Summary Measures of Performance for Predictive Models

Dan W Joyce

10th October 2019

Simulated Data

To make the arguments that follow, we use a simulated data set – and to situate in a clinical context – we use the example of predicting 'caseness' of a psychotic disorder (Y) on the basis of a hypothetical biomarker (B), the patient's Age and duration of attenuated psychotic symptoms (DAP). These labels aren't important, but simply make the simulations more relatable.

We construct the simulated data such that the correlation between the biomarker, B, and the outcome Y is 0.75 but for Age and DAP, the correlation with Y is low at 0.2. We construct Age and DAP to be highly correlated such that older people tend to have longer DAP. The outcome is a dichotomised "positive" and "negative" caseness – this simulates the 'ground truth' clinical assessment.

This simulated data represents a realistic example with favourable conditions; the biomarker is strongly associated with outcome while the other variables have little assocation.

The covariance matrix for the simulation is thus:

```
##        B Age DAP    Y
## B   1.00 0.2 0.2 0.75
## Age 0.20 1.0 0.8 0.20
## DAP 0.20 0.8 1.0 0.20
## Y   0.75 0.0 0.0 1.00
```

In the simulated sample of 400, Y is represented as a binary variable which would usually represent the clinical 'ground truth' (e.g. each participant meets criteria for a psychotic disorder, or does not). This is typical of the current literature. Clinicians and patients understand binary categories: a patient "has" or "does not" have a disease, illness or disorder. The discussion of whether mutually-exclusive assignment of binary 'caseness' is appropriate is important, but beyond the scope of this article.

Model development and validation proceeds by simulating a sample from the population (e.g. using the covariance structure above), and then assuming that half of the data was collected and used for training, and the remaining are the testing (validation) set. We have the ideal situation where the training and validation sample are from the same population. This is the most trivial example of out-of-sample validation, but there is no loss of generality for the remaining demonstrations.

We have two separate and well-balanced 200 training and validation samples containing 51 and 55 positive cases respectively.

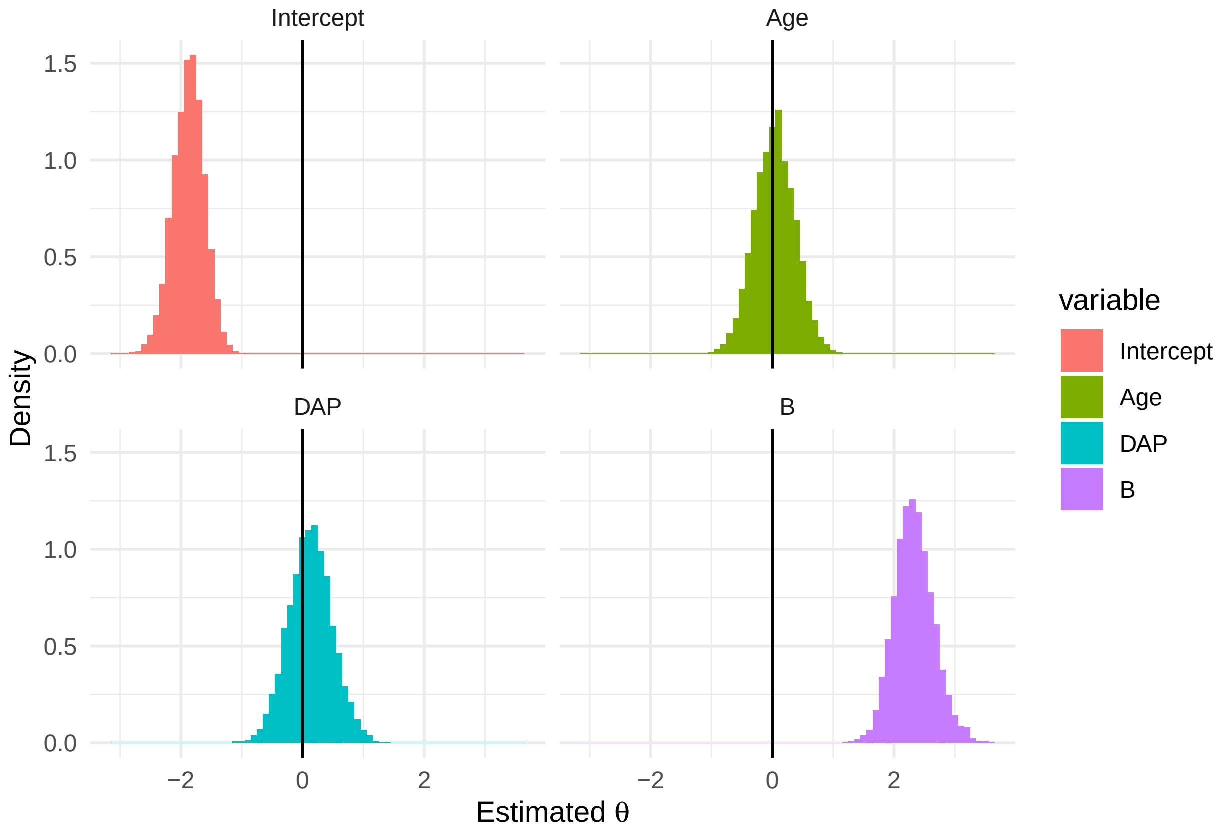
Model Training

We fit the logistic regression model using a contemporary variant of Markov Chain Monte Carlo (via the Stan platform). Using the Bayesian formulation of logistic regression is important because:

- it delivers a *distribution* of values for the parameters (Age, DAP and B or collectively, θ) *given* the training data D – this is the posterior distribution of the parameters and denoted $p_{train}\left(\theta \mid D\right)$. Here, D is the training data consisting of pairs (x, y) where x are the values representing measurements of each patient's Age, DAP and biomarker, B with y representing the patient's caseness (has, or does not have, a psychotic disorder).

- this posterior distribution of parameters explicitly captures uncertainty in the parameter estimates given the training data (and the model, in this case, a logistic regression where y is modelled as a Bernoulli trial)
- the more conventional maximum likelihood estimate for the parameters yields only a single (most likely) estimate for each parameter and means it is harder to explore the impact of model uncertainty on future predictions

Having access to posterior distributions means we can propagate uncertainty in the trained model explicitly in future predictions made with the model.



The plot above shows the posterior distribution of each parameter (given the data, which was mean centered and scaled to unit variance). Let θ_0 be the 'intercept' parameter and θ_B, θ_{Age} and θ_{DAP} be the estimated model parameters for the biomarker B, Age and DAP respectively.

Inspecting the plot, we can see that:

- For each parameter, the 'mass' of the distribution is centered (i.e. the mode) over the most likely value for the parameter and the dispersion (i.e. the 'width' of the distribution) is proportional to the uncertainty. As an example, we can be confident that the parameter θ_B has a value of between 1.5 and 3.2 with the most likely value being around 2.2.
- The posterior distributions allow one to directly answer questions such as "what is the probability that the intercept is effectively 0?" or "what is the probability that B (the coefficient for the biomarker) has a value great than 2?"
- In contrast, a traditional (e.g. maximum likelihood) method would give us only a **single, point**

estimates for the parameters for the intercept, Age, DAP and B. For example, the mode of the above distributions. So, we would instead have discrete values θ_0 = -1.9, θ_B = 2.2, $\theta_{Age} = 0$ and $\theta_{DAP} = 0.1$.

Making Predictions

To make the discussion of prediction concrete, we take a single new patient (not in the training set, D) for whom we want a prediction about their being a positive (conversely, a negative) case. The patient has measured biomarker, Age and DAP (centered and scaled rather than in their original units) that we denote x_{new} and we want the model to provide a continuous score y_{new} that is proportional the probability the new patient is a positive case. For any given patient, the continuous score y_{new} will depend on the 'trained' model parameters θ and the values of the predictor variables x_{new}.

In what follows, we will use a single new patient as an example, where the values x_{new} are:

```
##     B Age DAP
## 1 0.2 0.9 0.7
```

Using Only Point-Estimates for Model Parameters

To make a prediction, we would 'feed' these values for x_{new} into the trained logistic regression model and obtain the continuous score y_{new}. If we use the point estimates for the coefficients of the logistic model (θ_0, θ_B, θ_{Age} and θ_{DAP}) we can make a prediction:

$$\begin{aligned} y_{new} &= \frac{1}{1+\exp\left(-(\theta_0 + \theta_B B + \theta_{Age} Age + \theta_{DAP} DAP)\right)} \\ &= \frac{1}{1+\exp\left(-(-1.9 + 2.2 \times 0.2 + 0 \times 0.9 + 0.1 \times 0.7)\right)} \end{aligned}$$

Resulting in $y_{new} = 0.2$

In this classical point-estimate approach, we end up with a **single continuous score** for the new patient but there is no accounting for model uncertainty. Further there is no account of *observation uncertainty* if we equate the continuous score with the probability of x_{new} being a positive case. In the case of prediction, this is not unreasonable because by definition, logistic regression is designed to model the probability of an event (being a positive case) given the parameters and predictor variables.

We will return to this example shortly, but note that this continuous score y_{new} (for any new patient) is then traditionally compared to an operating threshold to determine the 'caseness' (positive or negative) to obtain the familiar summary measures of performance such as sensitivity, specificity, accuracy and so on.

Prediction Using the Posterior Distribution of Parameters

Recall that the posterior distribution of parameters $p_{train}\left(\theta \mid D\right)$ contains information about the uncertainty in the model's parameter estimates and this can be propogated forward and inform our confidence in predictions for a given patient. The model was trained using MCMC to deliver samples from the posterior distribution $p_{train}\left(\theta \mid D\right)$.

Predictions for new patients $\tilde{D}$ (e.g. in either validating the model, or deployment) arise from samples from the **posterior predictive distributions** (PPDs):

$$p(\tilde{D} \mid D) = \int p_{pred}\left(\tilde{D} \mid \theta\right) p_{train}\left(\theta \mid D\right) d\theta$$

After training, we have access to $1 \dots M$ samples from the posterior distribution $p_{train}\left(\theta \mid D\right)$ – shown in the histograms above – with a single sample denoted $\theta^{(m)}$:

$$\theta^{(m)} \sim p_{train}\left(\theta \mid D\right)$$

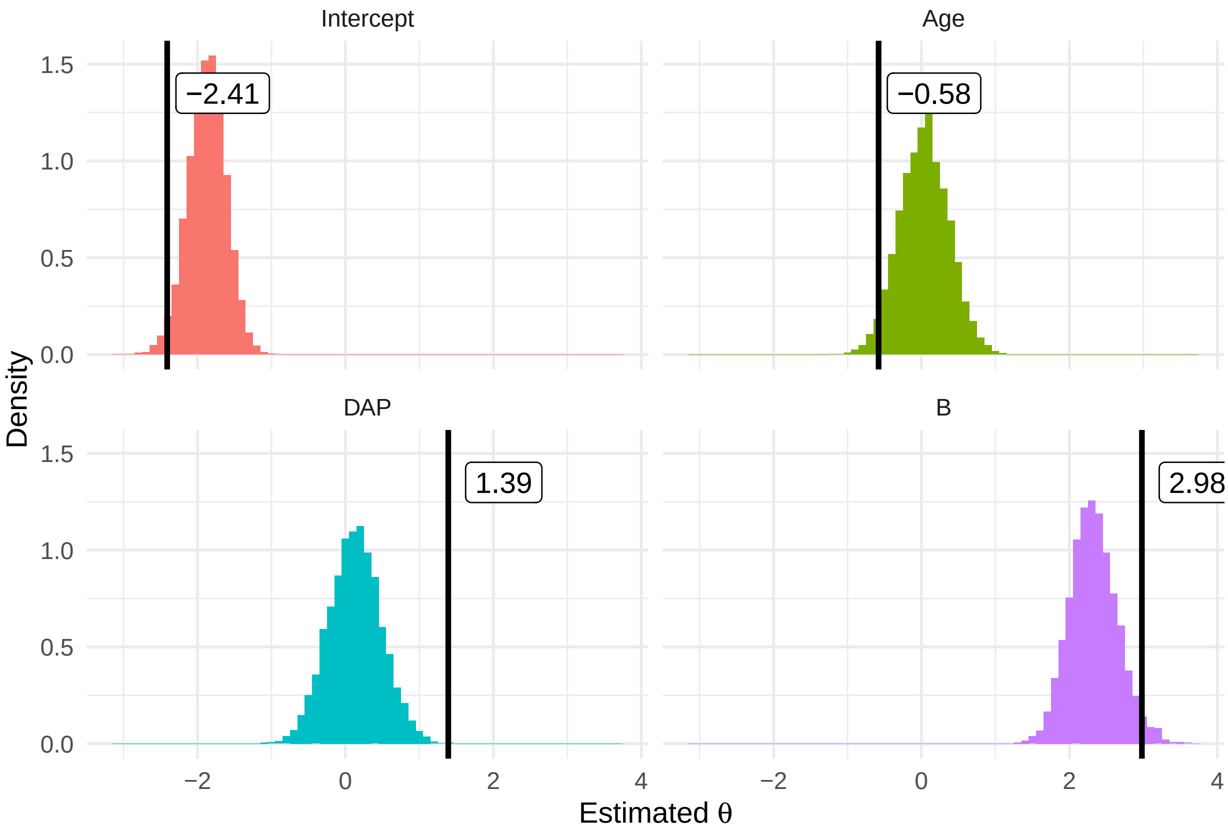
Using $\theta^{(m)}$ we then simulate a sample for y_{new} – the 'output' of the trained model for x_{new} (a realisation of the new data, $\tilde{D}$):

$$y_{new}^{(m)} \sim p_{pred}\left(x_{new} \mid \theta^{(m)}\right)$$

And we can then visualise and summarise the resulting posterior distribution of $y_{new}^{(m)}$.

Using the concrete example for x_{new} above, we proceed as follows:

1. Obtain a single sample of the model parameters $\theta^{(m)}$ from the posterior distribution of the parameters $p_{train}\left(\theta \mid D\right)$ – in the diagram below, the black lines indicate one sample for each parameter (Intercept, Age, DAP and B)



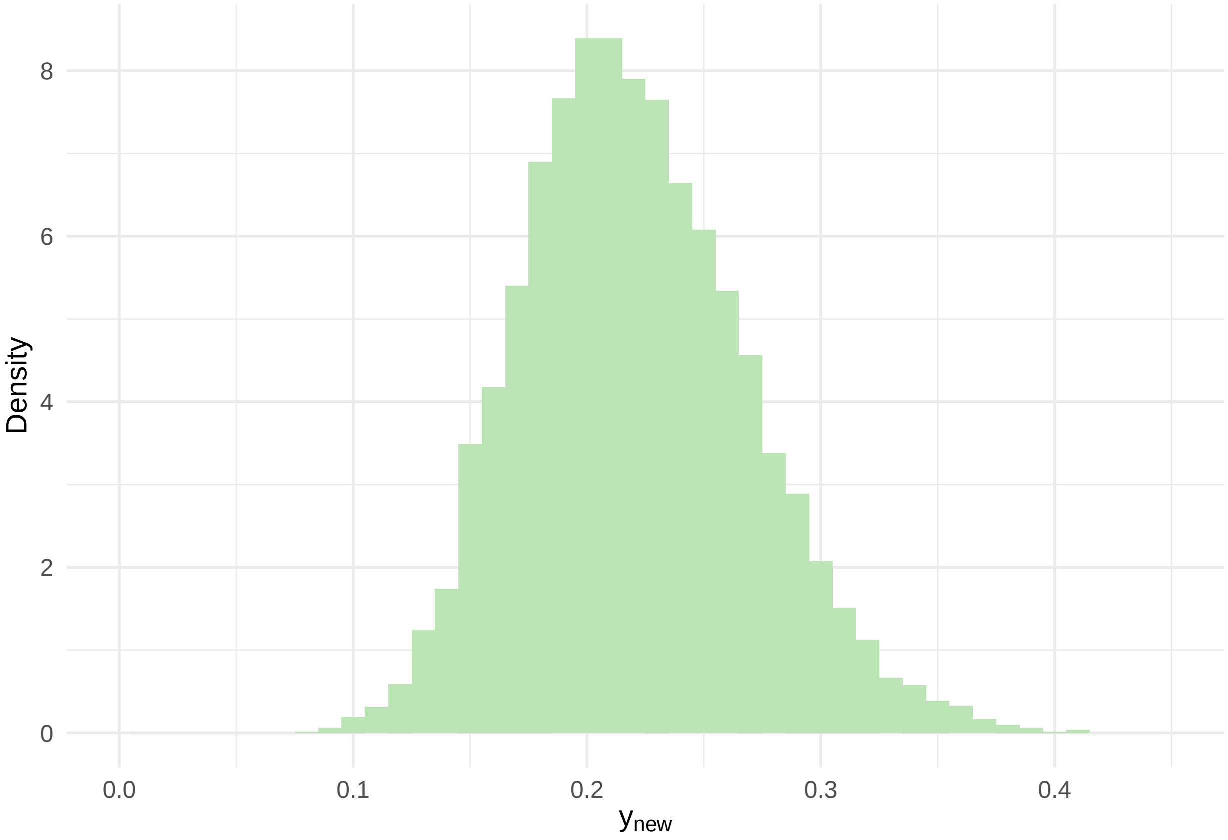
2. With this sample of the model parameters, 'run' the model with the new patient x_{new} to obtain the continuous score $y_{new}^{(m)}$:

$$\begin{aligned} y_{new}^{(m)} &= \frac{1}{1 + \exp(-(\theta_0^{(m)} + \theta_B^{(m)} B + \theta_{Age}^{(m)} Age + \theta_{DAP}^{(m)} DAP))} \\ &= \frac{1}{1 + \exp(-(-1.78 + 2.38 \times 0.2 + (-0.13 \times 0.9) + 0.28 \times 0.7))} \end{aligned}$$

Yielding $y_{new}^{(m)} = 0.2$

3. Repeat 1) and 2) for all $\theta^{(m)}$
4. Histogram the resulting M values of $y_{new}^{(m)}$

For the case of the single patient above, this results in a distribution of the probability of being a positive case given the uncertainty in the model:



Summary

- If we use only a point estimate of the model parameters (e.g. from a model trained by maximum likelihood) we can usually only obtain a single point estimate for the probability of any new patient being a positive case y_{new}
- The Bayesian formulation yields a distribution of values for y_{new} given the uncertainty in the posterior distribution of parameters $p_{train}(\theta \mid D)$

Decisions

Point Predictions and Dichotomising Decision Rules

For a given new patient x_{new} where we have a point estimate y_{new} we can compare this to a cutoff (operating threshold) and obtain a discrete answer to the question "Is this new patient a positive or negative case?"

This decision rule, $y_{new} >$ cutoff, assigns every new patient to a binary "yes" or "no" label. When validating the model, we want to know how well the model is performing. We know the assignment of positive/negative cases in the validation set, so for example, if a patient from the validation sample is assigned "yes" by the decision rule – when in fact the patient was a negative case – this represents a false positive.

Point Predictions using Bayesian Posterior Distributions

In contrast, when we have samples $y_{new}^{(m)}$ (the histogram above) we require a summary of these values to declare a positive or negative case. In Bayesian statistics, these point summaries are a consequence of the choice of *loss function* in a statistical decision-theoretic framework. Common loss functions give rise to familiar summary statistics, for example:

- If we use a quadratic loss function, the summary *single* value for y_{new} can be shown to be the expected value (mean) of $y_{new}^{(m)}$ which is 0.221
- If we use a linear loss function we can define quantiles of $y_{new}^{(m)}$, the most familar being the median which is 0.217
- If we use a zero-one loss function, the summary can be shown to be the mode of $y_{new}^{(m)}$ which is 0.205

With these summary statistics, we can now apply the cutoff (decision rule) if we want to obtain a discrete 'yes'/'no' prediction. However, as we show below, this may not be desirable when a predictive model is deployed clinically.

Interval Predictions using Posterior Distributions

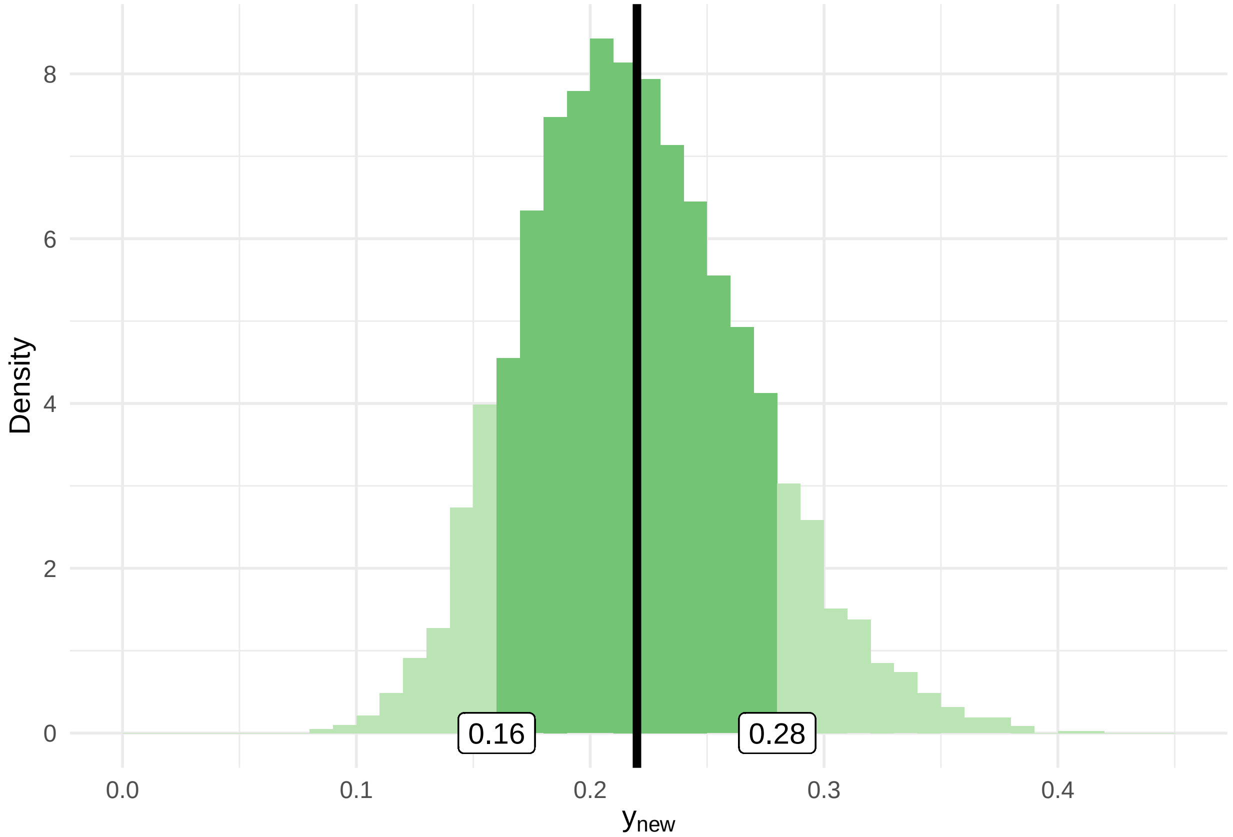
We can also summarise the distribution of $y_{new}^{(m)}$ by making statements about uncertainty in the predicted value. With reference to the example above, for a single patient x_{new}:

- This patient is likely negative, because the probability that the patient is a positive case is low – the median predicted value y_{new} is 0.22

Given the model, we can estimate uncertainty on this prediction by computing a **credible interval** – for example, the 80 percent (or 0.8) credible interval for the above patient's prediction is the number of samples $y_{new}^{(m)}$ in the quantile [0.1, 0.9]. The credible interval can be interpreted as:

- there is an 80% probability that the predicted value of y_{new} is in the interval [0.16, 0.28]

The same information can be presented graphically as:



Where the darker shaded green area highlights the 80% credible interval and the black solid lines shows the median.

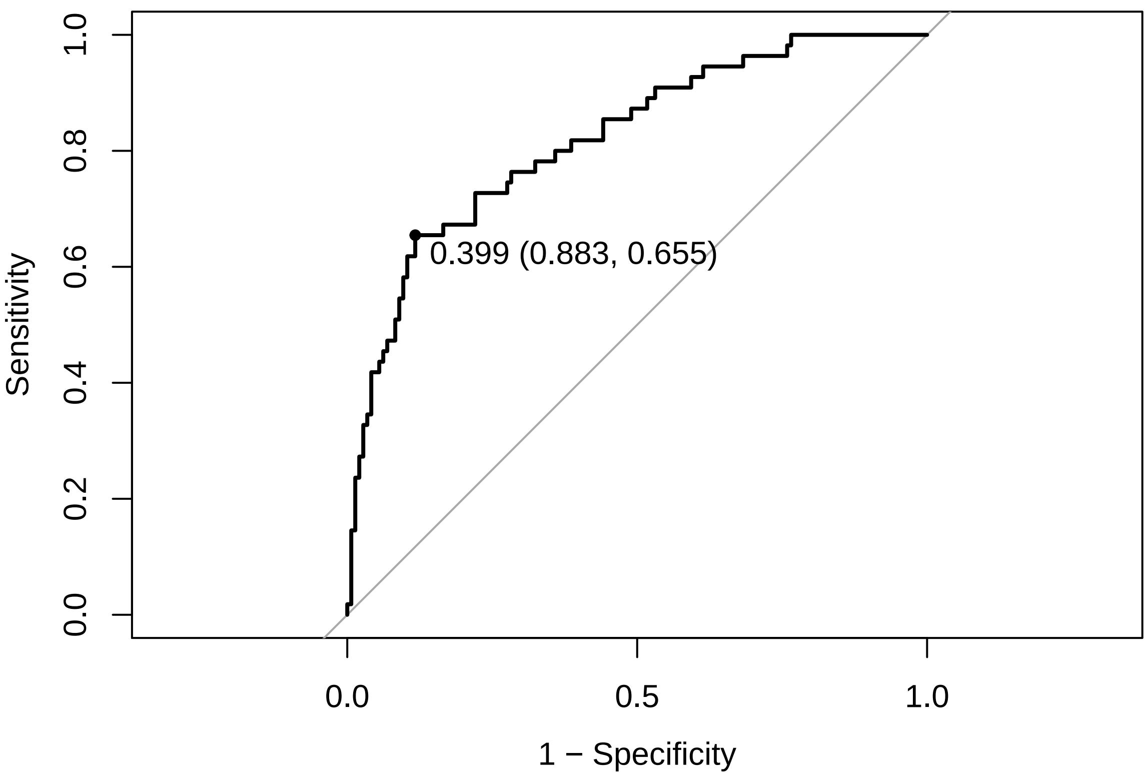
Summary Measures of Performance (SMP)

In the preceding sections, we showed how using samples and simulations from posterior distributions allows propogation of model uncertainty into individual predictions in contrast to point summaries.

We now consider the classical SMP found in a majority of the literature on predictive models. For each patient in the validation sample, the trained model yields a *continuous score* y_{new} (in this case, the probability of a psychotic disorder given their Age, DAP and biomarker level, B). This score is then compared to a *threshold* (the operating threshold, or 'cutoff') and then declared to be a positive (or negative) case.

Most often, the classical SMP are derived using a cutoff found by locating the point on the receiver operating characteristic (ROC) curve that *maximises* the balance or tradeoff of true positives (sensitivity) against false positives (1-specificity) in the training or validation data set. Here, we find the cutoff using the validation data because we want performance to be optimised (but not over-optimistic) on data not seen during training that we propose will be representative of patients seen when the algorithm is deployed. Note, we also tried optimising the cutoff using the training set; this resulted in a lower operating threshold (approximately 0.29) and increased the SMP further e.g. AUROC = 0.91, sensitivity of 0.67, specificity of 0.81, balanced accuracy of 0.74 and Brier score of 0.14.

So, while we *have* access to samples representing the posterior distribution for each patient in the validation set, deriving a ROC curve *requires* that we obtain and use a single point prediction for each patient (i.e. to determine the true/false positives and negatives). Taking the mean of the posterior samples $y_{new}^{(m)}$ for each patient's prediction, we can derive the following ROC curve and optimal operating threshold (cutoff):



With the optimal operating threshold (cutoff) fixed at 0.399, we then derive and report the resulting validation set performance using the 'classic' SMP found in most of the literature:

- AUROC (area under the ROC curve) = 0.82
- Sensitivity of 0.65 and a specificity of 0.88
- Balanced accuracy of 0.77
- Somers' D_{xy} of 0.54 (with -1 and +1 representing complete dis-agreement and complete agreement respectively, between model predictions and the validation set patient's actual caseness)
- Brier score of 0.14 (with 0 being the best, 1 being worst)

Of all of these measures, only the Brier score uses the predicted continuous score y_{new} for each patient in the validation set – the others depend on first applying the decision rule $y_{new} >$ cutoff, where each case is predicted to be only positive (or negative).

Summary

Summary measures of performance (SMP) attempt to capture the expected performance when the trained model is deployed with new (unseen) data, however:

- Most require a dichotomised (positive/negative) prediction – the exception being the Brier score
- Consequently, a point-prediction is required – i.e. a single-valued summary of the posterior predictive distribution (e.g. the mean)
- Information about model uncertainty – contained in the posterior predictive distribution – is therefore discarded and cannot be used when evaluating a prediction for an individual patient

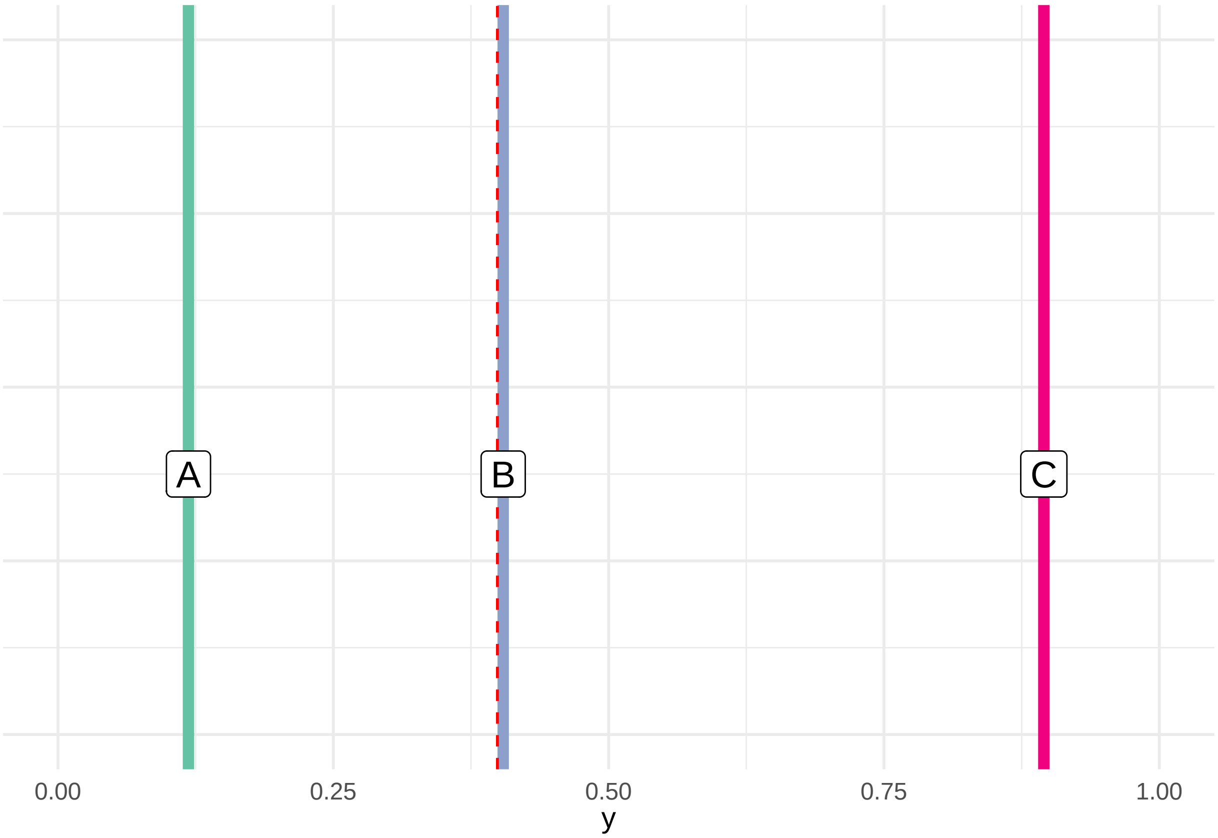
Predictions for Deployment

In the example above, according to the SMP, the predictive model achieved reasonable performance on the validation data, not seen during training. On the basis of this, one might be persuaded of the clinical utility of a model. We propose that when deployed in support of clinical decision making, the same measures and decision rules will obscure important information relevant to the patient and clinician.

To illustrate, we take 3 example patients (from the validation set) and explore how predictions are made using either point predictions (as is done with the SMP) and using the posterior distributions.

Deployment : Point Predictions

First, consider the information available when we submit three new patients (A, B and C) to the trained model and use the same point-predictions that are used to derive the SMP:



Each patient is assigned a predicted value y as shown in the plot above. The red dotted line represents the optimal operating threshold (cutoff) that yielded the SMP reported for the trained model. Using this scheme, we can state:

- Patient A : $y_A = 0.118$ which is **less** than the cutoff 0.399 so Patient A is **negative**
- Patient B : $y_B = 0.404$ which is **greater** than the cutoff 0.399 so Patient B is **positive**
- Patient C : $y_C = 0.895$ which is **greater** than the cutoff 0.399 so Patient C is **positive**

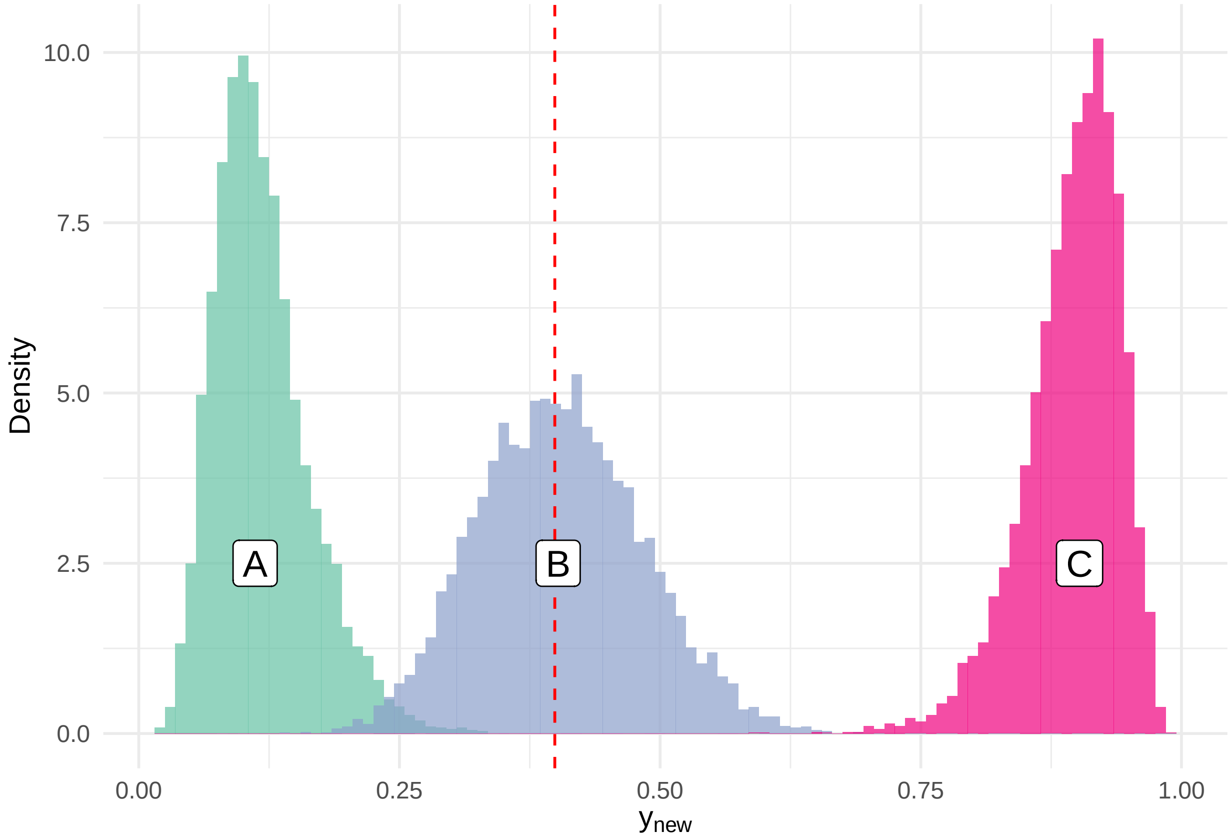
It is clear that for Patient B, the margin is narrow : y_B is greater than the threshold by only around 0.005.

This illustrates that dichotomising the prediction (using the decision rule $y_{new} >$ cutoff as we required to derive the SMP) discards information contained in the actual prediction delivered by the trained model.

There is no accounting for uncertainty.

Deployment : Posterior Distributions

Instead of taking a single point prediction, we now repeat the process outlined in Section 4.3 for the 3 new patients, A, B and C. The posterior distributions are:



The red dotted line again shows the cutoff used to derive the SMP quoted above.

Now, we can answer the following questions relevant to the predictions for each patient using quantiles of $y^{(m)}$ for A, B, and C:

- The **probability** that Patient A is **positive** is the number of samples $y_A^{(m)}$ greater than the cutoff = 0 – it can clearly be seen that the mass of the posterior distribution for A is concentrated far below the cutoff.
- The **probability** that Patient B is **positive** is the number of samples $y_B^{(m)}$ greater than the cutoff = 0.517
- The **probability** that Patient C is **positive** is the number of samples $y_C^{(m)}$ greater than the cutoff = 0

Alternatively, we can answer questions about the credible interval (here, the 80% interval) for the probabilities of being positive:

- For Patient A, there is an **80% probability** that the predicted value of being positive is between [0.066, 0.181]
- For Patient B, there is an **80% probability** that the predicted value of being positive is between [0.304, 0.508]
- For Patient C, there is an **80% probability** that the predicted value of being positive is between [0.834, 0.946]

Any treatment decision for Patient B should account for the obvious uncertainty in the trained model's prediction for them.

Summary

- The measures used to derive SMP discard information contained in the posterior predictive distribution
- When using a dichotomising decision rule, there are cases where prediction uncertainty is directly relevant (for example, those near the cutoff)

Visualising Overall Accuracy for Predictions

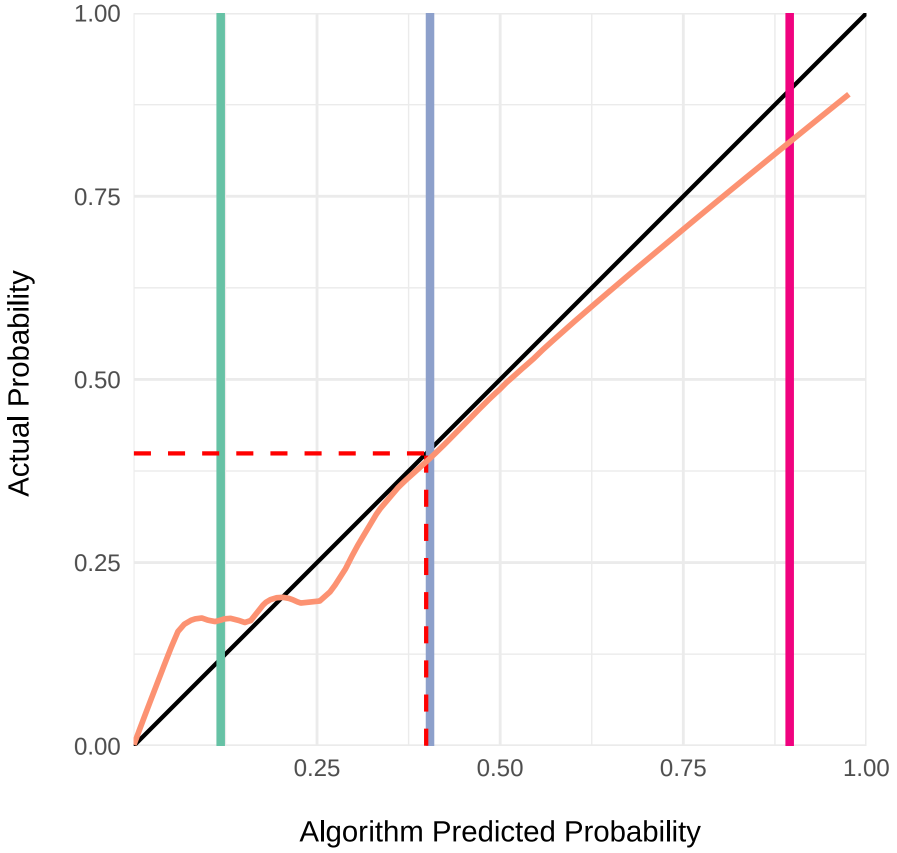
In the preceding sections, we've addressed the differences between point predictions, dichotomising decision rules and using the 'full' posterior distribution for new individual patients.

We might still require an illustration of how predictions from the trained model compare over a population sample. Traditionally, the **calibration curve** has been used to display this information. A calibration curve is constructed by scatterplotting the validation data as follows:

- on the horizontal axis, plot the **predicted** values for each case – i.e. those obtained from the trained model
- on the vertical axis, for each case, plot the **actual** probability of being positive
- fit a smoothing curve through the scatterplot
- "perfect" calibration is the straight 45-degree diagonal line from (0,0) to (1,1)

Calibration for Point Predictions

Here, we show the calibration curve for the point-prediction version of the model:



Inspecting the calibration plot, we can see that:

- When the model predicts that the probability of being positive is **low** (to the left of the horizontal axis), in the validation set, the actual probability is marginally higher – the model is under-estimating the probability of being positive
- To the right of the horizontal axis, when the model is predicting the probability of being positive is **high**, in the validation set, the actual probability is somewhat lower – the model is over-estimating the probability of being positive
- Near the decision-rule cutoff of 0.399 (the red dotted lines) the model appears well-calibrated

The coloured vertical lines correspond to the point predictions for patient A, B and C.

Calibration Plot incorporating Uncertainty in the Model

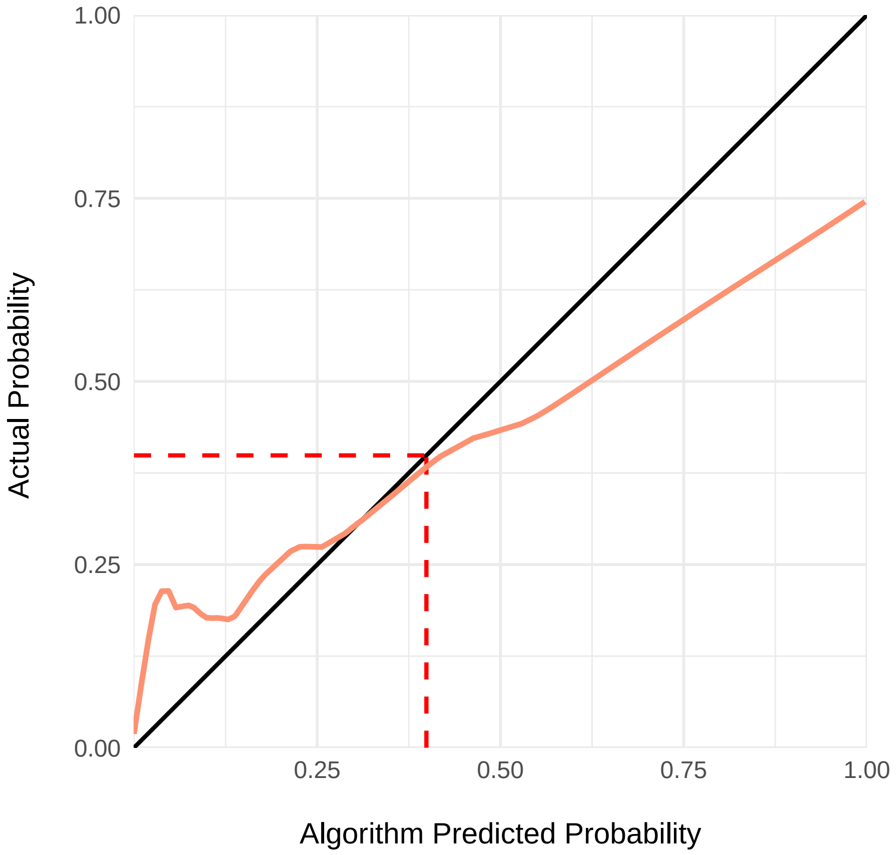
When making individual predictions we propogated uncertainty in the trained model using the posterior distribution of the parameters and showed how credible intervals can be located for the predicted probabilities.

We can similarly visualise uncertainty in calibration by constructing a calibration curve for *each* of the M samples $\theta^{(m)}$ from $p_{train}\left(\theta \mid D\right)$.

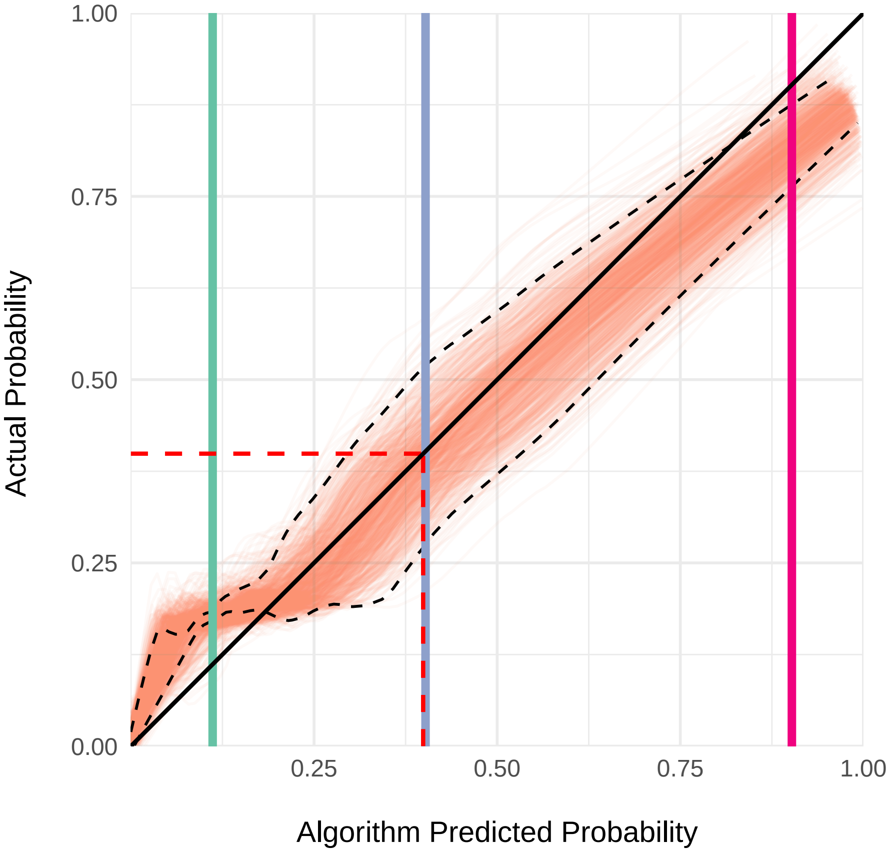
The procedure is as follows:

1. Take a sample of the parameters $\theta^{(m)} \sim p_{train}\left(\theta \mid D\right)$ – for example, in Section 3.2, we took a sample for the parameters (intercept, Age, B and DAP) where θ_0 = -2.41, θ_{Age} = -0.58, θ_B = 2.98 and θ_{DAP} = 1.39

2. With this parameter sample, compute the calibration curve **for every** patient in the validation set $\tilde{D}$



3. Repeat steps 1 and 2 to derive and plot the calibration curve for all M samples $\theta^{(m)}$



In addition to the spaghetti plot of individual calibration curves, the black dotted line shows the 80% credible interval of the calibration curves and we have plotted the median of the predictive distribution for the three example patients (A, B and C).

In contrast to the point-prediction calibration we can see that:

- For negative cases toward the left of the horizontal axis, the model systematically under-estimates the actual probabilities but fairly consistently
- Near the cutoff, there is substantial variation in the calibration when we account for uncertainty in the model
- For positive cases toward the right of the horizontal axis, the model systematically over-estimates the actual probabilities but again, fairly consistently

When we have an individual prediction for a new patient, we can not only see the uncertainty in the model's prediction (as in Section 6.2), but can compare that prediction with the model's overall calibration for other patients (in the validation set) with similar predictions.