Model updating after interventions paradoxically introduces bias
Supplementary Materials

James Liley[1,2,⋆] **Samuel R. Emerson**[3] **Bilal A. Mateen**[1,4,5]
Catalina A. Vallejos[1,2,⋆] **Louis J. M. Aslett**[1,3,⋆] **Sebastian J. Vollmer**[1,6,⋆]
[1] Alan Turing Institute, London, UK; [2] MRC Human Genetics Unit, Univ. of Edinburgh, UK;
[3] Department of Mathematical Sciences, Durham Univ., UK; [4] Kings College Hospital, London, UK;
[5] Wellcome Trust, London, UK; [6] Warwick Mathematics Institute, Univ. of Warwick, UK
⋆ Co-corresponding authors

]

7 Example of functions and variables in a realistic setting

We consider the model proposed by Rahimian et al. (2018) for prediction of emergency admission to a hospital in a given time period on the basis of electronic health records (EHRs). Such a model is not in common use in the location considered (England), so the data in the original paper is not affected by the problems we describe in the main manuscript.

For clarity[9], we presume a prediction window of ten months (February-November), and that predictions are made and distributed to primary health practitioners in January, with a new model being trained on the basis of each year's data in December, to be implemented the following January. In this setting, distribution of the score may open a second causal pathway between covariates and outcome as shown in figure 1, and is thus susceptible to the problems of naive updating.

In this setting, variables and functions may be interpreted as follows:

1. Y the event 'an emergency admission in the following year'
2. $X_e(0)$ the values of all variables which affect $E(Y)$ at the time when the predictive score is computed (the start of each year)
3. An 'epoch': the time in which a given model is in use; eg, each year.
4. 'Time': $t = 0$ when the predictive score is computed (the start of January); $t = 1$ represents the time after which any interventions are made (the start of Feburary).
5. X_e^s covariates affecting $\mathbb{E}(Y)$ which are included in the predictive score but which cannot be directly modified in the time frame: age, time since most recent emergency admission
6. X_e^a covariates affecting $\mathbb{E}(Y)$ included in the predictive score which can be modified in the time frame: current medications.
7. X_e^ℓ covariates affecting $\mathbb{E}(Y)$ which are not included in the predictive score, and possibly can be modified in the time frame: blood pressures, cardiac function
8. f_e the underlying causal process for Y given patient status; that is, the probability of admission in the subsequent year, given covariates.
9. g_e^a Hypothetical prescribed interventions made on X^a in response to a predictive score; for instance, reduce drug dosages. We roughly assume that this intervention is symmetric; for a patient at low emergency risk, a higher drug dose is acceptable.
10. g_e^ℓ Hypothetical prescribed interventions made on X^ℓ in response to a predictive score; for instance, treat low or high blood pressure.

It is clear that if such a risk score were used universally, and data was collected from the period in which a model was in place was then, then the data would be affected by the effect of the predictive score itself.

The model does not fully describe this setting. The trichotomisation into X^ℓ, X^a, and X^s is not perfect; intervention on X^L could also affect some variables in X^a and vice versa. Interventions are likely to be random-valued to some extent.

8 Alternative system described by naive updating

We note that the definition of h (equation (9)), and hence the following comments on recursion dynamics, can be used to describe a related setting in which we track the same samples over epochs, and the effect of interventions g^a, g^ℓ remain in place. Formally, we retain definitions of $X^s, X^a, X^\ell, e, t, f_e, g_e^a, g_e^\ell, \rho_e$ and all assumptions except 4,7. In place, we assume that f_e, g_e^a, g_e^ℓ are fixed across epochs, but instead of resampling $X_e(0)$ from μ_e, we have

$$X_{e+1}(0) = X_e(1) \tag{11}$$

[9] Analogous times and variables can be described for other prediction periods and updating patterns

thus, while values $X_0(0)$ are sampled from the distribution μ_0, values $X_e(0)$ are then determined for $e > 0$. We illustrate this in figure 4

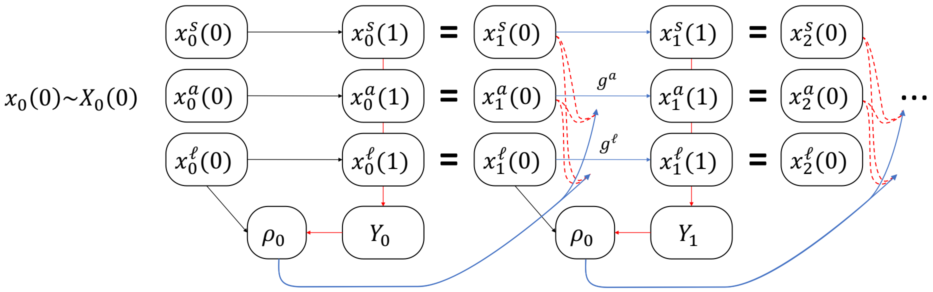

Figure 4: Diagram showing alternative setup for naive updating. Values x^s, x^a, x^ℓ are sampled at $(e,t) = (0,0)$, and used to determine ρ_0. Values are conserved until $t = 1$, and remain the same at the start of epoch 1 ($(e,t) = (1,0)$). Values are intervened on by g^a, g^ℓ according to $\rho_0\left(x_1^s(0), x_1^a(0)\right)$), and resultant values at $(e,t) = (1,1)$ are conserved until the start of the next epoch at $(e,t) = (2,0)$. Lowercase leters indicates that, while quantities random-valued, they inherit all randomness from their values at $(e,t) = (0,0)$. Colour and line conventions are as for figure 2

Now formulas (8), (9) will hold, and the recursion will proceed as detailed in theorem 1.

9 Proofs and counterexamples

9.1 Optimising both ρ and g^a, g^ℓ is equivalent to a general resource allocation problem

Consider the constrained optimisation problem in section 2.3. We show that if we allow ρ and g^a, g^ℓ to vary independently, then the constrained optimisation is equivalent to the solution of a problem in which the use of a predictive score is redundant.

Theorem 2. *Suppose that the triple $(\rho_{opt}, g^a_{opt}, g^\ell_{opt})$ minimises quantity (4) subject to constraint (5) in section 2.3, where all are arbitrary functions of two variables in the appropriate range. Let h^a_{opt} and h^ℓ_{opt} be solutions to a second constrained optimisation problem: find $h^a(x^s, x^a)$ and $h^\ell(x^s, x^a, x^\ell)$ which minimise*

$$\begin{aligned}\mathbb{E}_{X_e(0)}\{f(X^s,&\\ &h^a(X^s_e(0), X^a_e(0)),\\ &h^\ell(X^s_e(0), X^a_e(0), X^\ell_e(0)))\}\end{aligned} \tag{12}$$

subject to

$$\begin{aligned}\mathbb{E}_{X_e(0)}\{c^a(X^a_e(0),&\\ &X^a_e(0) - h^a(X^s_e(0), X^a_e(0)))+\\ c^\ell(X^\ell_e(0),&\\ &X^\ell_e(0) - h^\ell(X^s_e(0), X^a_e(0), X^\ell_e(0)))\} \leq C\end{aligned} \tag{13}$$

with c^a, c^ℓ, f as for section 2.3.

Then the minima of quantity (4) in the main text and of quantity (12) achieved by $(\rho_{opt}, g^a_{opt}, g^l_{opt})$ and $(h^a_{opt}, h^\ell_{opt})$ are the same.

Proof. Given a triple $(\rho_{opt}, g^a_{opt}, g^l_{opt})$, we explicitly construct an $(h^a_{opt}, h^\ell_{opt})$ which attains the same minimum, and vice versa.

Given $(\rho_{opt}, g^a_{opt}, g^l_{opt})$, the corresponding forms of h^a_{opt}, h^ℓ_{opt} are simply

$$\begin{aligned} h^a_{opt}(x^s, x^a) &= g^a_{opt}\left(\rho(x^s, x^a), x^a\right) \\ h^\ell_{opt}(x^s, x^a, x^\ell) &= g^\ell_{opt}\left(\rho(x^s, x^a), x^\ell\right) \end{aligned} \tag{14}$$

Given h^a_{opt}, h^ℓ_{opt}, the correspondence is slightly more complex. Set ρ_{opt} as a bijective function from $\mathbb{R}^{n_s+n_a}$ to $\mathbb{R}$; for instance, set it to 'splice' the decimal digits of arguments together. Now set g^a_{opt}, g^ℓ_{opt} to firstly 'decrypt' the value of ρ_{opt} back into constituent parts (x^s and x^a), and then compute $h^a_{opt}(x^s, x^a)$ and $h^\ell_{opt}(x^s, x^a, x^\ell)$ as outputs.

This shows that the two constrained optimisation problems are equivalent. □

We note that this implies that optimising (ρ, g^a, g^ℓ) jointly is equivalent to a more general treatment-allocation problem which does not involve a predictive score.

9.2 Counterexample showing naive updating can cause better models to appear worse

For this counterexample we shall use the following set up:

$$f(x^s, x^a, x^\ell) = f(x^s, x^a) = (1 + e^{-x^s - x^a})^{-1} \tag{15}$$

$$\rho_0(x^s, x^a \mid X^\star_0, Y^\star_0) = \begin{cases} \frac{\sum_{i=1}^n (Y^\star_0)_i \mathbb{1}\{\sum_{j=1}^2 (X^\star_0)_{ij} > 0\}}{\sum_{i=1}^n \mathbb{1}\{\sum_{j=1}^2 (X^\star_0)_{ij} > 0\}} & x^s + x^a > 0 \\ \frac{\sum_{i=1}^n (Y^\star_0)_i \mathbb{1}\{\sum_{j=1}^2 (X^\star_0)_{ij} \leq 0\}}{\sum_{i=1}^n \mathbb{1}\{\sum_{j=1}^2 (X^\star_0)_{ij} \leq 0\}} & x^s + x^a \leq 0 \end{cases} \tag{16}$$

$$\rho_1(x^s, x^a \mid X^\star_1, Y^\star_1) = (1 + e^{-\hat{\beta}_0 - x^s \hat{\beta}_1 - x^a \hat{\beta}_2})^{-1} \text{ where } \hat{\beta} = \operatorname{argmax}\{\mathcal{L}(\beta | X^\star_1, Y^\star_1)\} \tag{17}$$

$$m_{\tilde{f}_e}(\rho_e | X^\star_e, Y^\star_e) = \mathbb{E}_\mu\left[|f(X^s, g^a(\rho_{e-1}, X^a)) - \rho_e(X^s, X^a \mid X^\star_e, Y^\star_e)|\right] \tag{18}$$

$$g^a(\rho, x^a) = (1 - \rho)(x^a + 3) + \rho(x^a - 3) \tag{19}$$

For simplicity, we shall view the latent variables as having no effect on the true risk score f, which corresponds to the scenario where (if no interventions are made), it is possible with the data we observe to fully specify f. For the purpose of the counterexample it is reasonable to do this as model performance only requires $m_{\tilde{f}_e}$, which has no dependence on latent covariates.

We also state, that due to the omission of latent covariates, $X_e(0) = (X^s_e(0), X^a_e(0)) \sim N_2(0, I_2)$, which is then used to generate (through the statistical program R) an initial training data set at epoch 0, of size $n = 100$, which is summarised below:

index	$(\mathbf{X^\star_0})_{\cdot\mathbf{1}}$	$(\mathbf{X^\star_0})_{\cdot\mathbf{2}}$	$\mathbf{Y^\star_0}$
1	1.185	1.272	1
2	0.881	-0.995	0
3	0.122	-0.956	0
	⋮		
98	-0.826	1.779	1
99	0.853	0.151	1
100	0.177	0.805	1

This training data can then inputted into ρ_0 to give the following function:

$$\rho_0(x^s, x^a \mid X^\star_0, Y^\star_0) = \begin{cases} 0.733 & x^s + x^a > 0 \\ 0.200 & x^s + x^a \leq 0 \end{cases} \tag{20}$$

When intervening on any covariates at epoch 1 the function given in equation (20) will be used to produce $X_1(1)$ and subsequently Y_1.

We now consider $\mathbb{E}_{(X_0^\star,Y_0^\star)}\left[m_{\tilde{f}_0}(\rho_0|X_0^\star,Y_0^\star)\right]$, which we approximate using a Monte Carlo estimate with 1000 samples. However, $m_{\tilde{f}_0}(\rho_0|X_0^\star,Y_0^\star)$ also requires approximation, and so a Monte Carlo estimate with the same number of samples is also used for this function. The procedure is as follows:

1. For i from 1 to 1000:
   (a) Obtain a training data set , $(X_0^\star,Y_0^\star)_i$, by taking n samples of $(X_0(0),Y_0)$.
   (b) Use this training data set to obtain a $(\rho_0)_i$ of the form given in equation (20).
   (c) For j from 1 to 1000:
      i. Sample $(x^s,x^a)_j \sim X_0(0)$.
   (d) $m_{\tilde{f}_0}(\rho_0|(X_0^\star,Y_0^\star)_i) \approx \frac{1}{1000}\sum_{j=1}^{1000}|f((x^s,x^a)_j) - \rho_0((x^s,x^a)_j \mid (X_0^\star,Y_0^\star)_i)|$
2. $\mathbb{E}_{(X_0^\star,Y_0^\star)}\left[m_{\tilde{f}_0}(\rho_0|X_0^\star,Y_0^\star)\right] \approx \frac{1}{1000}\sum_{j=1}^{1000} m_{\tilde{f}_0}(\rho_0|(X_0^\star,Y_0^\star)_i)$

With this in mind, we give the following approximation: $\mathbb{E}_{(X_0^\star,Y_0^\star)}\left[m_{\tilde{f}_0}(\rho_0|X_0^\star,Y_0^\star)\right] \approx 0.124$.

If we assert that interventions never take place, then we can use the same procedure described above to obtain $\mathbb{E}_{(X_0^\star,Y_0^\star)}\left[m_{\tilde{f}_0}(\rho_1|X_0^\star,Y_0^\star)\right] \approx 0.056$. So here we can clearly see that in the setting where interventions are never made, $\mathbb{E}_{(X_0^\star,Y_0^\star)}\left[m_{\tilde{f}_0}(\rho_0|X_0^\star,Y_0^\star)\right] > \mathbb{E}_{(X_0^\star,Y_0^\star)}\left[m_{\tilde{f}_0}(\rho_1|X_0^\star,Y_0^\star)\right]$, and so the model closer to the truth is the logistic regression model at epoch 1. If agents were allowed to make interventions (based on (20)) however, we would consider $\mathbb{E}_{(X_1^\star,Y_1^\star)}\left[m_{\tilde{f}_1}(\rho_1|X_1^\star,Y_1^\star)\right] \approx 0.197$ instead. Now, since $\mathbb{E}_{(X_0^\star,Y_0^\star)}\left[m_{\tilde{f}_0}(\rho_0|X_0^\star,Y_0^\star)\right] < \mathbb{E}_{(X_1^\star,Y_1^\star)}\left[m_{\tilde{f}_1}(\rho_1|X_1^\star,Y_1^\star)\right]$, we would come to the incorrect conclusion that the model closer to the truth is the model used at epoch 1. Consequently we can state that, given the setup provided in section 3.1,

$$\begin{aligned}&\mathbb{E}_{(X_0^\star,Y_0^\star)}\left[m_{\tilde{f}_0}(\rho_0|X_0^\star,Y_0^\star)\right] > \mathbb{E}_{(X_0^\star,Y_0^\star)}\left[m_{\tilde{f}_0}(\rho_1|X_0^\star,Y_0^\star)\right] \not\Rightarrow \\ &\mathbb{E}_{(X_0^\star,Y_0^\star)}\left[m_{\tilde{f}_0}(\rho_0|X_0^\star,Y_0^\star)\right] > \mathbb{E}_{(X_1^\star,Y_1^\star)}\left[m_{\tilde{f}_1}(\rho_1|X_1^\star,Y_1^\star)\right]\end{aligned} \tag{21}$$

Additionally, we show that for this example:

$$\begin{aligned}&\mathbb{E}_{(X_0^\star,Y_0^\star)}\left[m_{\tilde{f}_0}(\rho_0|X_0^\star,Y_0^\star)\right] > \mathbb{E}_{(X_0^\star,Y_0^\star)}\left[m_{\tilde{f}_0}(\rho_1|X_0^\star,Y_0^\star)\right] \not\Rightarrow \\ &\mathbb{E}_{(X_0^\star,Y_0^\star)}\left[m_{\tilde{f}_0}(\rho_0|X_0^\star,Y_0^\star)\right] > \mathbb{E}_{(X_1^\star,Y_1^\star)}\left[m_{\tilde{f}_0}(\rho_1|X_1^\star,Y_1^\star)\right]\end{aligned} \tag{22}$$

as $\mathbb{E}_{(X_1^\star,Y_1^\star)}\left[m_{\tilde{f}_0}(\rho_1|X_1^\star,Y_1^\star)\right] \approx 0.215 > 0.124 \approx \mathbb{E}_{(X_0^\star,Y_0^\star)}\left[m_{\tilde{f}_0}(\rho_0|X_0^\star,Y_0^\star)\right]$. This statement is given here because for $\tilde{f}_0$, and therefore $m_{\tilde{f}_0}$, it is possible to gain estimates through a holdout test data set. Whilst the comparison is not between a risk score (ρ_e) and the function it is trying to estimate ($\tilde{f}_e$), the effect of deteriorating performance as epochs increase is still captured. Going further, it is assumed that if stakeholders were implementing naive model updating, they would assume that ρ_e is estimating $\tilde{f}_0$ for all epochs as the belief is that interventions do not effect the model. Therefore, comparison with $\tilde{f}_0$ will heighten the impression to stakeholders that using an updated model structure is causing performance to deteriorate, especially for epoch 0 to epoch 1, where for this comparison ρ_0 is actually estimating $\tilde{f}_0$.

We expect from a stakeholders view that comparison (using estimates) between the two models at successive epochs usually leads to the inequality $m_{\tilde{f}_0}(\rho_{e-1} \mid X_{e-1}^\star, Y_{e-1}^\star) < m_{\tilde{f}_0}(\rho_e \mid X_e^\star, Y_e^\star)$, and therefore the conclusion is that the new model leads to worse performance. We advise that a conclusion is only reached after further comparison is done between $m_{\tilde{f}_0}(\rho_{e-1} \mid X_e^\star, Y_e^\star)$ and $m_{\tilde{f}_0}(\rho_e \mid X_e^\star, Y_e^\star)$, as this gives an indication whether the drop in performance is due to the model structure or the intervention effect.

Finally, we advise caution when considering the effect of latent variables when estimating $m_{\tilde{f}_0}(\rho_e|X_e^\star,Y_e^\star)$. This is due to that fact that when holdout test data is used to obtain an estimate, it is an estimate of f rather than an estimate of $\tilde{f}_0$. If the latent variables have a small influence on f than $f \approx \tilde{f}_0$ and we can make inferences as shown above, but if latent variables have a large influence on f then our comparison is not based on $m_{\tilde{f}_0}$ but instead on m_f. This creates a problem as now how well we perceive our model's performance can be determined largely by how well a model arbitrarily captures the latent covariate information using just the set and actionable covariates. It therefore becomes substantially more difficult to determine whether the cause of a models poor performance is due to the model, the intervention effect or insufficient data. As a general rule however, large values of $m_{\tilde{f}_0}(\rho_0|X_0^\star, Y_0^\star)$ should indicate that either the initial model is very poor or that there is insufficient data, but in either case careful consideration of what could possibly influence the underlying mechanism should be made before a risk score is built and given to agents, to ensure that latent variables affect the model as little as possible.

9.3 Proof of theorem 1

If $h'(z_0) \leq -1$ then the single fixed point of h is unstable and ρ_e cannot converge to it unless it was always equal to z_0. There can be no other z with $h(z) = z_0$ since $h'(z) < 0$ by assumption.

Since $\rho_e \in [0, 1]$ and $h'(z) < 0$, ρ_e must tend toward a stable oscillation between two values, or converge to a single value.

If the bounds on partial derivatives hold, then from the triangle and Cauchy-Schwarz inequalities, for $z \in R$

$$\begin{aligned}
|h'(z)| &\leq \mathbb{E}_{X^L}\left[\sum_i^{p^a} |\delta_i^{g^a}\delta_i^{f^a}| + \sum_i^{p^L} |\delta_i^{g^\ell}\delta_i^{f^\ell}|\right] \\
&= \sum_i^{p^a} |\delta_i^{g^a}| \mathbb{E}_{X^\ell}\left[|\delta_i^{f^a}|\right] + \sum_i^{p^\ell} \mathbb{E}_{X^\ell}\left[|\delta_i^{g^\ell}\delta_i^{f^\ell}|\right] \\
&\leq \sqrt{\sum_i^{p^a} (\delta_i^{g^a})^2 \sum_i^{p^a} \mathbb{E}_{X^\ell}\left[\delta_i^{f^a}\right]^2} \\
&\quad + \sqrt{\sum_i^{p^\ell} \mathbb{E}_{X^\ell}\left[\left(\delta_i^{g^\ell}\right)^2\right] \sum_i^{p^\ell} \mathbb{E}_{X^\ell}\left[\left(\delta_i^{f^\ell}\right)^2\right]} \\
&\leq \sqrt{k_1 k_3} + \sqrt{k_2 k_4} < 1
\end{aligned} \tag{23}$$

so the map $h : \rho_e \to \rho_{e+1}$ is a contraction, and the convergence of the recurrence $\rho_e \to \rho_{e+1}$ follows from the Banach fixed-point theorem, as long as $\rho_e \in R$ for some value of e.

9.4 Counterexample showing failure of naive updating to generally solve constrained optimisation problem

For this counterexample, we do not need to consider latent covariates, and will assume they do not exist.

Under the setting in section 2.2, if ρ_n converges to $\rho_\infty(x^s, x^a)$ for some x^s, x^a under naive updating, then we have

$$\rho_\infty(x^s, x^a) = h(\rho_\infty(x^s, x^a) = f(g(\rho_\infty(x^s, x^a), x^a), x^s) \tag{24}$$

Suppose x^s and x^a each have dimension 1, and consider the example:

$$\begin{aligned}
f(x^a, x^s) &= \text{logit}(x^a + x^s) = \frac{1}{1 + \exp\left(-(x^a + x^s)\right)} \\
g(\rho, x^a) &= x^a - \log(1 + \rho) \\
c^a(x) &= x
\end{aligned}$$

For a given function ρ, the objective and cost are, respectively

$$\begin{aligned}
\text{obj}\{\rho\} &= E\left\{(1 + (1 + \rho)\exp(-(X^s + X^a)))^{-1}\right\} \\
\text{cost}\{\rho\} &= E\left\{\log(1 + \rho)\right\}
\end{aligned} \tag{25}$$

Using an oracle predictor of $Y|X$, as in the previous section, ρ_n converges to the fixed point of the recursion $z \to f(g(z,x^a),x^s)$, which is

$$\rho_\infty(x^s,x^a) = \frac{1}{2}\left(\sqrt{\left(e^{x+y}+1\right)^2 + 4e^{x+y}} - \left(e^{x+y}+1\right)\right) \tag{26}$$

To see why this is not optimal, suppose X^a, X^s have a discrete distribution taking either of the values $(0,-1)$, $(0,1)$ with probability $1/2$. Then

$$\begin{aligned} \text{cost}\{\rho_\infty\} &= \frac{\log(2)}{2} \approx 0.346 \\ \text{obj}\{\rho_\infty\} &= \frac{1+e}{1+e+\sqrt{1+6e+e^2}} \approx 0.428 \end{aligned}$$

However, consider some ρ_0 with $\rho_0(0,-1)=0$, $\rho_0(0,1)=1$. Now

$$\begin{aligned} \text{cost}\{\rho_0\} &= \frac{\log(2)}{2} = \text{cost}\{\rho_\infty\} \\ \text{obj}\{\rho_0\} &= \frac{1}{2}\left(\frac{1}{1+e} + \frac{e}{2+e}\right) \approx 0.423 < \text{obj}\{\rho_\infty\} \end{aligned} \tag{27}$$

9.5 Simple example of updating leading to oscillation

Define $g(\rho, x^a)$ as above, and instead define

$$f(x^a, x^s) = \text{logit}\left(-k(x^a + x^s)\right) \tag{28}$$

As usual, we presume that to estimate ρ, we regress Y on X_0^s, X_0^a, and we do it accurately enough to presume ρ is an oracle. Now

$$\begin{aligned} h(x) &= \frac{1}{1+(1+x)^k \exp\left(-k(x^s+x^a)\right)} \\ h'(x) &= -k\frac{e^{k(x^s+x^a)}(1+x)^{k-1}}{\left(e^{k(x^s+x^a)} + (1+x)^k\right)^2} \end{aligned} \tag{29}$$

Consider a setting when $x^s = x^a = 0$ and $k = 8$. Now $h(0) = 1/2 > 0$ and $h(1/5) \approx 0.189 < 1/5$. For $x \in (0,1)$ we have $h'(x) < 0$, so the equation $h(x) = x$ has a single solution in $(0, 1/5)$. But on $(0, 1/5)$, we have $h'(x) < -1$. So if x_0 is the unique root of $h(x) - x$ on $x \in (0,1)$ then $h'(x_0) < 0$

Now as long as $\rho_0(x^s, x^a)$ is not exactly the value of x for which $h(x) = x$, if we update ρ_n using h, it can never converge as the fixed point of the map h is unstable.

Conceptually, although no intervention changes x^a very much, the function f is very sensitive to small changes in x^a when $k = 8$, so a small change in x^a will necessarily cause a larger change in $f(x^a, x^s)$ when ρ is near the fixed point of h.

10 Comparison of solution/avoidance strategies

We briefly compare advantages and disadvantages of the general strategies identified in section 4 to avoid or overcome problems associated with naive updating.

Any of the three strategies can be used to avoid the naive updating problem if they enable an unbiased estimate of

$$\mathbb{E}\left[f_e\left(x^s, x^a, X^\ell\right)\right] \tag{30}$$

to be obtained, where the expectation is over X^ℓ either before or after intervention. The expectation (30) can be recognised as the quantity for which ρ_e is treated as an estimator. More frequent covariate observation as per section 4.1 allows this by enabling observation of $X_e(1)$, so such an unbiased estimate may be obtained by regression of Y_e on observed $X_e(1)$. The strategy in section 4.2 defines a hold-out subset of samples $X_e^\star$, $Y_e^\star$ for which $X^{\star}_e(1) = X^{\star}_e(0)$, so an unbiased estimate of (30) can be obtained by regression of $Y^{\star}_e$ on (observed) $X^{\star}_e(0)$ will work. Finally, the strategy in section 4.3 specifies g^a_e and g^{ℓ}_e, so an unbiased estimate of (30) can be made by regressing Y_e on $X^S_e(0)$, $g^a_e(\rho_e, X^a_e(0))$.

Although all three solutions avoid the problems of naive updating, they 'solve' somewhat different problems and require different experimental designs. The class of strategies described in section 4.1 (a range of modelling approaches generally requiring more frequent covariate observation) can solve the constrained optimisation problem in section 2.3 over ρ. The strategy described in section 4.2 (retention of a 'hold-out' set on which no interventions are made) simply enables unbiased observation of f_e. The strategy described in section 4.3 (explicit control of interventions g^a, g^{ℓ}) solves the constrained optimisation problem over g^a, g^{ℓ}.

However, solutions may be quantitatively compared with an aim of recommending which (if any) might be most appropriate in a given circumstance. If possible, the strategy in section 4.1 should be used if possible, as it enables the greatest flexibility in approach. The strategy in 4.3 should be used alternatively or additionally if appropriate.

The strategy in section 4.2 is advisable as a general approach if covariates cannot be observed more frequently and interventions cannot be controlled (that is, neither of the other strategies are actionable).

10.1 Illustration of solutions

We consider how each strategy may appear in the context of the setting described in Supplementary section 7.

The strategy in section 4.1 would comprise re-observing covariates in February ($t = 1$) after interventions are made. Under this closer observation (allowing inference of g^a and $\mathbb{E}(f)$), ρ_e could be set so as to optimise healthcare provision.

The strategy in section 4.2 would require nomination a random sample of the population on which scores would not be calculated, and hence on which no intervention could be made on the basis of a risk score. This would enable observation of 'native' covariate effects on risk.

The strategy in section 4.3 would implement specific interventions: for instance, 'if $\rho_e > 50\%$, stop drug X'. Interventions could then be tuned to optimise healthcare provision.

11 Open problems

We propose the following short list of open problems in this area.

1. Determine a framework to modulate both g^{ℓ} and g^a with the aim of solving the constrained optimisation problem in section 2.3.
2. Determine the dynamics and consequences of other model-updating strategies. What happens if training data is aggregated at each step, rather than only the most recent data being used?
3. Derive results of successive adjuvancy in more general circumstances.
4. How do the dynamics of the model change when assumptions differ? Can f, g^{ℓ} and g^a be extended to be random-valued, and possibly agglomerated into a single intervention function?
5. How can assumptions be changed to approximate more general machine learning settings?

References

Rahimian, F., Salimi-Khorshidi, G., Payberah, A. H., Tran, J., Solares, R. A., Raimondi, F., Nazarzadeh, M., Canoy, D., and Rahimi, K. (2018). Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLoS Medicine*, 15(11):e1002695.