Calibrating predictive model estimates to support personalized medicine

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, Lucila Ohno-Machado

▸ An additional appendix is published online only. To view this file please visit the journal online (http://jamia.bmj.com/content/19/2.toc).

Division of Biomedical Informatics, School of Medicine, University of California, San Diego, La Jolla, California 92093, USA

Correspondence to
Dr Xiaoqian Jiang, Division of Biomedical Informatics, School of Medicine, University of California, San Diego, La Jolla, CA 92093, USA: x1jiang@ucsd.edu

Received 31 March 2011
Accepted 8 September 2011
Published Online First
7 October 2011

ABSTRACT

Objective Predictive models that generate individualized estimates for medically relevant outcomes are playing increasing roles in clinical care and translational research. However, current methods for calibrating these estimates lose valuable information. Our goal is to develop a new calibration method to conserve as much information as possible, and would compare favorably to existing methods in terms of important performance measures: discrimination and calibration.

Material and methods We propose an adaptive technique that utilizes individualized confidence intervals (CIs) to calibrate predictions. We evaluate this new method, adaptive calibration of predictions (ACP), in artificial and real-world medical classification problems, in terms of areas under the ROC curves, the Hosmer-Lemeshow goodness-of-fit test, mean squared error, and computational complexity.

Results ACP compared favorably to other calibration methods such as binning, Platt scaling, and isotonic regression. In several experiments, binning, isotonic regression, and Platt scaling failed to improve the calibration of a logistic regression model, whereas ACP consistently improved the calibration while maintaining the same discrimination or even improving it in some experiments. In addition, the ACP algorithm is not computationally expensive.

Limitations The calculation of CIs for individual predictions may be cumbersome for certain predictive models. ACP is not completely parameter-free: the length of the CI employed may affect its results.

Conclusions ACP can generate estimates that may be more suitable for individualized predictions than estimates that are calibrated using existing methods. Further studies are necessary to explore the limitations of ACP.

INTRODUCTION

Predictive models are increasingly being used in clinical practice (eg, risk calculators based on the Framingham Study produce estimates for the probability of a particular individual developing cardiovascular disease in the next 10 years, while others based on a variety of different studies produce estimates for the development of breast cancer,[1] or mortality during hospitalization in an ICU[2]). In predictive models based on binary outcomes, the outputs constitute probability estimates that the event of interest will occur (eg, a particular patient has an 8% chance of having myocardial infarction given her risk factors). In this context, we measure the calibration of the individualized prediction by checking how close this prediction is to the true underlying probability of the event for that particular patient. Given that each patient is unique, it is not possible to determine what this true underlying probability is, and therefore certain proxies have to be used, such as the probability of the event in a group of similar individuals. If the prediction is close to the proportion of events in this group, then the individualized estimate is considered well calibrated. Calibration is important for these types of personalized medicine tools, since estimates (ie, predictions) are often used to determine a patient's individual risk.[3–5] A high risk can guide important clinical decisions, such as initialization of anti-lipid pharmacotherapy for an individual at high risk for cardiovascular disease,[6 7] or referral for chemoprevention trials for a woman with high chances of developing breast cancer.[8] Outside the USA, some authors have proposed the use ICU mortality calculators for critical decisions such as discontinuation of certain types of therapy.[9] As molecular markers from genomics and proteomics start to be incorporated into predictive models and become directly available to consumers,[10–12] understanding the shortcomings of individualized predictions and developing new methods to calibrate individual predictions becomes paramount. Calibration is even more crucial to ensure accurate probability estimations in personalized medicine, which includes individualized estimates for risk assessment, diagnosis, therapeutic intervention success, and prognosis.[13]

Oftentimes, adequate calibration is coupled with adequate discrimination in a predictive model; however, a highly discriminative classifier (eg, a classifier with a large area under the receiver operating characteristic (ROC) curve, or AUC[14]) is not necessarily well calibrated.[15] For example, a model that predicts all positive outcomes (ie, those with outcome labels '1') to occur with probability 0.99 and all negative outcomes to occur with probability 0.98 has perfect discrimination, but will have poor calibration because negative predictions are probably too high, and therefore, miscalibrated. Several machine learning approaches, for example, naive Bayes and decision tree, have been shown by other authors to have poor calibration in a variety of datasets.[16 17] Even logistic regression (LR) models, which are widely used in medicine, are not always well calibrated. Consequently, several methods have been proposed to improve the calibration of popular statistical and machine learning models.[17–19]

Zadrozny and Elkan applied *binning* to smooth predictions.[17] The method calibrates probability estimates produced by a given predictive model using histograms. Specifically, we first sort the predicted values of a model and divide them into 10 equal size groups, which are called *bins*. Any point whose original estimate falls within the upper and lower bounds of a single bin then receives a probability estimate (ie, prediction) that equals the fraction of positive cases in the bin, no matter whether it was close to the lowest or highest estimate in the bin. A major shortcoming of this approach is thus that at most 10 different estimates are produced for all cases, and the discrimination in each bin is no longer preserved. Alternatively, Platt suggested a parametric approach that transforms classifier's outputs into posterior probabilities[18] by fitting these outputs to a sigmoid function (ie, the estimates produced by a predictive model are transformed by the logistic function). The parameters of the sigmoid function are estimated using maximum likelihood estimation. However, the method is not likely to produce adequate probability estimates if the predictive model estimates are distributed in a biased fashion (eg, at the extremes, or all near the separating plane). To address the shortcomings of *binning* and *Platt scaling*, Zadrozny and Elkan[19] proposed another calibration approach that utilizes isotonic regression (IR). Their method involves finding a weighted least square fit $\{\tilde{y}_i\}_{i=1}^{n}$ with the following objective function: $\min_{\{\tilde{y}_i\}} \sum_{i=1}^{n} (\tilde{y}_i - y_i)^k$ subject to $\tilde{y}_i \le \tilde{y}_{i+1}$ and $\hat{y}_i \le \hat{y}_{i+1}, \forall_i$. Here $\{\hat{y}_i\}_{i=1}^{n}$ are pre-calibration probability estimates that are used to order the class labels $\{y_i\}_{i=1}^{n}$. When the tunable parameter k equals 2, an efficient pair-adjacent violators algorithm can be used to solve the problem in $O(n)$,[20] but efficient solutions for other values of k do not exist. Furthermore, as other authors have pointed out, the results of the IR are not continuous and tend to overfit the training data.[21] Note that both *IR* and *Platt scaling* use monotonic transformations of a model's predictions, preserve their rankings, and consequently preserve their AUCs.

To tackle the limitations of existing methods, we investigated an alternative generalized approach that uses individualized CIs to improve calibration without increasing model complexity. Because the CI can be calculated based on the local density of training cases in the neighborhood of a test case, our approach is applicable to any probabilistic predictive model. In this article, we limit our discussion to the widely used LR model, whose parameter estimations are straightforward. The calibration procedure for other learning models can be designed in a similar manner.

Specifically, we process each prediction by first finding a subset of training cases (labeled) whose predictions fall into the CI of the test case prediction that is being processed. We then substitute the test case prediction by the fraction of positive cases in this subset. We use a small subset of cases when the predictive model is confident about the prediction, and we use a large subset of cases when the predictive model is less confident about the prediction. Our method, adaptive calibration of predictions (ACP), therefore uses a non-monotonic transformation to calibrate predictive models. Figure 1 illustrates the adjusted estimations of probabilities using four different calibration approaches and the predictions of an LR model on a linearly separable dataset.

METHODS

A brief review of LR is provided in the online supplementary appendix. LR produces a probability estimate of a binary outcome for each case, as well as a CI for this estimate. For example, it can produce a risk estimate of 18% for development of cardiovascular disease for a patient with certain blood pressure and cholesterol measurements, smoking status, age, and gender. The 95% CI for this prediction is, for example, (10% to 26%). In our method, we first collect all cases whose predictions fall within the aforementioned CI, then take the average of their class labels to calibrate this particular prediction, obtaining a calibrated estimate of 12% for this individual.

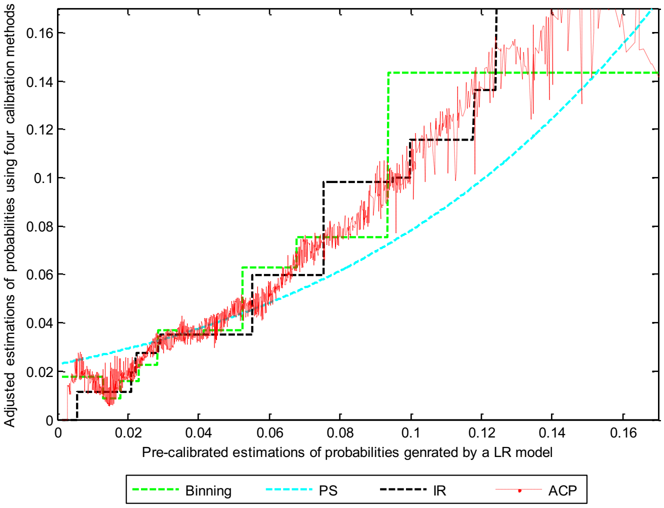


Figure 1 Illustration of calibration functions of four different approaches, including binning, Platt scaling (PS), isotonic regression (IR), and our proposed method, adaptive calibration of predictions (ACP). LR, logistic regression.

Adaptive calibration for LR

Cases that are close to each other should have approximately the same estimated probabilities in a calibrated model.[22] Intuitively, if we want to estimate $f(X^*)$ for a novel case, X^*, we can select a neighborhood of X^* and calibrate the raw probabilistic estimate of $P(Y^*=1|X^*)$ using cases taken from this neighborhood. A simple estimator is therefore:

$$f(X^*) = \frac{1}{|\mathcal{N}(X^*)|} \sum_{X^i \epsilon \mathcal{N}(X^*)} Y^i \qquad (1)$$

where X^i corresponds to the i-th neighboring case of X^* and Y^i corresponds to its class label (ie, binary outcome). Here $\mathcal{N}(X^*)$ denotes the neighborhood of X^*. Depending on the construction criteria for this neighborhood, equation (1) could represent, for example, a nearest neighbor estimator if we select a fixed number of n cases, or a Parzen window estimator if we choose a fixed bandwidth, for example, ϵ=max($|X^*–X^i|$) s.t. $\forall X^i \in \mathcal{N}(X^*)$. Given a n or ϵ, the estimator induced by equation (1) corresponds to the fraction of positive cases. However, it is non-trivial to select a single n or ϵ for the entire test population. In the first place, the computational complexity would be high because these estimators need to find the neighborhood for every test case X^* at run-time, as in the method proposed by Osl *et al*.[13] Furthermore, there might not be a single n or ϵ that works well for all test cases.

We propose ACP to overcome these difficulties. For a generalized linear model $f'(X)=g(W^TX)$, including LR, where W stands for weight parameters and $g(\cdot)$ is a link function, we can infer the variances Σ of W=$<\omega_0,\ldots,\omega_K>$ in addition to the means (ie, the coefficients). These variances are used to produce the CIs for individual predictions. Specifically, the standard deviations on each dimension of the parameter vector are multiplied by their corresponding attribute values for a subject (ie, test case), and then transformed through the inverse logit function (or alternatively converted through the delta method[23]) to obtain the desired CI for the prediction for that particular case. We then calibrate the probability estimates of a predictive model as follows

$$P_{acp}(Y^*=1|X^*)=\frac{1}{|CI'(X^*)|}\sum_{\{i:P(Y^i=1|X^i)\epsilon CI'(X^*)\}} Y^i \qquad (2)$$

where $|CI'(X^*)|$ is the 95% CI for a prediction $P(Y^*=1|X^*)$, and $|CI'(X^*)|$ denotes the total number of points whose predicted values are included in this interval. This above formula is directly applicable for situations in which the predicted probabilities of the training data range from zero to one. In this case, we can use a 95% CI to obtain the $|CI'(X^*)|$.

However, using a fixed CI, $|CI'(X^*)|$ can be problematic if estimated probabilities of the training data cover a much narrower range, that is, $[a, b]$ such that $a>0$ and $b<1$. The problems are: (1) the 95% CI of the test case could easily fall out of the range r=$[a, b]$ where no training cases exist; and (2) the 95% CI could be too wide, covering the entire spectrum of estimated probabilities of training cases and therefore making all calibrated predictions have the same value (ie, adjusted estimates of probabilities for test cases simply equal to the fraction of positive cases among all the training data).

To address these problems, we need to adjust $|CI'(X^*)|$. We can express the LR model as

$$Z(P(Y=1|X)) = \ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = w_0 + \sum_{k=1}^{K} w_k x_k \qquad (3)$$

where $P(Y = 1|X)$ is the estimated probability for a given X. Here $Z(\cdot)$ denotes the logit function. Let x_0=1, then

$$Z(P(Y=1|X)) = \sum_{k=0}^{K} w_k x_k \qquad (4)$$

The LR model provides estimated parameters W and their co-variances Σ. We can thus compute the co-variance matrix

$$\mathrm{var}\left(\sum_{k=0}^{K} w_k x_k\right) = \left(\Sigma^{\frac{1}{2}}\right)'X^2\left(\Sigma^{\frac{1}{2}}\right) \qquad (5)$$

and the 95% CI for a given observation is $Z(P(Y=1|X))\pm 1.96*\sqrt{\left(\Sigma^{\frac{1}{2}}\right)'X^2\left(\Sigma^{\frac{1}{2}}\right)}$. As mentioned earlier, this CI could be too wide for ACP when the range of estimated probabilities r = $|b-a|$ is smaller than one. Therefore, we rescale the CI to be $CI = Z(P(Y=1|X))\pm 1.96*\mathrm{r}*\sqrt{\left(\Sigma^{\frac{1}{2}}\right)'X^2\left(\Sigma^{\frac{1}{2}}\right)}$. As $P(Y^*=1|X^*) = \frac{e^{Z(P(Y^*=1|X^*))}}{1+e^{Z(P(Y^*=1|X^*))}}$, we can convert the CI of $Z(P(Y^*=1|X^*))$ into the probability using the inverse logit function.[24] The adjusted CI for $P(Y^*=1|X^*)$ is,

$$CI'(X^*) = \left(\frac{e^{Z(P(Y^*=1|X^*))-1.96*\mathrm{r}*\sqrt{\left(\Sigma^{\frac{1}{2}}\right)'(X^*)^2\left(\Sigma^{\frac{1}{2}}\right)}}}{1+e^{Z(P(Y^*=1|X^*))-1.96*\mathrm{r}*\sqrt{\left(\Sigma^{\frac{1}{2}}\right)'(X^*)^2\left(\Sigma^{\frac{1}{2}}\right)}}}, \frac{e^{Z(P(Y^*=1|X^*))+1.96*\mathrm{r}*\sqrt{\left(\Sigma^{\frac{1}{2}}\right)'\left(X^*\right)^2\left(\Sigma^{\frac{1}{2}}\right)}}}{1+e^{Z(P(Y^*=1|X^*))+1.96*\mathrm{r}*\sqrt{\left(\Sigma^{\frac{1}{2}}\right)'\left(X^*\right)^2\left(\Sigma^{\frac{1}{2}}\right)}}}\right)$$

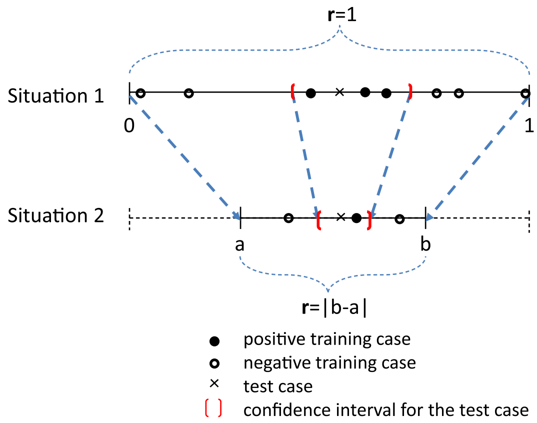
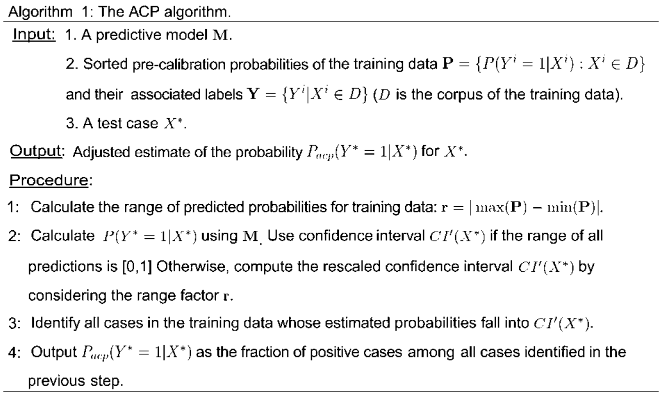


Figure 2 Illustration of the adaptive calibration of predictions (ACP) procedure. In situation 1, pre-calibration probabilities of training data range from zero to one, and equation (2) can be applied directly. In situation 2, pre-calibration probabilities of training data cover a much narrower interval (*a, b*). In this situation, utilizing the original CI would be problematic, as most observations might be within this range. To avoid that, we rescale the CI of the test case by considering the range factor r = |*b*−*a*|.

Figure 2 illustrates two situations in applying the ACP procedure.

In summary, we used four steps to convert a pre-calibration probability estimation $P(Y^* = 1|X^*)$ into a locally adjusted $P_{acp}(Y^* = 1|X^*)$ through adaptive calibration, as indicated in algorithm 1.

Figure 3 shows an example of applying ACP. The green lines in the first column of figures represent the convex hull (or set) of cases whose estimated probabilities fall into the CI of the test case X^* (indicated by a black star). Note that r equals |0.43−0.11| = 0.32 in all three cases, as it depends on predictions for the training data.



```
Algorithm 1: The ACP algorithm.
Input: 1. A predictive model $\mathbf{M}$.
       2. Sorted pre-calibration probabilities of the training data $\mathbf{P} = \{P(Y^i = 1|X^i) : X^i \in D\}$
       and their associated labels $\mathbf{Y} = \{Y^i|X^i \in D\}$ ($D$ is the corpus of the training data).
       3. A test case $X^*$.
Output: Adjusted estimate of the probability $P_{acp}(Y^* = 1|X^*)$ for $X^*$.
Procedure:
1: Calculate the range of predicted probabilities for training data: $\mathbf{r} = |\max(\mathbf{P}) - \min(\mathbf{P})|$.
2: Calculate $P(Y^* = 1|X^*)$ using $\mathbf{M}$. Use confidence interval $CI'(X^*)$ if the range of all
   predictions is [0,1] Otherwise, compute the rescaled confidence interval $CI'(X^*)$ by
   considering the range factor $\mathbf{r}$.
3: Identify all cases in the training data whose estimated probabilities fall into $CI'(X^*)$.
4: Output $P_{acp}(Y^* = 1|X^*)$ as the fraction of positive cases among all cases identified in the
   previous step.
```

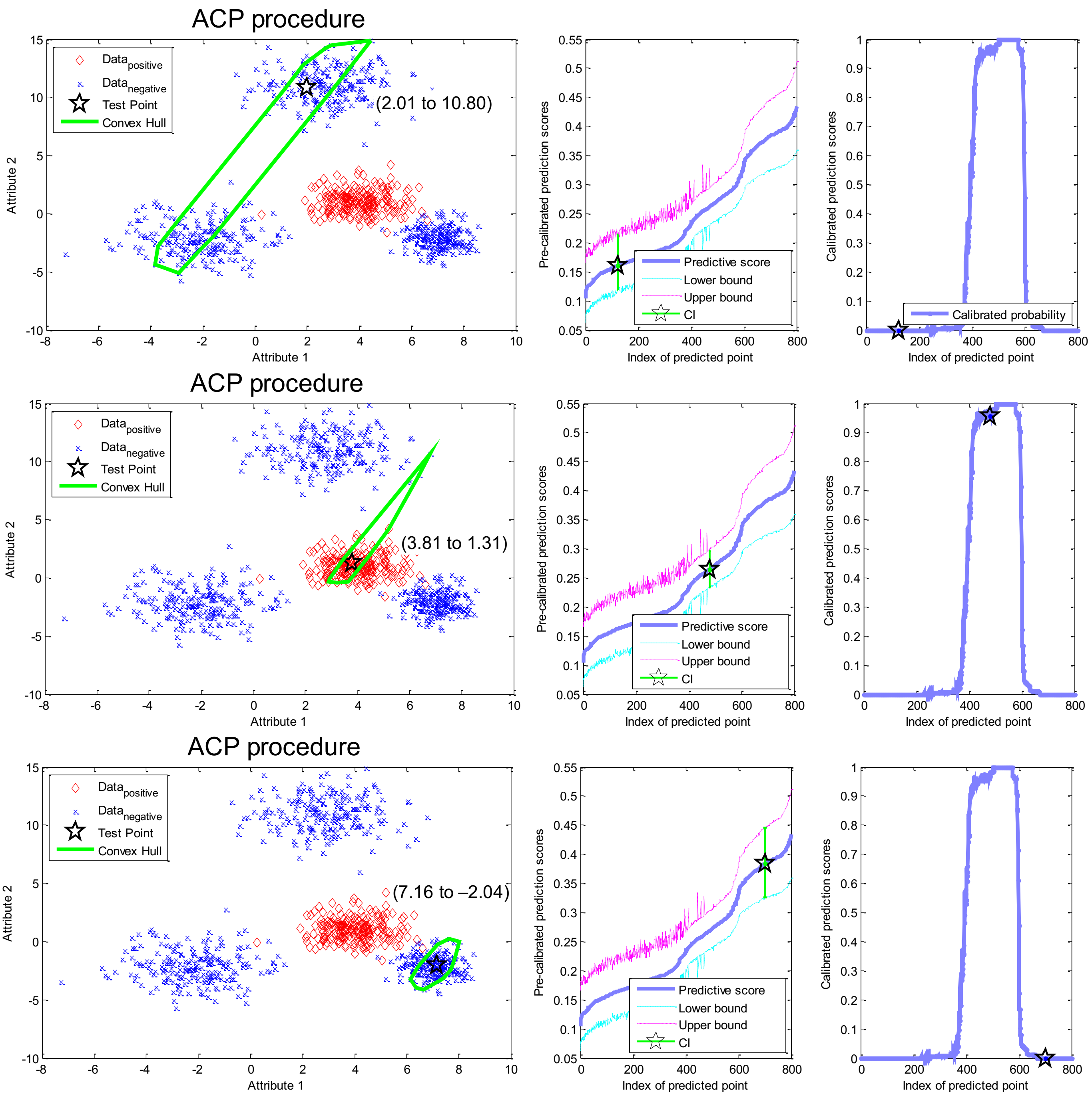


Figure 3 Examples of applying the adaptive calibration of predictions (ACP) method to test cases. We sampled 800 cases from four Gaussian distributions (600 negative and 200 positive cases) to create the training data. In the first column, each figure illustrates a test case and its convex hull (ie, the set that was used to calibrate the prediction). In the second column, we show the sorted probabilities and CIs for training cases, as well as the estimated probability and the CI for the test case. Finally, in the last column, we show adjusted probability estimates of all training cases and the test case after application of ACP. For comparison, we kept the order of cases on the x-axis of the figure in column three consistent with the orders in column two.

The probability of each test case was adjusted using training cases with similar probabilities. As a result, the ACP model was capable of handling non-linearly separable cases for which the original LR model failed. Therefore, although it used LR as a guide to order these cases by their probability estimates, and to generate a neighborhood using the LR CI around the test case, it was capable of calibrating predictions in regions where class labels of cases in these neighborhoods were very heterogeneous. Hence, in this particular case, it was able to 'fix' the predictions of an LR model. We certainly would not recommend LR usage in non-linearly separable problems (ie, without including interaction terms), and hence this example was used just to illustrate how ACP works, and how it can dramatically change predictions in certain cases, *but not to advocate for its use to remediate a model that does not fit the data in the first place*. Figure 4 illustrates the separation boundaries of the LR and ACP methods in a simple two-dimensional space, respectively.

Assuming the sizes of the training and test data are n and m, ACP needs $O(n \log n)$ to sort estimated probabilities, and uses a hash function to find the neighbors for each calibration at

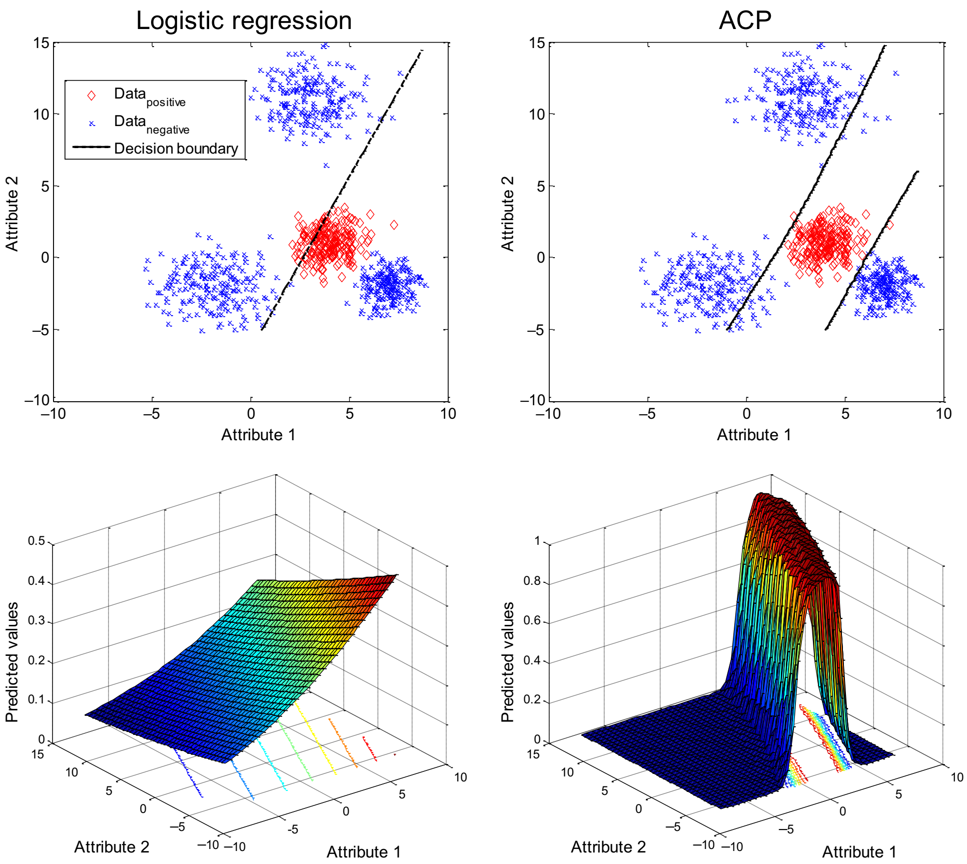


Figure 4 Visual comparison of adaptive calibration of predictions (ACP) and logistic regression (LR) models using a simulated 2D dataset. In the first row, blue crosses correspond to negative cases and red diamonds correspond to positive cases. The black lines indicate the decision boundaries of LR and ACP models at their cut-offs. In the second row, the surface plots illustrate the distribution of estimated probabilities for both models.

a cost of $O(1)$. For methods in comparison, *binning* requires $O(n \log n)$ to construct K bins, and an additional $O(1)$ for each calibration. *Platt scaling* requires $O(nT)$ to build a one-dimensional LR model using Newton's method,[25] where T is the number of iterations required for convergence. The calibration for each subject costs $O(1)$ for Platt scaling. Finally, IR requires $O(n)$ to build the step functions[20] and an additional $O(1)$ for each calibration. Table 1 summarizes the time complexity of these calibration methods.

RESULTS

We evaluated the performance of different calibration methods using both synthetic and real medical data. For comparison, we used three indices, the AUC,[14] the decile-based Hosmer–Lemeshow goodness-of-fit test (HL test),[26 27] and root mean squared error (MSE). These first two are measurements of discrimination (AUC) and goodness-of-fit for the LR (HL test), respectively and the latter (MSE) is related to analysis of residuals. See Lasko *et al*[28] and Zou *et al*[29] for a review of AUC, and Hosmer and Lemeshow[24] for the HL test.

To visualize estimated probabilities before and after applying calibration methods, we used a reliability diagram, which is produced using the following steps. First, we sort pre-calibrated predictions in ascending order. Next, these predictions along with their respective class labels are grouped into 10 bins. Like subgroups in the HL test, we have two choices for constructing these bins: (1) equal number of elements in the bins, sorted by probability estimates, that is, $\lfloor n/10 \rfloor$ elements per cell; or (2) fixed, equal length intervals of probability estimates, that is, $0<p<=0.1$, $0.1<p<=0.2$, etc. In this article, the first option was used to be consistent with the decile-based HL test. Finally, we plot average predictions versus average class labels (ie, proportion of positive cases) within each bin. The closer the plotted points are to the diagonal line, the better the calibration. We compared ACP with LR without calibration (LR), LR with binning (binning), LR with Platt scaling (PS), and LR with IR (IR) using both artificial and medically relevant data.

Table 1 Time complexity of calibration methods for training a calibration model with n cases and using it to predict all of m test cases

	Binning	Platt scaling	Isotonic regression	ACP
Time complexity	$O(n \log n+m)$	$O(nT+m)$	$O(n+m)$	$O(n \log n+m)$

Note that n and m stand for the size of training and test data, and T is the number of iterations required for convergence.
ACP, adaptive calibration of predictions.

Table 2 Performance measures over 1000 runs

	AUC (mean±SD)	MSE (mean±SD)	HL test (pass rate)
(a) Dataset A: almost linearly separable data			
LR	0.983±0.004	0.048±0.006	91.1%
Binning	0.980±0.004	0.052±0.005	98.6%
PS	0.983±0.004	0.051±0.007	43.3%
IR	0.985±0.004	0.045±0.006	98.8%
ACP	0.985±0.004	0.045±0.006	99.9%
(b) Dataset B: not linearly separable data			
LR	0.502±0.002	0.222±3e-5	0%
Binning	0.954±0.006	0.074±0.005	779%
PS	0.502±0.002	0.222±2e−5	0%
IR	0.718±0.005	0.180±0.002	0%
ACP	0.967±0.005	0.065±0.006	997%

ACP, adaptive calibration of predictions; AUC, area under the ROC (receiver operating characteristic) curve; HL, Hosmer–Lemeshow; IR, isotonic regression; LR, logistic regression; MSE, mean squared error; PS, Platt scaling.

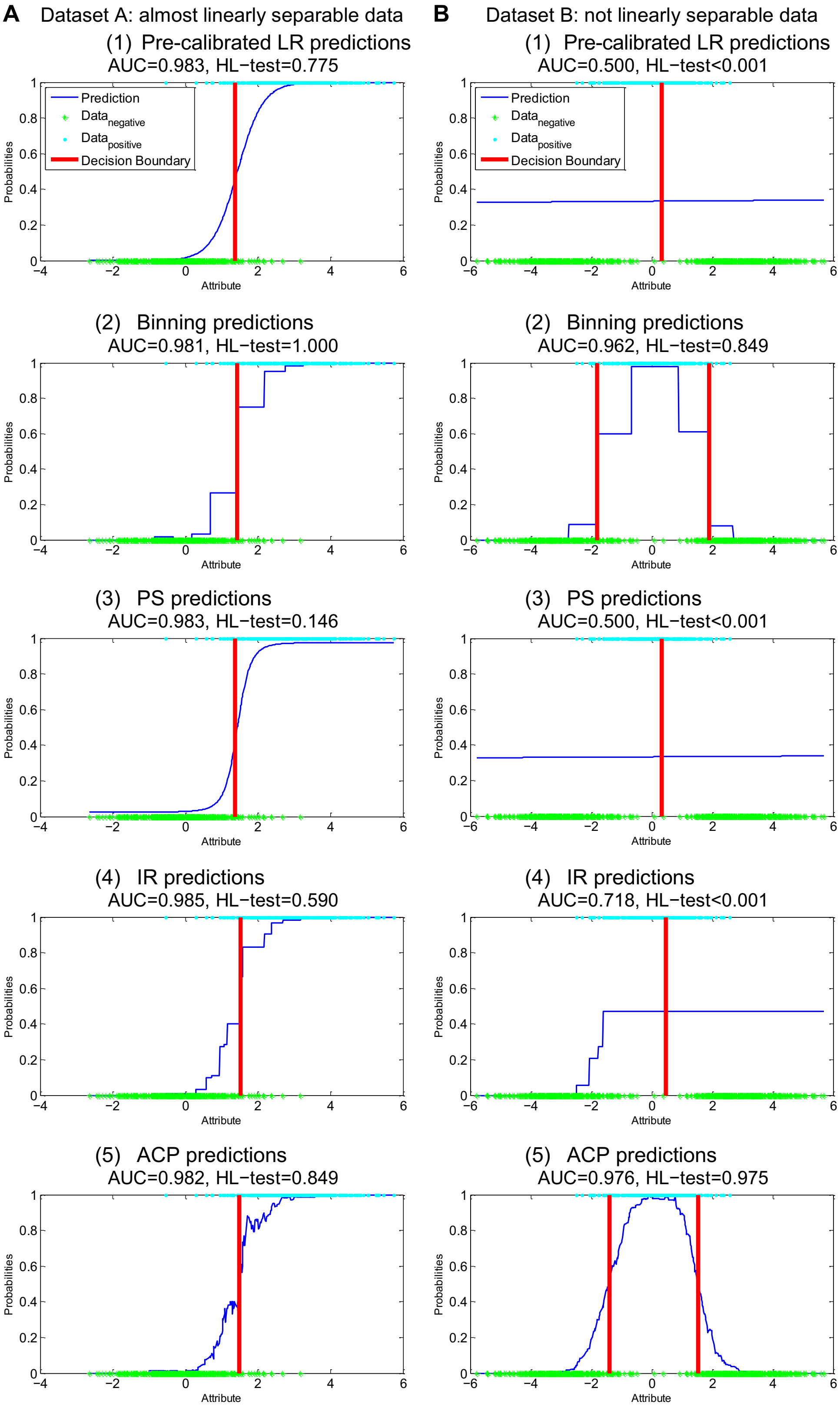


Figure 5 Visualization of probabilities generated by logistic regression (LR) and four different calibration methods using synthetic datasets. ACP, adaptive calibration of predictions; AUC, area under the ROC (receiver operating characteristic) curve; HL, Hosmer—Lemeshow goodness-of-fit test; IR, isotonic regression; PS, Platt scaling.

Synthetic data

As an illustration, we sampled one-dimensional data so that the probabilistic outputs of different approaches (LR, binning, PS, IR, and ACP) could be visualized. Our first simulated dataset is almost linearly separable, and sampled from two Gaussian distributions with unit variance but different means, $X_0 \in N(0,1)$, $X_1 \in N(3,1)$, and $X = X_1 \cup X_0$, where X_1 and X_0 correspond to data with class label '1' and '0'. The second dataset is not linearly separable, and was generated from $X_0 \in N(-3,1) \cup N(3,1)$, $X_1 \in N(0,1)$. Table 2 shows the results of applying different approaches on 1000 runs of simulated data. All approaches had comparable (ie, no significant difference) AUC and MSE for the almost linearly separable data, but ACP demonstrated better calibration, which 'passed' the HL test

Table 3 Summary of features and the target variable for the hospital discharge error data: eight features (potential predictors) are categorical and two are numerical

Name	Details
Potential predictors	
Specimen	0=blood, 1=urine, 2=sputum, 3=CSF
Specific days	Number of days between admission date and specimen collection date
Collected week	0=specimen collected on weekday, 1=specimen collected on weekend
Final week	0=final result on weekday, 1=final result on weekend
Visit type	1=admission, 0=non-admission
Service	0=<blank> (patient not admitted), 1=oncology, 2=general medicine, 3=medical subspecialties, 4=surgery and surgical sub-specialties, 5=other
Age	Age in years
Gender	0=male, 1=female
Race	0=white, 1=black, 2=Asian, 3=Hispanic, 4=other, 5=unknown/declined
Insurance	0=Medicare, 1=Medicaid, 2=commercial, 3=other
Outcome variable	
Potential error	0=not a potential follow-up error, 1=a potential follow-up error

in 999 out of the 1000 runs (ie, the p value was greater than 0.05).

Regarding the second dataset that is not linearly separable, ACP stood out in all three indices, and had a statistically significant higher AUC than LR ($p<1e-15$), binning ($p<1e-15$), PS ($p<1e-15$), and IR ($p<1e-15$) using a right-tailed paired t test. ACP also had a lower MSE compared to LR ($p<1e-10$), binning ($p<1e-10$), PS ($p<1e-10$), and IR ($p<1e-10$) using a left-tailed paired t test. Finally, ACP showed a higher rate of 'passing' the HL test than LR ($p<1e-10$), binning ($p<1e-10$), PS ($p<1e-10$), and IR ($p<1e-10$). All methods except ACP and binning performed poorly on dataset B. The reason is that their monotonic constraints limit their transformation power. Therefore, if pre-calibrated predictions have low AUCs, their 'calibrated' outcomes using PS and IR remain poorly discriminative. In other words, these outcomes still make many mistakes in ranking, which implicitly lead to large MSE and poor calibration (table 2).

Figure 5A (1–5) illustrates results from a single simulation of the almost linearly separable data with all five approaches. These methods showed similar discrimination ability, and all of them accepted the null hypothesis that the model was calibrated using the HL test at significance level 0.05. Similarly, figure 5B (1–5) shows the results from a single simulated run on data that are not linearly separable. LR and PS had AUCs around 0.5, which are close to the performance of a random classifier. Both approaches also failed to pass the HL test. IR demonstrated better AUC at 0.718, but its estimations of probabilities were not calibrated. In contrast, ACP showed superior performance in both discrimination and calibration (AUC=0.976, HL test p=0.975), which slightly outperformed the second best approach, binning (AUC=0.962, HL test p=0.849).

Experiments with clinical data

We also conducted experiments using clinical data. As opposed to the synthetic data, the gold standard individualized probability is unknown here, but results that use the same evaluation measures as in the synthetic data suggest that our approach may have advantages over binning, Platt scaling, and IR.

Hospital discharge data

This experiment was conducted on a de-identified dataset used for predicting potential follow-up errors related to microbiology cultures ordered while patients were hospitalized, for which a predictive model was previously published.[30] These errors include, among others, continued prescription of antibiotics that do not cover the microorganisms identified in cultures. Identifying the cases most likely to have inappropriate follow-up can help providers be on alert for these potential errors.

The data represented a retrospective analysis of microbiology cultures performed at a teaching hospital in 2007. The potential predictors consist of eight categorical variables and two numerical variables, which are shown in table 3. The outcome was a binary variable indicating a potential follow-up error.

Figure 6 illustrates the distribution of each feature variable (predictor) and the target variable. From a total of 8668 hospital discharge cases, 385 were considered to be potential errors in a review of charts executed by trained professionals. The dataset is highly imbalanced: non-errors dominate the observations.

In the modeling process, we represented each categorical feature by a set of binary variables. For example, the categorical feature Specimen corresponds to three Boolean variables (indicating *urine*, *sputum*, or *CSF*, with the baseline being *blood*). The fully expanded feature space had 22 dimensions. Similarly to our synthetic experiment, we applied different calibration models and compared their performance. We randomly split the data into training (66%) and test sets (34%) for evaluation. Figure 7 shows that AUCs for all five methods were comparable at around 0.742. Regarding calibration, the HL test results showed that binning (p=0.019), PS (p=0.005), and IR (p<0.001) did not generate well-calibrated outputs, while ACP and LR generated calibrated probabilities (ie, LR: p=0.096, and ACP: p=0.615).

We repeated the random split process 100 times and applied all calibration approaches. Their results are listed in table 4. All five methods had comparable AUCs around 0.71. ACP showed slightly lower MSE than other approaches, but this was not statistically

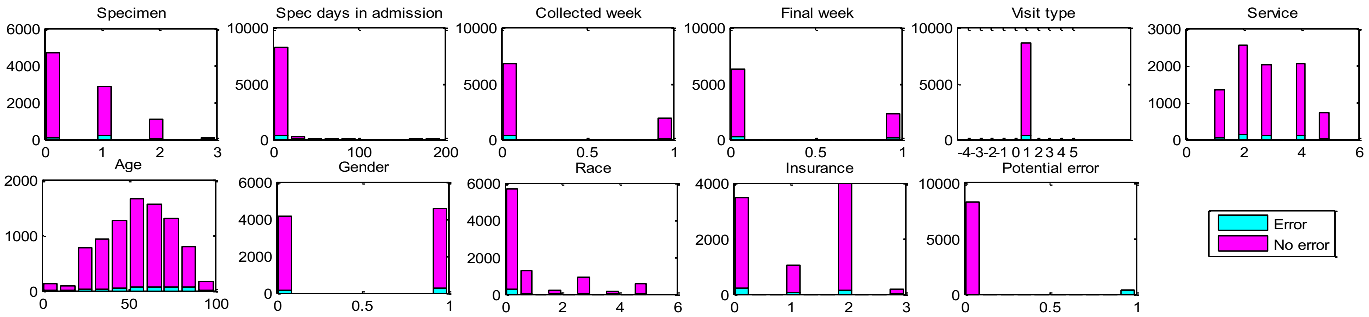


Figure 6 Histograms of feature and target variable values for the hospital discharge data. Blue bars indicate potential follow-up errors and red bars represent normal cases. There are eight categorical variables and two numerical variables.

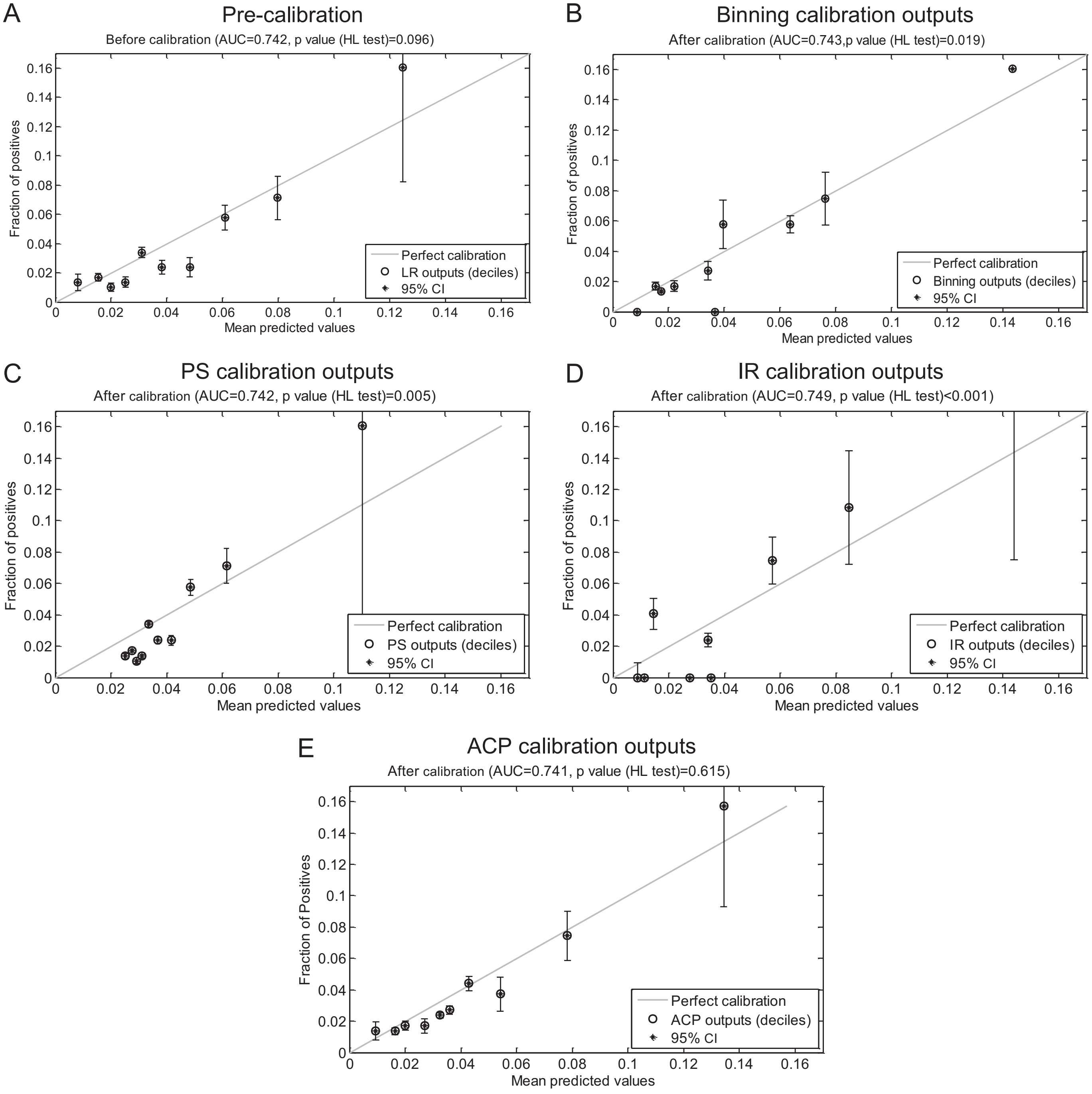


Figure 7 Various calibration methods are applied to the hospital discharge data. In the first sub-figure, the black dots indicate the averaged probabilities of an LR model, plotted against their corresponding fraction of positive cases in 10 equal-element cells. In the rest of the sub-figures, the graphs show reliability diagrams after application of different calibration methods. ACP, adaptive calibration of predictions; AUC, area under the ROC (receiver operating characteristic) curve; HL, Hosmer–Lemeshow; IR, isotonic regression; LR, logistic regression; PS, Platt scaling.

significant (ie, LR: p=0.56; binning: p=0.46; PS: p=0.056; and IR: p=0.45). Regarding calibration, ACP did not reject the null hypothesis that estimated probabilities are calibrated 65 out of the 100 times using HL tests, followed by PS (48), LR (46), binning (4), and IR (0). The p values of the HL test given by the ACP method are significantly higher than the p values of PS (p=0.01), binning (p<0.01), PS (p=0.04), and IR (p<0.01).

The binning approach had very poor calibration performance on this data. A major reason is that the binning approach merges thousands of pre-calibrated prediction values into only 10 values. Although its calibration on the training data is perfect,

Table 4 Performance of ACP and other calibration methods over 100 random splits

	AUC (mean±SD)	HL test (pass rate)	MSE (mean±SD)	Time (seconds)
LR	0.71±0.019	46%	0.041±0.0026	*5.75±0.318*
Binning	0.70±0.020	4%	0.041±0.0025	*1.24±0.066*
PS	0.71±0.019	48%	0.042±0.0027	*2.97±0.318*
IR	0.71±0.018	0%	0.041±0.0025	*1.33±0.069*
ACP	0.71±0.019	65%	0.040±0.0025	*2.28±0.121*

ACP, adaptive calibration of predictions; AUC, area under the ROC (receiver operating characteristic) curve; HL, Hosmer–Lemeshow; IR, isotonic regression; LR, logistic regression; MSE, mean squared error; PS, Platt scaling.

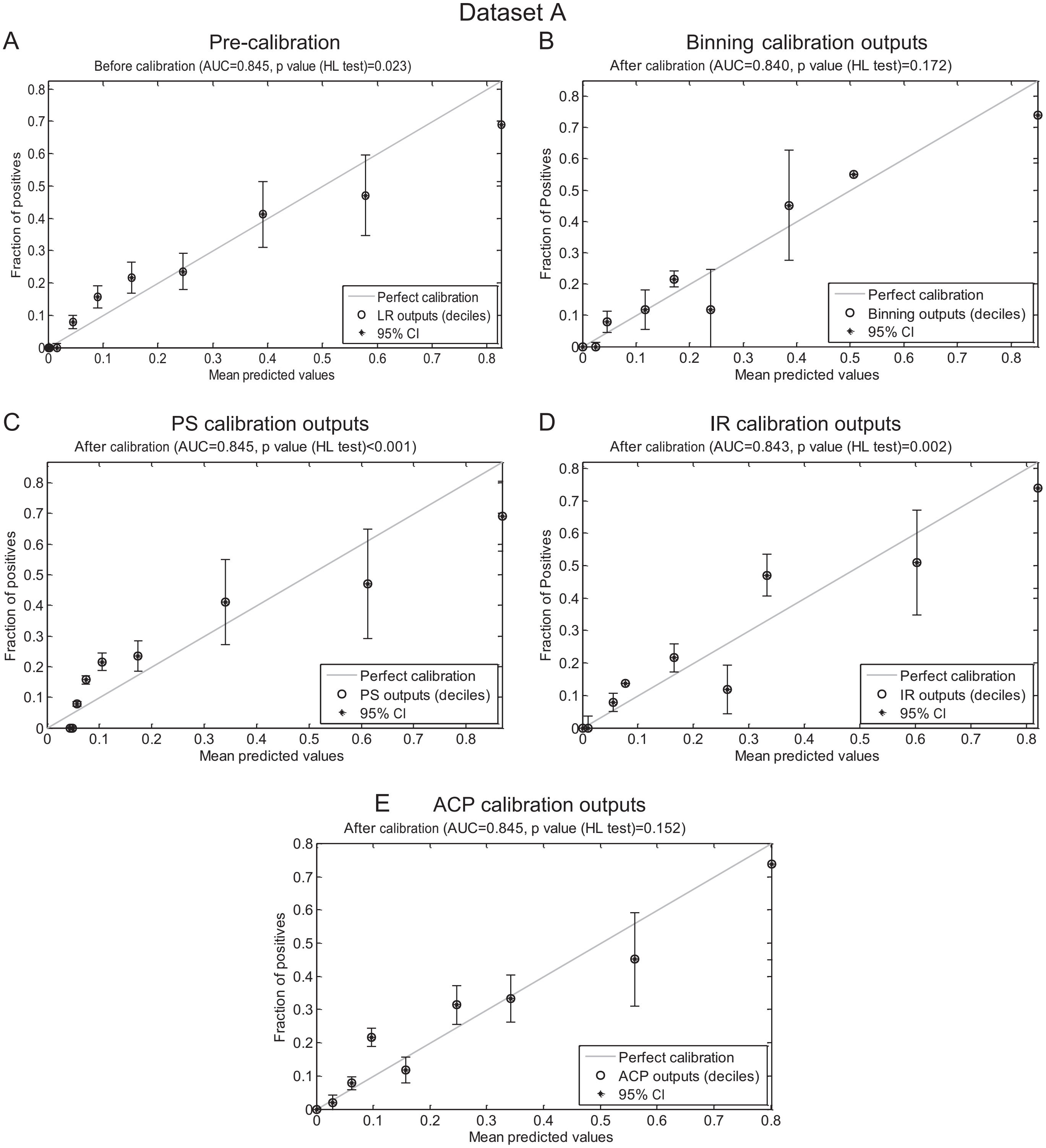


Figure 8 Comparison of various calibration approaches using dataset A. The ACP and binning methods passed the HL test while other approaches did not generate calibrated outputs. ACP, adaptive calibration of predictions; AUC, area under the ROC (receiver operating characteristic) curve; HL, Hosmer–Lemeshow; IR, isotonic regression; LR, logistic regression; PS, Platt scaling.

performance on the test data could be very poor due to overfitting. In extreme cases, some bins would host only negative test examples but no positive examples at all, which causes a large deviation.

Myocardial infarction data

The myocardial infarction (MI) datasets contain information from patients with and without MI, who were seen at two emergency departments in the UK.[31] These data were originally collected to determine which and how many data items were required to construct a decision support algorithm for early diagnosis of acute MI, using clinical and electrocardiograph data available at presentation. Variables such as nausea, chest pain characteristics, EKG and physical exam findings, demographics, and past history of MI were used to predict current MI. Although outdated, these data are representative of the types

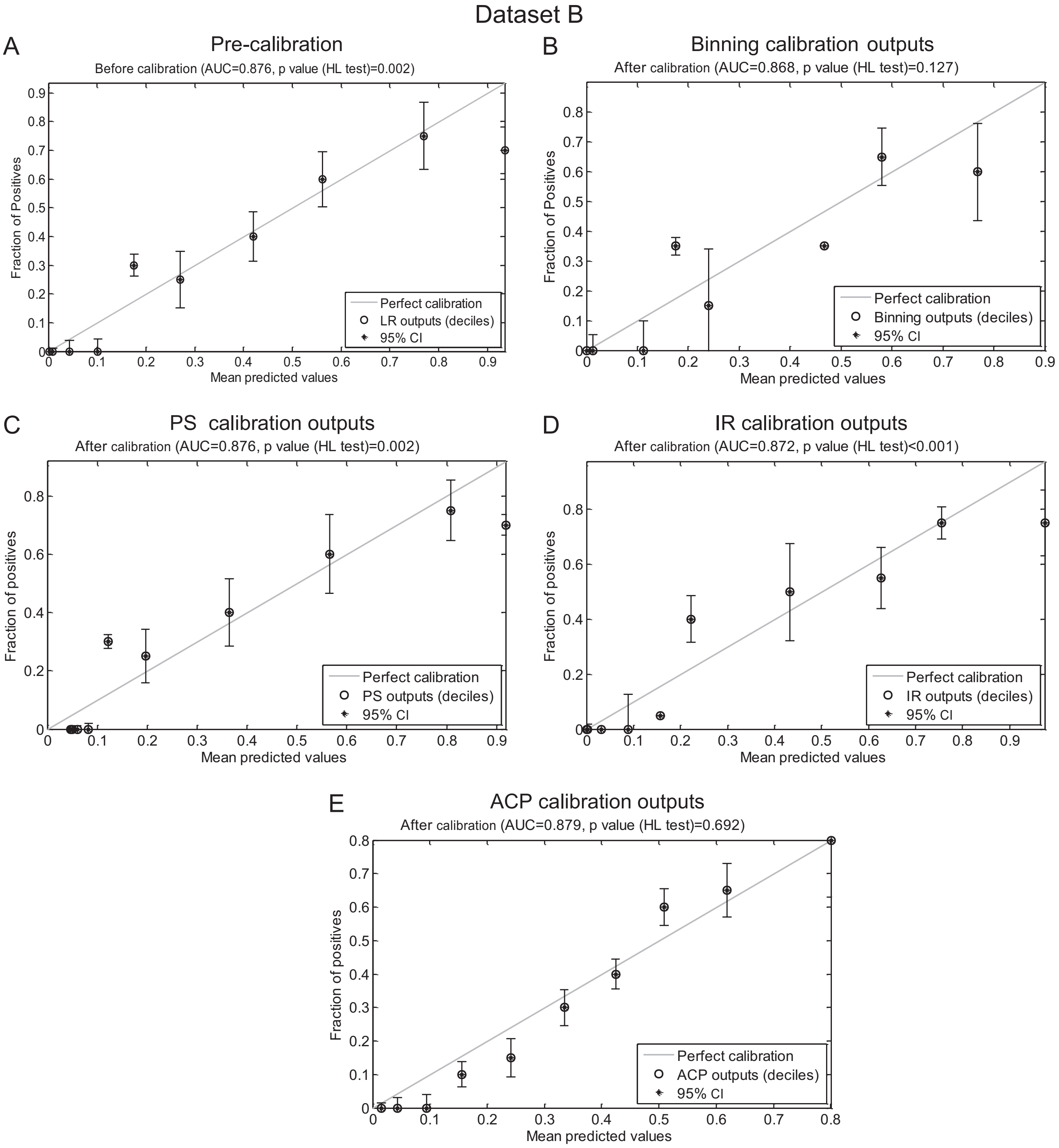


Figure 9 Comparison of various calibration methods using dataset B. PS and IR failed to pass the HL test, and their outputs visually deviated further away from the perfect calibration line (gray) compared to the ACP method. The ACP method generated calibrated predictions and had the largest AUC among all approaches. ACP, adaptive calibration of predictions; AUC, area under the ROC (receiver operating characteristic) curve; HL, Hosmer–Lemeshow; IR, isotonic regression; LR, logistic regression; PS, Platt scaling.

of problems being addressed by predictive models and are used here for illustration purposes. We used a random split to divide dataset A (ie, patients observed in emergency departments in Edinburgh) into a training (60%) and a test (40%) set. Similarly, dataset B (ie, patients observed in emergency departments in Sheffield) was divided into a training (60%) and a test (40%) set.

Figures 8 and 9 illustrate the use of various calibration methods on the test data, which were randomly split from the datasets A and B. In both experiments, PS and IR failed to pass the HL test at the significance level of 0.05. On the other hand, binning passed the HL test, but its AUCs were lower than the other methods. For both cases, ACP showed superior calibration without decreasing AUCs.

Table 5 Performance of ACP and other calibration methods using 100 random splits of datasets A and B

	A				B			
	AUC (mean±SD)	HL test (pass rate)	MSE (mean±SD)	Run time (mean±SD)	AUC (mean±SD)	HL test (pass rate)	MSE (mean±SD)	Run time (mean±SD)
LR	0.851±0.017	15%	0.117±0.008	*0.379±0.040*	0.850±0.025	6%	0.151±0.018	*0.247±0.047*
Binning	0.846±0.017	29%	0.120±0.008	*1.014±0.067*	0.839±0.028	8%	0.155±0.019	*1.041±0.116*
PS	0.852±0.017	11%	0.120±0.009	*1.147±0.040*	0.850±0.025	9%	0.156±0.019	*1.115±0.047*
IR	0.848±0.017	12%	0.119±0.008	*1.030±0.057*	0.845±0.026	3%	0.156±0.019	*1.009±0.065*
ACP	0.858±0.017	46%	0.116±0.007	*0.980±0.056*	0.850±0.026	38%	0.145±0.014	*0.927±0.062*

Note that the run time of both experiments is measured in seconds.
ACP, adaptive calibration of predictions; AUC, area under the ROC (receiver operating characteristic) curve; HL, Hosmer–Lemeshow; IR, isotonic regression; LR, logistic regression; MSE, mean squared error; PS, Platt scaling.

We repeated the random split process 100 times and applied all four calibration approaches on both datasets. The results are listed in table 5. For dataset A, ACP showed the largest AUCs, which were significantly higher than the AUCs of binning ($p<3.6e-17$) and IR ($p<2.2e-11$) using a right-tailed paired t test. ACP also had lower MSEs compared to LR ($p=0.004$), binning ($p<1e-5$), PS ($p<1e-5$), and IR ($p=1e-5$) using a left-tailed paired t test. ACP also had the highest rate of passing the HL test compared to the other methods, and its p values were significantly higher than LR ($p<1e-5$), binning ($p<1e-5$), PS ($p<1e-5$), and IR ($p<1e-5$). For dataset B, the AUCs of ACP were higher than the AUCs of binning ($p<1.5e-13$) and IR ($p<1.5e-9$) and comparable to LS ($p=0.56$) and PS ($p=0.56$). ACP also showed significantly lower MSE compared to LR ($p=0.007$), binning ($p=0.0003$), PS ($p<1e-3$), and IR ($p<1e-3$). Regarding calibration, ACP demonstrated a higher rate of passing the HL tests compared to other methods, and its p values were significantly higher than those of LR ($p<1e-3$), binning ($p=0.0005$), PS ($p<1e-3$), and IR ($p<1e-5$).

DISCUSSION

Calibration is a less studied but important aspect of a predictive model, particularly when estimates are used for personalized medicine. If uncalibrated predictions are used as surrogates of risk estimations, the medical decisions for individual patients could be incorrect. While previous efforts to calibrate probabilistic estimates provide global solutions, we proposed a novel approach that uses tailored information to calibrate adaptively and locally. Without increasing computational complexity, our approach demonstrated good performance in experiments using synthetic and clinical data. Furthermore, our framework can be extended to any probabilistic models that generate CIs associated with each prediction.

The ACP procedure is simple and straightforward. However, it is not always easy to determine the CIs of predictions for predictive models. For example, there is not a closed form solution for the CIs of SVM predictions, which is not a method originally designed to produce probabilistic estimates. Although it is always possible to estimate the CIs by bootstrapping,[32] the calculation becomes computationally more expensive.

LIMITATION

ACP is not parameter-free. It needs a threshold parameter for truncating the CIs, which is set to be 95% in our experiments as mentioned earlier in the Methods section. A larger value for this threshold parameter would allow ACP to have more local points included in the computation, and vice versa. We can systematically stretch or compress the 'neighborhood of interest' for our calibration model according to the threshold. Even though the model is not completely parameter free, it provides more flexibility in adjusting estimated probabilities when compared to calibration methods like binning, Platt scaling, and IR. More research in investigating the optimal threshold value for the CIs is certainly warranted. Furthermore, the examples presented here are relatively small and do not represent the full spectrum of predictive models that are increasingly being used in clinical practice and biomedical research. The connection to personalized medicine is, like personalized medicine itself, still tentative. However, the field is not likely to evolve unless calibration issues are resolved. While we believe that ACP provides a contribution in that direction, clearly much more research and extensive studies are needed.

Acknowledgments We thank Drs El-Kareh and Fraser for providing the clinical datasets used in this article. We thank two anonymous reviewers for insightful suggestions, and Dr Michele Day for editing this manuscript.

Funding XJ, JK, and LOM were funded in part by the National Library of Medicine (R01LM009520) and NHLBI (U54 HL10846), NIH.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement We will share the simulation data when the paper is accepted.

REFERENCES

1. **Petracci E,** Decarli A, Schairer C, *et al*. Risk factor modification and projections of absolute breast cancer risk. *J Natl Cancer Inst* 2011;**103**:1037–48.
2. **Zimmerman JE,** Kramer AA, McNair DS, *et al*. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med* 2006;**34**:1297–310.
3. **Karp I,** Abrahamowicz M, Bartlett G, *et al*. Updated risk factor values and the ability of the multivariable risk score to predict coronary heart disease. *Am J Epidemiol* 2004;**160**:707–16.
4. **Wei W,** Visweswaran S, Cooper GF. The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *J Am Med Inform Assoc* 2011;**18**:370–75.
5. **Ohno-Machado L,** Resnic FS, Matheny ME. Prognosis in critical care. *Annu Rev Biomed Eng* 2006;**8**:567–99.
6. **Stone NJ,** Bilek S, Rosenbaum S. Recent national cholesterol education program adult treatment panel III update: adjustments and options. *Am J Cardiol* 2005;**96**:53–9.
7. **Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults.** Executive summary of the third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). *JAMA* 2001;**285**:2486–97.
8. **Hooks MA.** Breast cancer: risk assessment and prevention. *South Med J* 2010;**103**:333–8.
9. **Chang RW,** Jacobs S, Lee B. Use of APACHE II severity of disease classification to identify intensive-care-unit patients who would not benefit from total parenteral nutrition. *Lancet* 1986;**1**:1483–7.
10. **Kullo IJ,** Cooper LT. Early identification of cardiovascular risk using genomics and proteomics. *Nat Rev Cardiol* 2010;**7**:309–17.
11. **Altman RB,** Miller KS. 2010 Translational bioinformatics year in review. *J Am Med Inform Assoc* 2010;**18**:358–66.
12. **Sarkar IN,** Butte AJ, Lussier YA, *et al*. Translational bioinformatics: linking knowledge across biological and clinical realms. *J Am Med Inform Assoc* 2011;**18**:354–7.

13. **Osl M,** Ohno-Machado L, Baumgartner C, *et al*. Improving calibration of logistic regression models by local estimates. *AMIA Annu Symp Proc* 2008:535—9.
14. **Hanley J,** McNeil B. The meaning and use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* 1982;**143**:29—36.
15. **Ayer T,** Alagoz O, Chhatwal J, *et al*. Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. *Cancer* 2010;**116**:3310—21.
16. **Domingos P,** Pazzani M. Beyond independence: Conditions for the optimality of the simple Bayesian classifier. *Mach Learn* 1997;**29**:103—30.
17. **Zadrozny B,** Elkan C. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. *The Eighteenth International Conference on Machine Learning*. 2001:609—16.
18. **Platt JC.** Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers* 1999:61—74.
19. **Zadrozny B,** Elkan C. Transforming classifier scores into accurate multiclass probability estimates. *The Eighth International Conference on Knowledge Discovery and Data Mining*. 2002:694—99.
20. **Burdakov O,** Grimvall A, Hussian M. A generalised pav algorithm for monotonic regression in several variables. *International Conference on Computational Statistics*. 2004:761—67.
21. **Wang X,** Li F. Isotonic smoothing spline regression. *J Comput Graph Stat* 2008;**17**:21—37.
22. **Altman NS.** An introduction to Kernel and nearest-neighbor nonparametric regression. *Am Stat* 1992;**46**:175—85.
23. **Oehlert G.** A note on the delta method. *Am Stat* 1992;**46**:27—9.
24. **Hosmer DW,** Lemeshow S. *Applied Logistic Regression*. New York: Wiley-Interscience, 2000.
25. **Minka T.** *A comparison of numerical optimizers for logistic regression*. Pittsburgh, PA: Carnegie Mellon University, Technical Report, 2003.
26. **Hosmer DW,** Hosmer T, Le Cessie S, *et al*. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 1997;**16**:965—80.
27. **Kramer AA,** Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit Care Med* 2007;**35**:2052—6.
28. **Lasko TA,** Bhagwat JG, Zou KH, *et al*. The use of receiver operating characteristic curves in biomedical informatics. *J of Biomed Inform* 2005;**38**:404—15.
29. **Zou K,** Liu A, Bandos A, *et al*. *Statistical evaluation of diagnostic performance: topics in ROC analysis*. Boca Raton, FL: Chapman & Hall/CRC Biostatistics Series, 2011.
30. **El-Kareh R,** Roy C, Brodsky G, *et al*. Incidence and predictors of microbiology results returning post-discharge and requiring follow-up. *J Hosp Med* 2011;**6**:291—6.
31. **Kennedy RL,** Burton AM, Fraser HS, *et al*. Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models. *Eur Heart J* 1996;**17**:1181—91.
32. **Efron B.** Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci* 1986;**1**:54—75.