^{-1} \{d^1 - m(n^1, \Theta)\} \\
\Psi(n) &= \text{var}_{\pi_n}\{\ell(n)\} = k(n, n) - k(n, n^1) [k(n^1, n^1) + \text{diag}\{(\sigma^1)^2\}]^{-1} k(n^1, n)
\end{align*}
\]

In algorithm 3, 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 \( \ell \) 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 \( d^- = \min_i\{d^1_i\} \), and

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

We see that by formulating the problem in terms of \( EI(\cdot) \), there is a natural stopping criterion on the size of \( n \): setting a threshold \( EI(\hat{n}) > \tau \) allows us to specify that for each iteration that we expect total cost to improve by at least \( \tau \) over our current known minimum \( d^- \). Examples of \( \mu(\cdot), \Psi(\cdot), EI(\cdot) \) are shown in Supplementary Figure 8. This leads to algorithm 4 for OHS estimation by Bayesian Emulation (a more precise version of algorithm 2 in the main manuscript).

Various results on the consistency of the expected improvement algorithm have been proved, albeit in differing settings; either with noiseless observations \( d \) (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 S6.2.

**Theorem S6.** If \( \ell(n), \sigma \), and \( m(n, \Theta) \) are almost surely bounded and \( d_i \sim N\{l(n_i), \sigma^2_i\} \) then for every \( n \in \{1, \ldots, N\} \), as the multiplicity of \( n \) in \( n \) tends to \( \infty \) we have \( \mu(n) \rightarrow \ell(n) \) and \( \Psi(n) \rightarrow 0 \) almost surely with respect to variation in \( d \).
Algorithm 4: Emulation OHS estimation; minimum cost improvement $\tau$

1. $n, d, \sigma^2 \leftarrow$ some initial values $n$ with $(d)_i = d(n_i) \approx \ell(n)$, $(\sigma^2)_i = \text{var}(d(n_i))$;
2. Coalesce $n, d, \sigma$ into $n^1, d^1, \sigma^1$ as above;
3. Estimate functions $\mu(n)$, $\Psi(n)$, $EI(n)$, with $\Theta = \Theta(n^1, d^1, \sigma^1)$;
4. while $\max_{n \in \{1, \ldots, N\}} \{EI(n)\} > \tau$ do
5. $\tilde{n} \leftarrow \arg \max_{n \in \{1, \ldots, N\}} EI(n)$;
6. Estimate $d(\tilde{n}) \approx k_2(\tilde{n})$;
7. $n \leftarrow (n \cup \tilde{n})$; $d \leftarrow (d \cup d\{\tilde{n}\})$; $\sigma^2 \leftarrow [\sigma^2 \cup \text{var}(d(\tilde{n}))]$;
8. Coalesce $n, d, \sigma$ into $n^1, d^1, \sigma^1$;
9. Re-estimate functions $\mu(n)$, $\Psi(n)$, $EI(n)$, with $\Theta = \Theta(n^1, d^1, \sigma^1)$;
10. end
11. return $n^\text{final} = \arg \min_{n \in \{1, \ldots, N\}} \{\mu(n)\}$

also noting the following simple result, proved in Supplement S6.3:

Theorem S7. Given the conditions of Theorem S6, for every $n \in \{1, \ldots, N\}$, as the multiplicity of $n$ in $n$ tends to $\infty$,
$$EI(n) \rightarrow 0$$
almost surely with respect to randomness in $d$

These results assert that $\mu(n)$ can eventually approximate any loss function sufficiently well given enough estimates of $\ell$ at all values of $n$. It is not obvious that this is guaranteed by algorithm 4, although we show that this generally does occur in the following, the proof of which is given in Supplement S6.4:

Theorem S8. If $\ell(n)$, $\sigma$, and $m(n, \Theta)$ are almost surely bounded and $d_i \sim N\{l(n_i), \sigma^2_i\}$ then under algorithm 4 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: \{$n : \text{pr}_{n_*} \{\ell(n) < \mu(n_*)\} \geq 1 - \alpha\}$, although this should not be interpreted as a credible set for $n_*$. This is implemented in our R package OptHoldoutSize, available on CRAN.

S6.1 Emulation of cost function with nugget term

Rather than explaining the variation of values in $d$ corresponding to a design point in $n^1$ 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):

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

(43)

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:

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

(44)

with

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

(45)

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 $\mathbf{d}^1$ is now:

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

(46)

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

$$\mu(n) = \mathbb{E}_{\pi_{n^1}}(\mathcal{E}(n)) = m(n) + k(n, \mathbf{n}^1)[k(\mathbf{n}^1, \mathbf{n}^1) + \text{diag}(\kappa(\mathbf{n}^1))]^{-1}(\mathbf{d}^1 - m(\mathbf{n}^1))$$

$$\Psi(n) = \text{var}_{\pi_{n^1}}(\mathcal{E}(n)) = k(n, n) + \kappa(n) - k(n, \mathbf{n})[k(\mathbf{n}^1, \mathbf{n}^1) + \text{diag}(\kappa(\mathbf{n}^1))]^{-1}k(\mathbf{n}^1, n)$$

(47)

(48)

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.
S6.2 Proof of Theorem S6

Theorem S6. If \( \ell(n), \sigma, \) and \( m(n, \Theta) \) are almost surely bounded and \( d_i \sim N \{ \ell(n_i), \sigma_i^2 \} \) then for every \( n \in \{1, \ldots, N\} \), as the multiplicity of \( n \) in \( \mathbf{n} \) tends to \( \infty \) we have \( \mu(n) \to \ell(n) \) and \( \Psi(n) \to 0 \) almost surely with respect to variation in \( \mathbf{d} \).

Proof. Assume W.L.O.G that \( (n^1)_1 = n \). Since \( \sigma \) is bounded, we have (from equation 42) \( \text{var}\left((d^1)_1\right) = |\sigma_1^2| \to 0 \), so \( (d^1)_1 \to \ell(n) \) almost surely. We now prove that \( k(n, n^1)[k(n^1, n^1) + \text{diag}((\sigma^1)^2)]^{-1} = (1, 0, \ldots, 0) \) when \( (\sigma^1)_1 = 0 \). Now:

\[
k(n, n^1)[k(n^1, n^1) + \text{diag}((\sigma^1)^2)]^{-1} = (1, 0, \ldots, 0)
\]

\[
\iff k(n, n^1) = (1, 0, \ldots, 0) * k(n^1, n^1 + \text{diag}((\sigma^1)^2))
\]

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

\[
\mu(n) = m(n, \Theta) + (1, 0, \ldots, 0)(d^1 - m(n^1, \Theta)) = m(n, \Theta) + d(n) - m(n, \Theta) = d(n) = \ell(n)
\]

almost surely, and

\[
\Psi(n) = k(n, n) - (1, 0, \ldots, 0)k(n^1, n) = k(n, n) - k(n, n) = 0
\]

in the limit.

\[\square\]

S6.3 Proof of Theorem S7

Theorem S7. Given the conditions of Theorem S6, for every \( n \in \{1, \ldots, N\} \), as the multiplicity of \( n \) in \( \mathbf{n} \) tends to \( \infty \),

\[
EI(n) \to 0
\]

almost surely with respect to randomness in \( \mathbf{d} \).

Proof. From Theorem S6 we have that \( \mu(n) \to \ell(n) < \infty \) and \( d^1_i \to \ell(n^1) \), so therefore in the limit we can state \( \Pr\{ \infty < d^- - \mu(n) \leq 0 \} = 1 \). Indeed, let \( j \) be the index such that \( n^1_j = n \). If in the limit \( d^- > \mu(n) = d(n) \) then this implies that \( d^1_j < \min_i\{d^1_i\} \) which is a contradiction. Also note from Theorem S6 that \( \Psi(n) \to 0 \) 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.

\[\square\]
S6.4 Proof of Theorem S8

Theorem S8. If $\ell(n)$, $\sigma$, and $m(n, \Theta)$ are almost surely bounded and $d_i \sim N\{l(n_i), \sigma_i^2\}$ then under algorithm 4 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 4 leads to the multiplicity of $\tilde{n}$ in $n$ tending to infinity, from which the result follows from Theorem S6.

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 $n$: namely that if the multiplicity of $n$ in $n$ diverges, the value of $EI(n)$ converges to 0; otherwise, it remains positive. We introduce the index $EI_{n}(n)$ to indicate the dependence of $EI(n)$ on $n$ and assume that the function $\ell(n)$ is fixed. For a multiset $n_i$, we denote mult$_{n_i}(n)$ as the multiplicity of $n$ in $n_i$.

Lemma S2. Suppose $m \times m$ matrix $A$ is symmetric. Denote by $I^1$ the $m \times m$ matrix with $I^1_{i,j} = 1$ if $i = j = 1$. 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

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

(52)

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 $MI^1M = M_1M^T_1$. Since $(A + pI)$ and $A$ differ only in the top row, we have $\text{adj}(A + pI)_1 = \text{adj}(A)_1$, where adj(·) indicates the adjugate matrix and ·$_1$ the top row. We now have

$$\frac{\partial}{\partial p} \left( x^T (A + pI^1)^{-1} x \right) = -x^T (A + pI^1)^{-1} \frac{\partial (A + pI^1)}{\partial p} (A + pI^1)^{-1} x$$

$$= -x^T (A + pI^1)^{-1} I^1 (A + pI^1)^{-1} x$$

$$= x^T \text{adj}(A + pI^1) I^1 \text{adj}(A + pI^1)x$$

$$= \frac{x^T \text{adj}(A + pI^1)I^1 \text{adj}(A + pI^1)x}{|A + pI^1|^2}$$

$$= -\frac{x^T \text{adj}(A + pI^1)I^1 \text{adj}(A + pI^1)x}{|A + pI^1|^2}$$

$$= -\frac{x^T \text{adj}(A)I^1 \text{adj}(A)^T x}{|A + pI^1|^2}$$

$$= -\frac{|A_x|^2}{|A + pI^1|^2}$$

as required.  \[\Box\]
Lemma S3. Let $S_1$ and $S_2$ be disjoint subsets of $[N] = \{1, \ldots, N\}$ with $S_1 \cup S_2 = [N]$. For a multiset $n$ denote

$$q_1(n) = \max_{n \in S_1} \text{mult}_n(n)$$
$$q_2(n) = \min_{n \in S_2} \text{mult}_n(n)$$

(53)

Suppose we have infinite sequences $n, d, \sigma$, where $d \sim N(l(n), \sigma^2)$ and $\sigma^2$ is upper-bounded, and let $n_i, d_i, \sigma_i$ denote the (multiset) first $i$ elements of each sequence. Suppose that $q_1(n_i) \leq m_1$ for all $i$ and $q_2(n_i) \to \infty$, and the set $\{k_2(n, \Theta_i) = k_2(n, \Theta(n_i, d_i, \sigma_i)) : n \in \{1, \ldots, N\}, i \in \mathbb{N}\}$ is almost surely asymptotically bounded. Then for sufficiently large $\sigma_u$:

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

(54)

almost surely.

Proof. We will in fact show that even $\liminf_{i \to \infty} EI_{n_i}(n) > 0$ for $n \in S_1$, but $\limsup$ will suffice for our purposes. We note that

$$EI_{n_i}(n) > 0 \iff \sqrt{\Psi_{n_i}(n)} \phi \left( \frac{d_{n_i} - \mu_{n_i}(n)}{\sqrt{\Psi_{n_i}(n)}} \right) > (\mu_{n_i}(n) - d_{n_i}^-) \Phi \left( \frac{d_{n_i}^- - \mu_{n_i}(n)}{\sqrt{\Psi_{n_i}(n)}} \right)$$

(55)

We will show that for all $n$, we have

$$P \left( -\infty < \liminf_{i \to \infty} (d_{n_i}^- - \mu_{n_i}(n)) \right) = 1$$

(56)

By the argument in theorem S6 and corollary S7 we have for $n \in S_2$ that $\lim_{i \to \infty} \Psi_{n_i}(n) = 0$, from which both sides of 55 converge to 0. For $n \in S_1$ we will show $\lim_{i \to \infty} \Psi_{n_i}(n) > 0$, in which case we may define

$$z_{n_i}(n) = \frac{\mu_{n_i}(n) - d_{n_i}^-}{\sqrt{\Psi_{n_i}(n)}}$$

(57)

from which inequality 55 reduces to

$$\phi(z_{n_i}(n)) > z_{n_i}(n) \Phi(-z_{n_i}(n))$$

(58)

which holds for all $0 \leq z_{n_i}(n) < -\infty$. Since $z_{n_i}(n)$ is asymptotically bounded between positive values, the result follows.

Beginning with $d_{n_i}$, we note that $d_{n_i}^-$ is the minimum of
1. Values of \( d_1^1 \) corresponding to values of \( n_1^1 \) in \( S_1 \); and

2. Values of \( d_1^1 \) corresponding to values of \( n_1^1 \) in \( S_2 \)

For sufficiently large \( s \), the sequence \( \{n_j = (n)_j : j > s\} \) never contains any \( n \in S_1 \) again; hence, the minimum of item 1 is determined after finitely many \( i \) again; hence, the minimum of item 1 is determined after finitely many \( i \) and its limit is finite. Since \( \sigma \) is upper-bounded, all values of \( d_1^1 \) in item 2 converge to finite values in \( \{\ell(n) : n \in S_2\} \) almost surely. Hence \( d_{n_i}^1 \) converges almost surely to a finite value.

Since \( \lim sup_{i \to \infty} \) and \( \lim inf_{i \to \infty} \) of \( m(n; \Theta (n_i, d_i, \sigma_i)) \) are almost surely finite, all terms in \( \mu(n) \) are asymptotically finite, from which equation (56) follows.

It remains to consider \( \Psi_{n_i}(n) \) for \( n \in S_1 \). Firstly take \( n \in n^1 \) and suppose W.L.O.G that \( n_{i_1}^1 = n \). Since \( n \in S_1 \) we have \( \lim_{i \to \infty} \text{mult}_{n_i}(n) > 0 \) so \( \lim_{i \to \infty}(\sigma_i^1)_{1_1} \) exists and is positive. Denoting \( \sigma' \) as \( \sigma_i^1 \) with 0 substituted for the first element, we have

\[
\frac{\partial}{\partial(\sigma_1^1)} \Psi_{n_i}(n) = \frac{\partial}{\partial(\sigma_1^1)}(k(n,n) - k(n,n_1^1)[k(n_1^1,n_1^1) + \text{diag}((\sigma_1^1)^2)]^{-1}k(n_1^1,n))
\]

\[
= \frac{|k(n_1^1,n_1^1) + \text{diag}((\sigma')^2)|^2}{|k(n_1^1,n_1^1) + \text{diag}((\sigma_1^1)^2)|^2} > 0
\]

by lemma S2; hence \( \Psi_{n_i}(n) \), considered as a function of \( (\sigma_1^1)^2 \), is increasing. Given that \( \lim_{i \to \infty}(\sigma_i^1)^2_{1_1} = 0 \) for \( n_{i_1}^1 \) \( \in S_2 \) and is positive for \( n_{i_1}^1 \) \( \in S_1 \), we conclude that \( \lim_{i \to \infty} \Psi_{n_i}(n) \) is positive when \( n \in S_1 \) and \( n \in n^1 \).

If \( n \neq n_1^1 \), \( n \) never occurs in any \( n_i \), then we firstly note that since \( k(n,n) < k(n,m) \) for any \( m \neq n \), we have:

\[
k(n,n) - k(n,n_1^1)[k(n_1^1,n_1^1)]^{-1}k(n_1^1,n) > 0 \quad (59)
\]

This omits the term \( \text{diag}((\sigma_1^1)^2) \) from the expression for \( \Psi_{n_i}(n) \). However, if we denote \( k_j' \), the matrix \( k(n_1^1,n_1^1) + \text{diag}((\sigma_1^1)^2) \) with the \( j \)th row replaced by \( k(n,n_1^1) \), we have from lemma S2:

\[
\frac{\partial}{\partial(\sigma_1^1)^2_{1_1}} \Psi_{n_i}(n) = \frac{|k_j'|^2}{|k(n_1^1,n_1^1) + \text{diag}((\sigma_1^1)^2)|^2} > 0
\]

for any element \( (\sigma_1^1)^2_{1_1} \) of \( (\sigma_1^1)^2 \); hence \( \Psi_{n_i}(n) \) is increasing in any such element and its positivity follows. This completes the proof of the lemma.

\[\Box\]

Now suppose that some \( n \in \{1, \ldots, N\} \) occurs only finitely often in \( n^1 \). 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 of \( S_1 \) occurs in \( n^1 \) with finite multiplicity there must be some \( j \) such that no \( n \in S_1 \) occurs amongst the values \( \{(n^1)_{j+1},(n^1)_{j+2}, \ldots\} \). But from lemma S3, there will almost surely eventually be some \( J > j \) for which some value in \( \{E(n^1)_{n_j} : n \in S_1\} \) exceeds all
values in \( \{E_{I_{nj}}(n) : n \in S_1\} \), and hence \((n^1)_{j+1} \in S_1\) (as long as \(\tau\) is sufficiently small), contradicting the choice of \(j\). So the event that an \(n \in \{1, \ldots, N\}\) occurs in \(n^1\) with finite multiplicity has probability 0. This completes the proof.
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 $\rho$ 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.

Figure 5 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 2 and 4 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 $\rho$) 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.
Figure 5: Examples of cost functions as per Theorem 1 arising naturally from a basic risk score, with varying underlying model (und.), risk score type ($\rho$) and one point-wise standard deviation (shaded regions). The contributions of terms $k_1n$ to $\ell(n)$ depend only on the underlying model and are the same in each column. OHS: optimal holdout set size.
S8 Estimation of parameters for optimal holdout size in ASPRE

S8.1 Estimation of $N$

As per our assumptions, we presume we are refitting ASPRE to 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

$$\frac{8 \times 10^4}{5 \times 10^6} = \frac{1}{125}$$

so we have

$$N \approx 5 \times 8 \times 10^4$$

$$= 400000$$

with standard error

$$SE(N) \approx 5\sqrt{5 \times 10^6 \times \frac{1}{125} \left(1 - \frac{1}{125}\right)}$$

$$\approx 1500$$

S8.2 Estimation of $k_1$ and $k_2$

We assume $\pi = 10\% = \frac{2707}{25797}$, the proportion of individuals assigned to the treatment group in Rolnik et al. (2017a) due to having an estimated risk of PRE > 1%.

To estimate $k_1$, 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 individuals were assessed, amongst which 239 developed PRE, for an overall incidence of $\frac{239}{8875} \approx 0.027$. 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. 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 that amongst the 10% of women designated highest-risk by the NICE guidelines, we have a PRE rate of

\[
\pi_1 \approx \frac{TPR \times (\text{Num. PRE})}{\text{Num. positive}} = \frac{0.2 \times 239}{0.1 \times 8875} \approx 0.054
\]

(64)

and amongst the 90% designated lower risk, a PRE rate of

\[
\pi_0 \approx \frac{(1 - TPR) \times (\text{Num. PRE})}{\text{Num. negative}} = \frac{0.8 \times 239}{0.9 \times 8875} \approx 0.024
\]

(65)

Given that true positive rates of the NICE guidelines are computed as a fraction with denominator 239, we presume standard errors of \(\pi_1\) and \(\pi_0\) of

\[
\text{SE}(\pi_1) \approx \sqrt{\pi_1(1 - \pi_1)} \times 0.1 \approx 0.0076
\]

\[
\text{SE}(\pi_0) \approx \sqrt{\pi_0(1 - \pi_0)} \times 0.9 \approx 0.0017
\]

(66)

Now, treating errors in \(\pi_0\), \(\pi_1\) and \(\alpha\) as pairwise independent

\[
k_1 = \pi_0(1 - \pi) + \pi_1 \pi \alpha \approx 0.0235
\]

\[
\text{SE}(k_1) = \text{SE}(\pi_0(1 - \pi) + \pi_1 \pi \alpha) \approx 0.0016
\]

(67)

We estimate the population prevalence \(\pi_{PRE}\) of untreated PE as the frequency observed in the original ASPRE data:

\[
\pi_{PRE} = \frac{1426}{57974} \approx 2.4\%
\]

(68)

Note that, although this is approximately equal to \(\pi_0\), they are different quantities: \(\pi_0\) 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 \( \pi \) of individuals are designated high-risk. Thus for any training set size \( n \)

\[
\pi_0(n) = \frac{\pi_{PRE} - \pi_1(n)}{1 - \pi}
\]

so the average cost to an individual in the intervention set may be expressed in terms of \( \pi_1(n) \):

\[
k_2(n) = \pi_0(n)(1 - \pi) + \pi_1(n)\pi\alpha
= \pi_{PRE} - \pi_1(n)(1 - \alpha)
\]

### S8.3 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 \( \pi_{PRE} \), 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 \( n \) and \( k_2/d \), we initially chose a set \( n \) of 20 random values from [500, 30000]. For each size \( n \) in \( n \), 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 \( k_2 \) as above. We fitted values \( \theta = \theta(n, k_2) = (a, b, c) \) parametrising \( k_2 \) as the maximum-likelihood estimator of \( \theta \) under the model

\[
(k_2)_i \sim N(k_2((n)_i, \theta), \sigma^2) \sim N(a(n)_i^{-b} + c, \sigma^2)
\]

for a fixed values \( \sigma \), noting that the estimate of \( \theta \) is independent of \( \sigma \). For the parametric algorithm, we then set all values of \( \sigma \) to the same value, chosen empirically as the sample variance of \( k_2 - k_2(n, \theta(n, k_2)) \)

\[
(k_2)_i \sim N(k_2((n)_i, \theta), \sigma^2) \sim N(a(n)_i^{-b} + c, \sigma^2)
\]

For the emulation algorithm, we set values \( d \) as

\[
d_i = k_1(n)_i + (k_2)_i(N - (n)_i)
\]

transforming values \( \sigma \) correspondingly for use in the emulation algorithm. We then sequentially chose 100 additional values \( n \) using both algorithm 3 and 4, setting \( \sigma \) as the
same value found in S8.3. After choosing the 120 values of $n$ using algorithm 3, we re-estimated $k_2/d$ 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.
S9 Supplementary discussion

In this section we discuss the use and consequences of several of our methods and results in greater depth.

Theorem 1 establishes that straightforward conditions on the system to be modelled lead to an optimal holdout set size which is straightforward to find, and indeed that the cost function is convex (in a discrete sense). An obvious limitation is the requirement for convexity of $k_2$; risk score learning curves, on which $k_2$ depends, are not necessarily convex for complex models (Viering and Loog, 2021). In practice, the cost function $\ell(n)$ is still generally well-behaved even if convexity of $k_2$ is violated, and may be approximated using our emulation algorithm (as demonstrated in Supplementary Figure 9b and Figure 3 in the main manuscript).

The use of a Gaussian process emulator to approximate the true loss function enables an automatic selection of the optimal holdout set size under fewer assumptions, although the efficiency of this method heavily depends on the quality of our emulator. 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.

Our simulations and theoretical findings illuminate several non-obvious properties of the optimal holdout set size. Firstly, from section 6, we note that the optimal holdout set size is often fairly small, and training a risk score until improvement in $k_2$ is negligible is counterproductive. More accurate risk scores do not necessarily lead to lower optimal holdout set size, but generally lead to lower loss. Given corollary 1, it is generally better to err on a higher side of the optimal holdout set size, since cost increases at most linearly.
Figure 6: 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 \((a, b, c) = (\frac{3}{2}, \frac{3}{2}, \frac{1}{4}), k_1 = \frac{4}{5}, N = 10^4\).
Figure 7: Dependence of minimum total cost on parameters of estimated learning curve \((a, b, c, \text{ with } k_2(n; a, b, c) = a n^{-b} + c), \text{ cost in intervention set } k_1, \text{ and total number of samples } N. \) Figures show change in minimal cost \(\ell(n^*)\) while varying one parameter and holding others constant at \((a, b, c) = (\frac{3}{2}, \frac{3}{2}, 1), k_1 = \frac{4}{5}, N = 10^4.\)
Figure 8: Left panel shows emulator constructed using three $k_2()$ values (see pipelines). Function $m(n, \Theta)$ is constructed using $\theta$ 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 panel 8a. Note local minima at existing sample points.
Figure 9: Parametric and emulation algorithms with parametric assumptions satisfied or unsatisfied. OHS: optimal holdout set size
References

ANDRIANAKIS, Y and CHALLENOR, PG. Parameter estimation for Gaussian process emulators. Technical report, Technical report, Managing Uncertainty in Complex Models (2011).

BOWER, RICHARD G, GOLDSTEIN, MICHAEL, and VERNON, IAN. Galaxy formation: a Bayesian uncertainty analysis. *Bayesian analysis*, 5(4):619–669 (2010).

BROCHU, ERIC, CORA, VLAD M, and DE FREITAS, NANDO. 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).

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

KILLEYA, MATTHEW RH. *Thinking inside the box: using derivatives to improve Bayesian black box emulation of computer simulators with application to compart mental models*. Ph.D. thesis, Durham University (2004).

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

O’GORMAN, NEIL, WRIGHT, DAVID, POON, LC, ROLNIK, DANIEL L, SYNGELAKI, ARGYRO, DE ALVARADO, MERCEDES, CARBONE, ILMA F, DUTEMEYER, VIVIEN, FIOLNA, MADGALENA, FRICK, ALEX, ET AL. 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(6):756–760 (2017).

ROLNIK, DANIEL L, WRIGHT, DAVID, POON, LCY, SYNGELAKI, ARGYRO, O’GORMAN, NEIL, DE PACO MATALLANA, CATALINA, AKOLEKAR, RANJIT, CICERO, SIMONA, JANGA, DEEPA, SINGH, MANDEEP, ET AL. ASPRE trial: performance of screening for preterm pre-eclampsia. *Ultrasound in obstetrics & gynecology*, 50(4):492–495 (2017a).

ROLNIK, DANIEL L, WRIGHT, DAVID, POON, LIONA C, O’GORMAN, NEIL, SYNGELAKI, ARGYRO, DE PACO MATALLANA, CATALINA, AKOLEKAR, RANJIT, CICERO, SIMONA, JANGA, DEEPA, SINGH, MANDEEP, ET AL. Aspirin versus placebo in pregnancies at high risk for preterm preeclampsia. *New England Journal of Medicine*, 377(7):613–622 (2017b).

RYZHOV, ILYA O. On the convergence rates of expected improvement methods. *Operations Research*, 64(6):1515–1528 (2016).
Tsymbal, Alexey. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 106(2):58 (2004).

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

Vernon, Ian, Liu, Junli, Goldstein, Michael, Rowe, James, Topping, Jen, and Lindsey, Keith. 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).

Viering, Tom and Loog, Marco. The shape of learning curves: a review. *arXiv preprint arXiv:2103.10948* (2021).

Zliobaite, Indre. Learning under concept drift: an overview. *arXiv preprint arXiv:1010.4784* (2010).