

Predicting dichotomised outcomes from high-dimensional data in biomedicine

Armin Rauschenberger & Enrico Glaab

To cite this article: Armin Rauschenberger & Enrico Glaab (2023): Predicting dichotomised outcomes from high-dimensional data in biomedicine, Journal of Applied Statistics, DOI: [10.1080/02664763.2023.2233057](https://doi.org/10.1080/02664763.2023.2233057)

To link to this article: <https://doi.org/10.1080/02664763.2023.2233057>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 26 Jul 2023.



[Submit your article to this journal](#)



Article views: 24



[View related articles](#)



[View Crossmark data](#)

Predicting dichotomised outcomes from high-dimensional data in biomedicine

Armin Rauschenberger  and Enrico Glaab 

Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg

ABSTRACT

In many biomedical applications, we are more interested in the predicted probability that a numerical outcome is above a threshold than in the predicted value of the outcome. For example, it might be known that antibody levels above a certain threshold provide immunity against a disease, or a threshold for a disease severity score might reflect conversion from the presymptomatic to the symptomatic disease stage. Accordingly, biomedical researchers often convert numerical to binary outcomes (loss of information) to conduct logistic regression (probabilistic interpretation). We address this bad statistical practice by modelling the binary outcome with logistic regression, modelling the numerical outcome with linear regression, transforming the predicted values from linear regression to predicted probabilities, and combining the predicted probabilities from logistic and linear regression. Analysing high-dimensional simulated and experimental data, namely clinical data for predicting cognitive impairment, we obtain significantly improved predictions of dichotomised outcomes. Thus, the proposed approach effectively combines binary with numerical outcomes to improve binary classification in high-dimensional settings. An implementation is available in the R package `cornet` on GitHub (<https://github.com/rauschenberger/cornet>) and CRAN (<https://CRAN.R-project.org/package=cornet>).

ARTICLE HISTORY

Received 23 August 2022

Accepted 28 June 2023

KEYWORDS

Linear/logistic regression; numerical prediction; binary classification; dichotomisation; high-dimensional data; ridge/lasso regularisation



MATHEMATICS SUBJECT CLASSIFICATIONS


62-04; 62J07; 62J05; 62H30; 62P10

1. Introduction

Many diagnostic and prognostic problems in biomedicine are essentially binary classification tasks. A binary outcome splits samples into two groups of interest. Some binary outcomes are *naturally* binary, whereas other binary outcomes are *artificially* binary [29]. We focus on artificial binary outcomes that result from the dichotomisation of numerical outcomes with a single threshold. Such binary variables indicate whether the underlying measurements are greater than a given cut-off value.

While there are strong reservations against outcome dichotomisation in the statistical literature [3,8,14,15,19,26], it remains popular in empirical research. The main problem

CONTACT Armin Rauschenberger  armin.rauschenberger@uni.lu  Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 7 avenue des Hauts Fourneaux, 4362 Esch-sur-Alzette, Luxembourg

 Supplemental data for this article can be accessed here: <https://doi.org/10.1080/02664763.2023.2233057>

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

for prediction is that dichotomising a numerical outcome implies a loss of information equivalent to discarding a certain proportion of the data [2], although it might simplify the understanding and communication of results [6,7] or increase robustness against contamination [25]. In our experience, researchers often underestimate the disadvantages or overestimate the advantages of dichotomisation.

However, many biomedical applications require predicted probabilities rather than predicted values. Suppose there is a critical transition if $y > c$, where y denotes a clinical outcome, and c denotes a threshold. Then we would want to predict $\mathbb{P}(y > c)$ rather than y . Typically, the prediction $y = c$ means that the probability of the critical transition is about 50%, but other predictions $y \neq c$ are more difficult to interpret, because they only tell whether the probability is below or above 50%. Even if the threshold is only an estimate of the tipping point where the critical transition occurs, we might want to predict the probability that the outcome will exceed this threshold, e.g. to make or to predict a treatment decision. (This also holds for *arbitrary* thresholds. Suppose a clinical protocol requires mechanical ventilation if the oxygen level falls below a certain value: Even if the patient could cope with lower values, we might want to predict whether the physician will use a ventilator.)

The analysis of modern biomedical data, typically including some hundred samples but many thousand features, requires new statistical methods. In this paper, we propose an approach to obtain improved predictions of dichotomised outcomes in high-dimensional settings (i.e. settings with many more features than samples).

The same problem has previously been addressed in low-dimensional settings [5,13,28] (i.e. settings with many fewer features than samples). Although the proposed approach is novel, we consider these and other related methods for possible extensions (see Section 5). There are methods that address different problems but also combine binary and numerical outcomes, such as risk estimation for dichotomised outcomes [27], bivariate regression for binary-continuous outcomes [4,9], odds ratios for linear regression [17], and ordinal logistic regression [12]. A recurrent idea is to exploit information from the numerical outcome and provide interpretation for the binary outcome.

This manuscript describes a straightforward approach to predict dichotomised outcomes from high-dimensional data. A more complex predictive method (e.g. random forests or neural networks for obtaining predicted values) together with a calibration method (e.g. Platt scaling or isotonic regression for transforming predicted values to predicted probabilities) could provide more predictive models, but these would be less interpretable ('black box'). We solve this specific prediction problem by combining two familiar methods (linear and logistic regression with lasso or ridge regularisation), leading to models that are not only predictive but also interpretable.

2. Approach

2.1. Overview

Our goal is to predict the (artificial) binary outcome, rather than the (natural) numerical outcome, from many features. For any sample, we either know or ignore both outcomes. Our strategy is to learn from the samples with observed outcomes how the features affect *both* outcomes, in order to predict the *binary* outcome of the samples with unobserved outcomes. A challenge in supervised learning (especially in high-dimensional settings) is

to avoid overfitting, which occurs if the model fits well to the observed data but not to unobserved data. This is how we model the two outcomes based on many features:

- two outcomes: In the generalised linear model framework, a suitable approach for binary outcomes is logistic regression, and a suitable approach for numerical outcomes is linear regression. Given both types of outcomes, we can fit both regression models. In most cases, linear regression is the better choice, because the numerical outcome is normally more informative than the binary outcome (see Examples 1 and 2 in Section 3.3). In some cases, however, logistic regression is the better choice, because it is more robust against departures from linearity and normality (see Examples 3 and 4 in Section 3.3). While logistic regression returns predicted probabilities, linear regression returns predicted values.
- many features: In low-dimensional settings without strong multicollinearity, we could estimate the regression coefficients by maximising the likelihood function. But in high-dimensional settings, which include many more features than samples, we need to regularise the regression coefficients. The lasso and ridge penalties, whose weighted sum is the elastic net penalty [32], increase with the absolute or squared values of the coefficients, respectively. Both penalties shrink the coefficients towards zero (regularisation), but only the lasso penalty sets coefficients equal to zero (variable selection).

We combine logistic and linear regression, with lasso or ridge regularisation, to predict dichotomised outcomes. This leads to two estimated effects for each feature, one from logistic regression and one from linear regression, and two predictions for each sample, one from logistic regression and one from linear regression. The proposed approach transforms the predicted values from linear regression to predicted probabilities and combines these predicted probabilities with those from logistic regression. Figure 1 illustrates the workflow.

2.2. Data

We observe one numerical outcome and p features for n samples. Let i in $\{1, \dots, n\}$ index the samples, and let j in $\{1, \dots, p\}$ index the features. For each i and j , let y_i denote

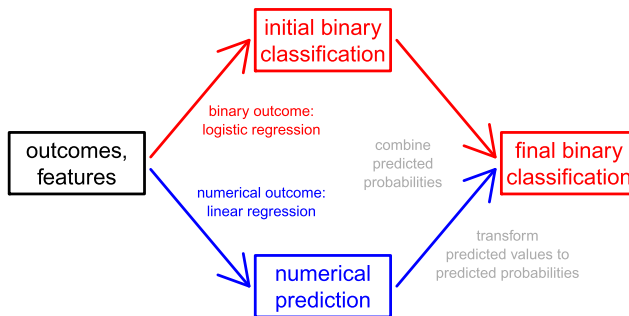


Figure 1. After modelling the (artificial) binary outcome with penalised logistic regression, and the (original) numerical outcome with penalised linear regression, we use the numerical prediction to improve the binary classification.

the outcome for sample i , and let x_{ij} denote feature j for sample i . Then the vector $\mathbf{y} = (y_1, \dots, y_n)^\top$ represents the outcome, and the $n \times p$ matrix \mathbf{X} represents the features. We focus on high-dimensional settings, where $p \gg n$. To prepare the data for penalised regression, we standardise all features (zero mean, unit variance).

Given a predefined threshold for dichotomising the numerical outcome, samples with an outcome above this threshold are in class 1, and all other samples are in class 0. For each sample i , let z_i indicate whether the numerical outcome y_i is greater than the threshold c , or formally $z_i = \mathbb{I}[y_i > c]$. Then the vector $\mathbf{z} = (z_1, \dots, z_n)^\top$ represents the binary outcome. Since the transformation of \mathbf{y} to \mathbf{z} is non-invertible, \mathbf{y} is at least as informative as \mathbf{z} (but typically much more informative).

2.3. Logistic regression

We relate the binary outcome to the features through logistic regression:

$$\text{logit}(\mathbb{P}[z_i = 1]) = \gamma_0 + \sum_{j=1}^p \gamma_j x_{ij},$$

where γ_0 is the unknown intercept, and $\{\gamma_1, \dots, \gamma_p\}$ are the unknown slopes. The latter represent the effects of the features on the log-odds of the binary outcome. Given the estimated coefficients $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_0, \dots, \hat{\gamma}_p)^\top$, the predicted probabilities are $\hat{\mathbf{z}} = (\hat{z}_1, \dots, \hat{z}_n)^\top$, where $\hat{z}_i = \text{logit}^{-1}(\hat{\gamma}_0 + \sum_{j=1}^p \hat{\gamma}_j x_{ij})$. For logistic regression, we use the logistic deviance as loss function:

$$L_{\log}(\boldsymbol{\gamma}) = -2 \sum_{i=1}^n \{z_i \log(\hat{z}_i) + (1 - z_i) \log(1 - \hat{z}_i)\},$$

which tends to zero if the predicted probabilities $\hat{\mathbf{z}}$ approach 1 for positives ($z_i = 1$) and 0 for negatives ($z_i = 0$).

2.4. Linear regression

We relate the numerical outcome to the features through linear regression:

$$\mathbb{E}[y_i] = \beta_0 + \sum_{j=1}^p \beta_j x_{ij},$$

where β_0 is the unknown intercept, and $\{\beta_1, \dots, \beta_p\}$ are the unknown slopes. The latter represent the effects of the features on the numerical outcome. Given the estimated coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_p)^\top$, the predicted values are $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_n)^\top$, where $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{ij}$. For linear regression, we use the mean squared error as loss function:

$$L_{\text{lin}}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

which tends to zero if the predicted values $\hat{\mathbf{y}}$ approach the numerical outcomes \mathbf{y} .

2.5. Parameter regularisation

We estimate the logistic and linear regression models by penalised maximum likelihood using lasso (L_1) or ridge (L_2) regularisation, which are generalised by elastic net regularisation [32]. Following the notation from [10], the penalties for logistic and linear regression are equal to

$$P_{\log}(\boldsymbol{y}|\lambda_0, \alpha) = \lambda_0 \sum_{j=1}^p (1 - \alpha) \frac{y_j^2}{2} + \alpha \left| y_j \right|,$$

$$P_{\text{lin}}(\boldsymbol{\beta}|\lambda_1, \alpha) = \lambda_1 \sum_{j=1}^p (1 - \alpha) \frac{\beta_j^2}{2} + \alpha \left| \beta_j \right|,$$

where λ_0 and λ_1 are the regularisation parameters ($\lambda_0 \geq 0$, $\lambda_1 \geq 0$), and α is the elastic net mixing parameter ($0 \leq \alpha \leq 1$). The elastic net penalty collapses to the lasso or ridge penalty if α equals 1 or 0, respectively. We use the lasso penalty to estimate sparse models and the ridge penalty to estimate dense models, but it would also be possible to select α by tuning or combine multiple α by stacking [23].

The penalised loss functions for logistic and linear regression are the sums of the respective loss and penalty functions:

$$M_{\log}(\boldsymbol{y}|\lambda_0, \alpha) = L_{\log}(\boldsymbol{y}) + P_{\log}(\boldsymbol{y}|\lambda_0, \alpha),$$

$$M_{\text{lin}}(\boldsymbol{\beta}|\lambda_1, \alpha) = L_{\text{lin}}(\boldsymbol{\beta}) + P_{\text{lin}}(\boldsymbol{\beta}|\lambda_1, \alpha).$$

Given an elastic net mixing parameter α and the regularisation parameters λ_0 and λ_1 , we can estimate the coefficients \boldsymbol{y} and $\boldsymbol{\beta}$.

2.6. Model combination

We aim to improve the predicted probabilities from penalised logistic regression by accounting for the predicted values from penalised linear regression. Since the predicted values from linear regression are unbounded real numbers, we transform them to the unit interval via the Gaussian cumulative distribution function:

$$\Phi(\hat{y}_i|\mu = c, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{t=-\infty}^{\hat{y}_i} \exp \left\{ -\frac{(t - c)^2}{2\sigma^2} \right\} dt,$$

where μ is the mean ($\mu = c$) and σ^2 is an optimisable variance ($\sigma^2 \geq 0$). This corresponds to the probit link, one of the two most common link functions for binary regression, with a fixed mean parameter for the threshold and a free variance parameter for calibration. The crucial difference to probit regression is that we do not model the *binary* outcome and transform the linear predictor to predicted probabilities but that we model the *numerical* outcome and transform predicted values to predicted probabilities. If the predicted value \hat{y}_i is greater than the threshold c , the probability $\Phi(\hat{y}_i|\mu = c, \sigma^2)$ is greater than 0.5. The variance σ^2 calibrates the probabilities: these diverge to 0 and 1 as σ^2 decreases, and converge to 0.5 as σ^2 increases. Intuitively, this transformation ‘confidently’ assigns samples to classes if σ^2 is small and ‘hesitantly’ if σ^2 is large.

For each sample i , we combine the predicted probability \hat{z}_i from logistic regression and the predicted value \hat{y}_i from linear regression:

$$\hat{p}_i = (1 - \pi)\hat{z}_i + \pi\Phi(\hat{y}_i|\mu = c, \sigma^2),$$

where π is an optimisable weight ($0 \leq \pi \leq 1$). The weighting provides a compromise between the probabilities from penalised logistic regression and the calibrated probabilities from penalised linear regression. By construction, the combined values $\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_n)^\top$ are interpretable as probabilities. As the weight π increases, the contribution of logistic regression decreases, and the contribution of linear regression increases. The combined predicted probability \hat{p}_i is completely determined by logistic or linear regression if π equals 0 or 1, respectively.

Again, we use the logistic deviance as loss function:

$$L_{\text{com}}(\pi, \sigma^2) = -2 \sum_{i=1}^n \{z_i \log(\hat{p}_i) + (1 - z_i) \log(1 - \hat{p}_i)\},$$

which tends to zero if the predicted probabilities $\hat{\mathbf{p}}$ approach 1 for positives ($z_i = 1$) and 0 for negatives ($z_i = 0$).

In short, we combine the predicted probabilities from logistic regression ($\hat{\mathbf{z}}$) and the predicted values from linear regression ($\hat{\mathbf{y}}$) to the predicted probabilities $\hat{\mathbf{p}}$, and we propose to interpret these combined predicted probabilities.

2.7. Parameter optimisation

We fix the elastic net mixing parameter α , tune the regularisation parameters λ_0 and λ_1 , estimate the coefficients $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, and then estimate the weight parameter π and the scale parameter σ^2 :

- tuning λ_0 and λ_1 : We generate two sequences of 100 decreasing values for λ_0 and λ_1 , with the largest values ($\rightarrow \infty$) yielding empty models, and the smallest values ($\rightarrow 0$) yielding full models. In k -fold cross-validation, we split the samples into k folds, repeatedly estimate the coefficients with $k-1$ included folds, and predict the outcomes for the excluded fold. In each iteration, we estimate the coefficients by minimising the penalised loss functions $M_{\log}(\boldsymbol{\gamma}|\lambda_0, \alpha)$ and $M_{\text{lin}}(\boldsymbol{\beta}|\lambda_1, \alpha)$ with respect to $\boldsymbol{\gamma}$ or $\boldsymbol{\beta}$, respectively, via coordinate descent along the regularisation path [10]. After the last iteration, we tune the regularisation parameters λ_0 and λ_1 to minimise the loss functions $L_{\log}(\boldsymbol{\gamma})$ and $L_{\text{lin}}(\boldsymbol{\beta})$.
- estimating $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, π and σ^2 : Given the tuned regularisation parameters $\hat{\lambda}_0$ and $\hat{\lambda}_1$, we re-estimate the coefficients by minimising $M_{\log}(\boldsymbol{\gamma}|\hat{\lambda}_0, \alpha)$ and $M_{\text{lin}}(\boldsymbol{\beta}|\hat{\lambda}_1, \alpha)$ with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. With the estimated coefficients $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$, we calculate the fitted probabilities $\hat{\mathbf{z}}$ from logistic regression and the fitted values $\hat{\mathbf{y}}$ from linear regression. To combine them, we estimate the weight and scale parameters by numerically minimising the loss function $L_{\text{com}}(\pi, \sigma^2)$ with respect to π and σ^2 .

This optimisation procedure first addresses penalised logistic (λ_0 , $\boldsymbol{\gamma}$) and linear (λ_1 , $\boldsymbol{\beta}$) regression separately, and then addresses their combination (π , σ^2). Alternatively, we

might use the expectation-maximisation (EM) algorithm to iteratively estimate $\{\gamma, \beta\}$ and $\{\pi, \sigma^2\}$. In contrast to the EM approach, our two-stage approach has practical advantages: the processing time is only slightly longer than for logistic and linear regression together, $\hat{\gamma}$ and $\hat{\beta}$ are interpretable as estimated effects of the features on the log-odds of the binary outcome or on the identity of the numerical outcome, respectively, and the local minima problem does not affect the estimation of the coefficients.

3. Simulation

3.1. Motivation

In this simulation study, we empirically show that combined regression outperforms not only logistic regression but also ‘calibrated linear regression’ at predicting dichotomised outcomes.

Logistic regression and calibrated linear regression are special cases of the proposed combined regression (with $\pi = 0$ or $\pi = 1$, respectively). While logistic regression requires the binary outcome and returns predicted probabilities, calibrated linear regression requires the numerical outcome and returns predicted values transformed to predicted probabilities.

We illustrate in four examples why the proposed combined regression – combining predicted probabilities from logistic regression and predicted values from linear regression – is suitable for predicting dichotomised outcomes.

3.2. Data generating process

Let n denote the sample size and let p denote the number of features. This is our process for generating features, effects and outcomes:

- features: Let x_{ij} represent feature j for sample i , for any j in $\{1, \dots, p\}$ and any i in $\{1, \dots, n\}$. Simulating all values from a standard Gaussian distribution ($x_{ij} \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$), we obtain the $n \times p$ feature matrix X .
- effects: Let β_j represent the effect of feature j , for any j in $\{1, \dots, p\}$. Simulating all effects from a mixture distribution of a Bernoulli trial with success probability 5% and a standard Gaussian distribution ($\beta_j \sim \mathcal{B}(n = 1, \pi = 0.05) \times \mathcal{N}(\mu = 0, \sigma^2 = 1)$), we obtain the p -dimensional vector $\beta = (\beta_1, \dots, \beta_p)^\top$. While around 95% of the features have no effects, around 5% of the features have negative or positive effects of different sizes ($\sum_{j=1}^p \mathbb{I}[\beta_j \neq 0]/p \approx 0.05$).
- linear predictors: Let η_i represent the linear predictor for sample i , for any i in $\{1, \dots, n\}$. Calculating all linear predictors from the effects and the features ($\eta_i = \sum_{j=1}^p \beta_j x_{ij}$), we obtain the n -dimensional vector $\eta = (\eta_1, \dots, \eta_n)^\top$.
- error terms: Let ϵ_i represent the error term for sample i , for any i in $\{1, \dots, n\}$. Simulating all error terms from a standard Gaussian distribution ($\epsilon_i \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$), we obtain the n -dimensional vector $\epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$.
- outcomes: Let y_i and z_i represent the numerical or binary outcome of sample i , respectively, for any i in $\{1, \dots, n\}$. In each example, the numerical outcome depends on the linear predictor and the error term in a different way (see below). In all examples,

the binary outcome z_i indicates whether the numerical outcome y_i is greater than the threshold zero ($z_i = \mathbb{I}[y_i > 0]$). The corresponding n -dimensional vectors are $\mathbf{y} = (y_1, \dots, y_n)^\top$ and $\mathbf{z} = (z_1, \dots, z_n)^\top$.

3.3. Examples

We provide one representative example where calibrated linear regression should outperform logistic regression, and three illustrative examples where logistic regression should outperform calibrated linear regression. In each example, the equation holds for any i in $\{1, \dots, n\}$.

- (1) standard setting: The numerical outcome equals the sum of the linear predictor and the error term.

$$y_i = \eta_i + \epsilon_i.$$

- (2) latent binary variable: The numerical outcome is clustered around a negative or positive value, depending on whether the linear predictor is below or above the threshold, respectively.

$$y_i = \begin{cases} -2 + \epsilon_i & \text{if } \eta_i < 0 \\ +2 + \epsilon_i & \text{if } \eta_i > 0. \end{cases}$$

- (3) asymmetric relationship: The numerical outcome is not linearly related to the linear predictor but with a square-root below the threshold and a square above the threshold.

$$y_i = \begin{cases} -\sqrt{|\eta_i + \epsilon_i|} & \text{if } \eta_i < 0 \\ +(\eta_i + \epsilon_i)^2 & \text{if } \eta_i > 0. \end{cases}$$

- (4) presence of outliers: The numerical outcome usually equals the sum of the linear predictor and the error term, but rarely there is contamination by a large negative or a large positive number.

$$y_i = \begin{cases} \eta_i + \epsilon_i & \text{with } \mathbb{P} = 95\% \\ \eta_i + \epsilon_i - 1.5\|\boldsymbol{\eta}\|_\infty & \text{with } \mathbb{P} = 2.5\% \\ \eta_i + \epsilon_i + 1.5\|\boldsymbol{\eta}\|_\infty & \text{with } \mathbb{P} = 2.5\%, \end{cases}$$

where the infinity norm $\|\boldsymbol{\eta}\|_\infty$ returns the largest absolute value of $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^\top$.

3.4. Hold-out method

As there is no restriction on the sample size for simulated data, we simulate data for $n_0 = 100$ training samples but $n_1 = 10,000$ testing samples ($n = n_0 + n_1 = 10,100$) in each repetition of the hold-out method. Using $p = 500$ features, we obtain a high-dimensional setting because the number of features is much larger than the number of training samples ($p \gg n_0$). After estimating the parameters of the three regression models with the 100 training samples, we predict the binary outcome for the 10,000 testing samples and compare the predicted probabilities ($0 \leq \hat{p}_i \leq 1$) with the observed classes ($z_i = 0$ or $z_i = 1$).

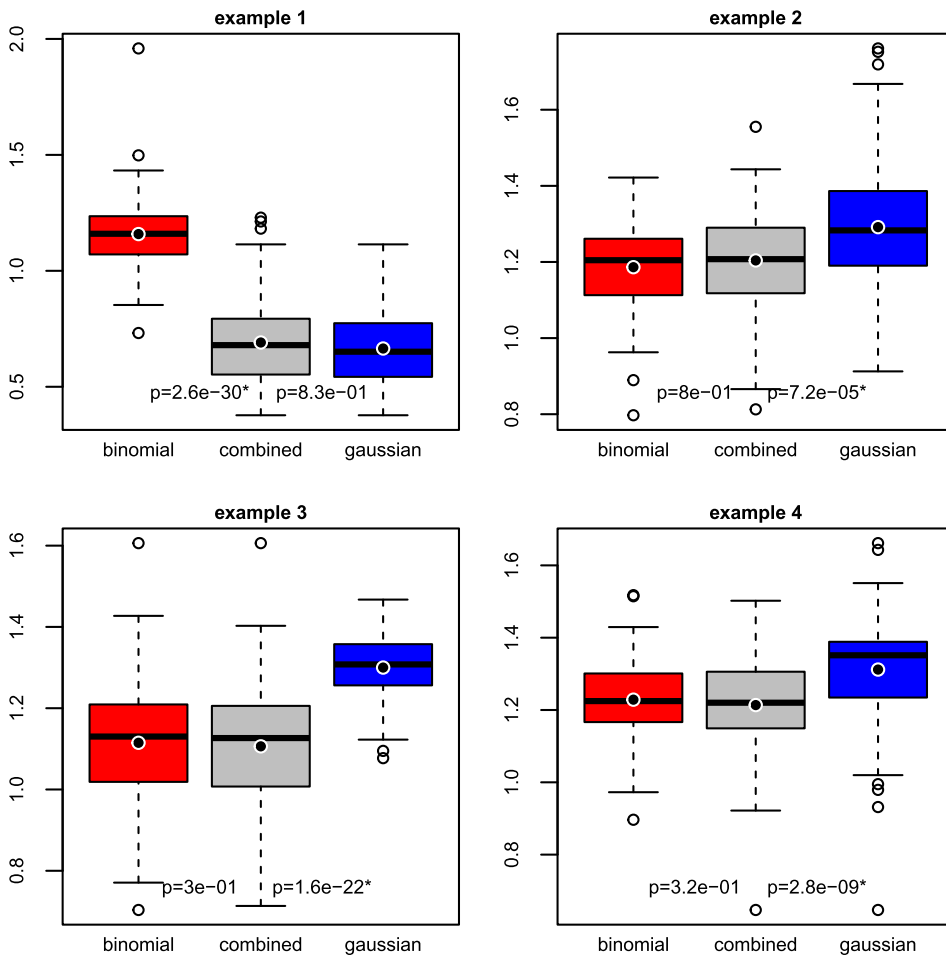


Figure 2. Out-of-sample logistic deviance (lower = better) from logistic regression ('binomial'), combined regression, and calibrated linear regression ('gaussian'), in four simulation settings. The black point added to the box plot represents the mean. A p -value with an asterisk indicates that the decrease in logistic deviance from logistic (left) or calibrated linear (right) to combined regression is statistically significant (one-sided Wilcoxon signed-rank test, Bonferroni-adjusted 5% significance level).

3.5. Predictive performance

For each example in Section 3.3, we performed 100 repetitions of the hold-out method (i.e. simulating 100 sets of training and testing data). Figure 2 summarises the distributions of out-of-sample logistic deviances from logistic regression, calibrated linear regression, and combined regression, each under lasso regularisation. We tested whether combined regression leads to a significantly lower logistic deviance than logistic regression and calibrated linear regression, using the one-sided Wilcoxon signed-rank test.

We find that combined regression is significantly more predictive than logistic regression in the first example and significantly more predictive than calibrated linear regression in the other examples, at the Bonferroni-adjusted 5% level ($p\text{-value} \leq 0.05/8$). Thus, combined regression is highly predictive because it combines the advantages of linear regression (efficiency) and logistic regression (robustness).

4. Application

The Montreal Cognitive Assessment (MoCA) is a screening tool for mild cognitive impairment (MCI) [20]. Although the total MoCA score is a discrete numerical variable ranging from 0 to 30, researchers often model a binary variable indicating the absence or presence of cognitive impairment. For example, Fullard *et al.* [11] use Cox proportional hazards regression to predict the conversion time to MCI, and Caspell-Garcia *et al.* [1] use logistic regression to predict MCI, given the commonly accepted definition of MCI as $\text{MoCA} \leq 25$. Identifying patients at risk of cognitive impairment is important to develop measures for early intervention and prevention, such as cognitive training and physical exercise programmes. Here, we predict cognitive impairment from clinical features, analysing data from a longitudinal cohort study, the Parkinson's Progression Markers Initiative (PPMI) [16].

- features: We extracted the features from the curated baseline data. While the raw data include several hundred unfiltered variables in the categories 'subject characteristics', 'biospecimen', 'digital sensor', 'enrolment', 'imaging', 'medical history', 'motor assessment', 'non-motor assessment', and 'remote data collection', the curated data include 130 relevant variables, either selected or derived from the raw data (Supplementary Table A1). The proportion of missing data is approximately 3%.
- outcomes: We extracted the outcomes from the curated follow-up data, which cover the clinical visits after approximately one, two and three years. The total MoCA score is available for 390, 373 and 363 patients, indicating cognitive impairment ($\text{MoCA} \leq 25$) for 34.4%, 32.4% and 32.2% of the patients, respectively. The apparent improvement likely results from non-random missingness, measurement variation, and training effects after repeated participation in cognitive assessments.

Our objective is to predict from clinical features at baseline which patients will have cognitive impairment after one, two or three years. While logistic regression only exploits the binary outcome of interest 'total MoCA score ≤ 25 versus ≥ 26 ', combined regression also exploits the underlying numerical outcome 'total MoCA score' to predict this probability.

We first imputed missing values in the feature matrix by chained random forests with predictive mean matching (R package [missRanger](#)) and then replaced categorical variables by dummy variables. Instead of imputing the missing values once and analysing one imputed data set ('single imputation'), we imputed the missing values ten times and analysed each imputed data set separately ('multiple imputation').

For each imputed data set, we estimated the predictive performance of logistic and combined regression by nested cross-validation, with an internal loop for training and validation and an external loop for testing. In this unbiased evaluation, we split the samples into five folds, repeatedly train and validate the models with four folds, and test the models with the other fold. To obtain comparable performance estimates, we used the same 5 external and the same 10 internal folds for logistic and combined regression.

Algorithm 1 includes the high-level pseudocode for multiple imputation and nested cross-validation. In all comparisons, we used either lasso or ridge regularisation for both logistic and combined regression. We then examined the percentage change in cross-validated logistic deviance from logistic to combined regression (Supplementary Table A2). For both penalties (L_1, L_2) and all years (1, 2, 3), we observe an improvement for most

imputations (8/10 or 10/10). This improvement also holds for other evaluation metrics, including the misclassification rate and the areas under the receiver operating characteristic and precision-recall curves (Supplementary Table A3).

Algorithm 1 Pseudocode

High-level pseudocode for multiple imputation, external cross-validation, parameter optimisation, and internal cross-validation. We use internal cross-validation to tune the hyperparameters and external cross-validation to estimate the predictive performance. Samples repeatedly switch between the *training* set (included folds in internal loop), the *validation* set (excluded fold in internal loop), and the *test* set (excluded fold in external loop).

(1) MULTIPLEIMPUTATION

input: incomplete data
for i from 1 to 10
 impute missing values
 EXTERNALCROSSVALIDATION
end for
output: mean performance metric

(2) EXTERNALCROSSVALIDATION

input: complete data
 split samples into 5 folds
for j from 1 to 5
 exclude fold j (test set)
 PARAMETEROPTIMISATION
 predict z for fold j
end for
output: performance metric
 (e.g. L_{com})

(3) PARAMETEROPTIMISATION

input: training data
for various λ_0 and λ_1
 INTERNALCROSSVALIDATION
end for
 tune λ_0 and λ_1 (min L_{lin} and L_{log})
 re-estimate β and γ (min M_{lin} and M_{log})
 estimate σ^2 and π (min L_{com})
output: parameter estimates

(4) INTERNALCROSSVALIDATION

input: training data, λ_0, λ_1
 split samples into 10 folds
for k from 1 to 10
 exclude fold k (validation set)
 estimate β and γ (min M_{lin} and M_{log})
 predict y and z for fold k
end for
output: loss (L_{lin} and L_{log})

We used the multi-split approach from [30] to test the prediction error difference between logistic and combined regression. First, we randomly split the samples 50 times into 80% for training and validation, and 20% for testing. Then, for each split, we calculated the squared deviance residuals, whose mean equals the logistic deviance, and compared the paired residuals from logistic and combined regression with the one-sided Wilcoxon signed-rank test. Finally, we calculated the median p -value from the 50 splits, which maintains the type I error rate [30]. For each penalty (L_1, L_2), each year (1, 2, 3), and each imputation (1-10), the median p -value is significant at the 5% level (Supplementary Table A2). Therefore, combined regression leads to significantly better predictions than logistic regression. In this application, however, combined regression does not lead to

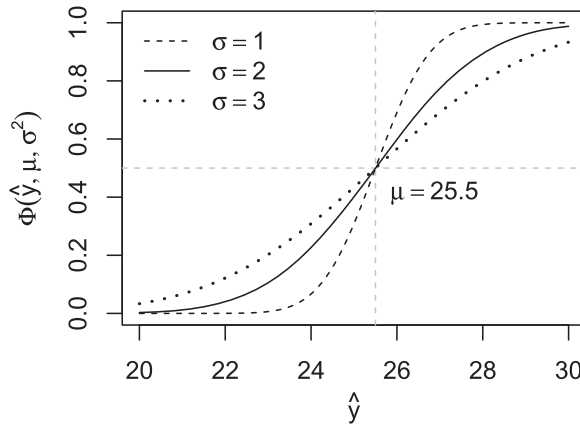


Figure 3. Transformation of predicted values (x -axis) to calibrated probabilities (y -axis) via the Gaussian cumulative distribution function with mean μ and variance σ^2 . Predicted values above μ (vertical line) imply probabilities above 0.5 (horizontal line). While the mean μ equals the threshold c , we need to estimate the variance σ^2 . The probabilities tend to 0 or 1 under small variances and to 0.5 under large variances.

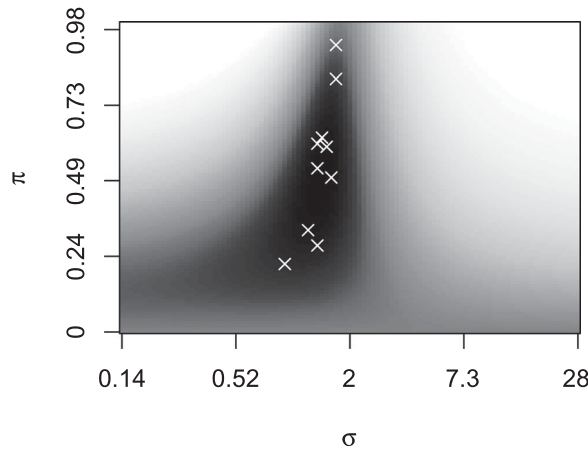


Figure 4. Logistic deviance given weight π (y -axis) and standard deviation σ (x -axis). The region with the lowest mean loss (dark) contains the selected tuning parameters (white crosses). Logistic regression obtains full weight if π equals 0 (bottom), and linear regression if π equals 1 (top). The latter renders predicted probabilities around 0 and 1 if σ is small (left) and around 0.5 if σ is large (right).

significantly better prediction than calibrated linear regression (i.e. combined regression with zero weight for the logistic part). Here, two ensemble learning methods (random forest, gradient boosting) perform worse than ridge and lasso regression (Supplementary Table A4).

To examine weighting and scaling, we refitted combined regression to all folds. Depending on the penalty (L_1, L_2), the year (1, 2, 3), and the imputation (1-10), we estimated weights (π) between 0.20 and 1.00 and variances (σ^2) between 0.16^2 and 1.70^2 (Supplementary Table A2). Together, these estimates determine the combination of the predicted

probabilities from logistic and linear regression. Figure 3 shows the transformation of predicted values from linear regression to calibrated probabilities, and Figure 4 shows the mean loss (for predicting the first-year outcome under lasso regularisation) at different combinations of weights and variances, where the mean is taken over the 10 imputations.

5. Discussion

We have proposed an approach for predicting dichotomised outcomes from high-dimensional data. Combining predicted probabilities from penalised logistic regression and predicted values from penalised linear regression, it achieves a high predictive performance, as shown by simulation and application. The general applicability includes biomedical prediction problems with clinically relevant thresholds.

Ideally, the threshold for dichotomisation is commonly established and splits the samples into two biologically relevant groups. If there is no practical or theoretical justification for setting the threshold equal to a specific value, the need for a probabilistic interpretation is questionable. Special care is required for data-dependent thresholds, because the same criterion typically leads to different thresholds in different data sets, and searching for the ‘optimal’ threshold typically leads to model overfitting.

Our approach integrates numerical information into binary classification, by first modelling binary and numerical outcomes separately and then combining the (calibrated) probabilities. This is related to transforming classifier scores to calibrated probabilities [18,24,31]. Given a threshold and predictions of the numerical outcome, we provide a probabilistic classification. Our aim is an interpretable combination of logistic and linear regression, but we recognise that non-parametric methods for mapping scores onto probabilities might improve the predictive performance.

Instead of applying linear regression on the numerical outcome and transforming the predicted values to probabilities, we could transform the numerical outcome to probabilities and apply logistic regression on the probabilities. Such an approach has previously been developed for low-dimensional settings [13,28]. However, due to the iteration between estimating the nuisance parameter and estimating the coefficients, an extension to high-dimensional settings would be computationally expensive. We estimate them separately but recognise that a simultaneous approach might provide superior performance.

Only the binomial distribution supports binary outcomes, but different distributions support quantitative outcomes. We chose the Gaussian distribution for modelling the quantitative outcome and for transforming predicted values to probabilities. This distribution is supported on the whole real line and has two parameters for thresholding and calibration. It is possible to use different distributions for modelling the observed outcomes or transforming the predicted outcomes. For the first, we could model counts with the Poisson or the negative binomial distribution. For the latter, we could increase flexibility with the three-parameter log-normal distribution [28] or the skew normal distribution.

Since the numerical outcome is normally more informative than the binary outcome, it is not surprising that modelling the numerical outcome next to the binary outcome improves the predictions of the binary outcome. A more important result is that modelling the numerical and the binary outcomes together can provide better predictions than modelling only the numerical outcome. Similarly, numerical features and binary transformations of the same features can be more predictive together than alone [21].

The proposed approach combines the predicted probabilities from logistic regression and the predicted values from linear regression, leaving their estimated coefficients untouched. If the aim was to merge the estimated coefficients from logistic and linear regression into a single set of estimated coefficients, one could use bivariate regression by stacked generalisation for the binary and the numerical outcome [22]. However, this would make the combination of predicted probabilities and predicted values less interpretable.

Although this study focuses on dichotomised outcomes, it is not our intention to advocate dichotomisation. Numerical outcomes should not be binarised, unless there is a strong reason to the contrary. On this condition, we recommend to exploit the binary *and* the numerical outcome. For binary classification in high-dimensional settings, the proposed approach combines both sources of information.

6. Conclusion

For predicting *numerical* outcomes, we suggest to use penalised linear regression to obtain predicted values. For *natural binary* outcomes, we suggest to use penalised logistic regression to obtain predicted probabilities. And for *artificial binary* outcomes (also known as dichotomised outcomes), we propose to combine penalised linear and logistic regression to obtain predicted probabilities.

Reproducibility

The R package `cornet` includes the code for the simulation and the application (<https://cran.r-project.org/package=cornet>). We obtained our results using R 4.3.0 with `cornet` 0.0.8 on a physical machine (aarch64-apple-darwin20, macOS Ventura 13.4). Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (<https://www.ppmi-info.org/data>).

Acknowledgments

We are grateful to Léon-Charles Tranchevent for helpful discussions, to Maharshi Vyas for technical support, and to the anonymous reviewers for constructive criticism.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

We acknowledge support by the Luxembourg National Research Fund (FNR) for the project 'Clinnova—a trans-regional digital health effort' (NCER/23/16695277) and for the project DIGIPD (INTER/ERAPERMED 20/14599012) as part of the European Union's Horizon 2020 research and innovation program. PPMI - a public-private partnership - is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners (<https://www.ppmi-info.org/fundingpartners>).

ORCID

Armin Rauschenberger  <http://orcid.org/0000-0001-6498-4801>

Enrico Glaab  <http://orcid.org/0000-0003-3977-7469>

References

- [1] C. Caspell-Garcia, T. Simuni, D. Tosun-Turgut, I. Wu, Y. Zhang, M. Nalls, A. Singleton, L.A. Shaw, J.-H. Kang, J.Q. Trojanowski, A. Siderowf, C. Coffey, S. Lasch, D. Aarsland, D. Burn, L.M. Chahine, A.J. Espay, E.D. Foster, K.A. Hawkins, I. Litvan, I. Richard, and D. Weintraub, *Multiple modality biomarker prediction of cognitive impairment in prospectively followed de novo Parkinson disease*, PLoS One 12 (2017), Article ID e0175674. doi: [10.1371/journal.pone.0175674](https://doi.org/10.1371/journal.pone.0175674)
- [2] J. Cohen, *The cost of dichotomization*, Appl. Psychol. Meas. 7 (1983), pp. 249–253. doi: [10.1177/014662168300700301](https://doi.org/10.1177/014662168300700301)
- [3] N.V. Dawson and R. Weiss, *Dichotomizing continuous variables in statistical analysis: A practice to avoid*, Med. Decis. Making 32 (2012), pp. 225–226. doi: [10.1177/0272989X12437605](https://doi.org/10.1177/0272989X12437605)
- [4] A.R. de Leon and B. Wu, *Copula-based regression models for a bivariate mixed discrete and continuous outcome*, Stat. Med. 30 (2011), pp. 175–185. doi: [10.1002/sim.4087](https://doi.org/10.1002/sim.4087)
- [5] M. de Paula and C.A.R. Diniz, *Generalized linear regression models incorporating original outcome distributions*, Commun. Stat. – Theory Methods 45 (2016), pp. 5762–5786. doi: [10.1080/03610926.2014.948726](https://doi.org/10.1080/03610926.2014.948726)
- [6] A. Dupuy and D. Nassar, *Dichotomization of primary outcomes serves external validity*, J. Invest. Dermatol. 134 (2014), pp. 266–267. doi: [10.1038/jid.2013.258](https://doi.org/10.1038/jid.2013.258)
- [7] D.P. Farrington and R. Loeber, *Some benefits of dichotomization in psychiatric and criminological research*, Crim. Behav. Ment. Health 10 (2000), pp. 100–122. doi: [10.1002/cbm.349](https://doi.org/10.1002/cbm.349)
- [8] V. Fedorov, F. Mannino, and R. Zhang, *Consequences of dichotomization*, Pharm. Stat. 8 (2009), pp. 50–61. doi: [10.1002/pst.331](https://doi.org/10.1002/pst.331)
- [9] G.M. Fitzmaurice and N.M. Laird, *Regression models for a bivariate discrete and continuous outcome with clustering*, J. Am. Stat. Assoc. 90 (1995), pp. 845–852. doi: [10.2307/2291318](https://doi.org/10.2307/2291318)
- [10] J. Friedman, T. Hastie, and R. Tibshirani, *Regularization paths for generalized linear models via coordinate descent*, J. Stat. Softw. 33 (2010), pp. 1–22. doi: [10.18637/jss.v033.i01](https://doi.org/10.18637/jss.v033.i01) ([glmnet](https://cran.r-project.org/web/packages/glmnet/)).
- [11] M.E. Fullard, B. Tran, S.X. Xie, J.B. Toledo, C. Scordia, C. Linder, R. Purri, D. Weintraub, J.E. Duda, L.M. Chahine, and J.F. Morley, *Olfactory impairment predicts cognitive decline in early Parkinson's disease*, Parkinsonism Relat. Disord. 25 (2016), pp. 45–51. doi: [10.1016/j.parkreldis.2016.02.013](https://doi.org/10.1016/j.parkreldis.2016.02.013)
- [12] F.E. Harrell, *General aspects of fitting regression models – avoiding categorization; ordinal logistic regression*, in *Regression Modeling Strategies*, Springer, Cham, 2015, pp. 311–325. doi: [10.1007/978-3-319-19425-7](https://doi.org/10.1007/978-3-319-19425-7)
- [13] S. Heritier and E. Ronchetti, *Robust binary regression with continuous outcomes*, Can. J. Stat. 32 (2004), pp. 239–249. doi: [10.2307/3315927](https://doi.org/10.2307/3315927)
- [14] O. Kuss, *The danger of dichotomizing continuous variables: A visualization*, Teach. Stat. 35 (2013), pp. 78–79. doi: [10.1111/test.12006](https://doi.org/10.1111/test.12006)
- [15] R.C. MacCallum, S. Zhang, K.J. Preacher, and D.D. Rucker, *On the practice of dichotomization of quantitative variables*, Psychol. Methods 7 (2002), pp. 19–40. doi: [10.1037/1082-989X.7.1.19](https://doi.org/10.1037/1082-989X.7.1.19)
- [16] K. Marek, D. Jennings, S. Lasch, A. Siderowf, and C. Tanner, *The Parkinson Progression Marker Initiative (PPMI)*, Prog. Neurobiol. 95 (2011), pp. 629–635. doi: [10.1016/j.pneurobio.2011.09.005](https://doi.org/10.1016/j.pneurobio.2011.09.005)
- [17] B.K. Moser and L.P. Coombs, *Odds ratios for a continuous outcome variable without dichotomizing*, Stat. Med. 23 (2004), pp. 1843–1860. doi: [10.1002/sim.1776](https://doi.org/10.1002/sim.1776)
- [18] M.P. Naeini and G.F. Cooper, *Binary classifier calibration using an ensemble of piecewise linear regression models*, Knowl. Inf. Syst. 54 (2018), pp. 151–170. doi: [10.1007/s10115-017-1133-2](https://doi.org/10.1007/s10115-017-1133-2)
- [19] O. Naggara, J. Raymond, F. Guilbert, D. Roy, A. Weill, and D.G. Altman, *Analysis by categorizing or dichotomizing continuous variables is inadvisable: An example from the natural history of unruptured aneurysms*, AJNR Am. J. Neuroradiol. 32 (2011), pp. 437–440. doi: [10.3174/ajnr.A2425](https://doi.org/10.3174/ajnr.A2425)
- [20] Z.S. Nasreddine, N.A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J.L. Cummings, and H. Chertkow, *The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment*, J. Am. Geriatr. Soc. 53 (2005), pp. 695–699. doi: [10.1111/j.1532-5415.2005.53221.x](https://doi.org/10.1111/j.1532-5415.2005.53221.x)

- [21] A. Rauschenberger, I. Ciocănea-Teodorescu, M.A. Jonker, R.X. Menezes, and M.A. van de Wiel, *Sparse classification with paired covariates*, Adv. Data Anal. Classif. 14 (2020), pp. 571–588. doi: [10.1007/s11634-019-00375-6](https://doi.org/10.1007/s11634-019-00375-6) ([palasso](#)).
- [22] A. Rauschenberger and E. Glaab, *Predicting correlated outcomes from molecular data*, Bioinform. 37 (2021), pp. 3889–3895. doi: [10.1093/bioinformatics/btab576](https://doi.org/10.1093/bioinformatics/btab576) ([joinet](#)).
- [23] A. Rauschenberger, E. Glaab, and M.A. van de Wiel, *Predictive and interpretable models via the stacked elastic net*, Bioinform. 37 (2021), pp. 2012–2016. doi: [10.1093/bioinformatics/btaa535](https://doi.org/10.1093/bioinformatics/btaa535) ([starnet](#)).
- [24] J. Schwarz and D. Heider, *GUESS: Projecting machine learning scores to well-calibrated probability estimates for clinical decision-making*, Bioinform. 35 (2019), pp. 2458–2465. doi: [10.1093/bioinformatics/bty984](https://doi.org/10.1093/bioinformatics/bty984) ([CalibratR](#)).
- [25] Y. Shentu and M. Xie, *A note on dichotomization of continuous response variable in the presence of contamination and model misspecification*, Stat. Med. 29 (2010), pp. 2200–2214. doi: [10.1002/sim.3966](https://doi.org/10.1002/sim.3966)
- [26] D.L. Streiner, *Breaking up is hard to do: The heartbreak of dichotomizing continuous data*, Can. J. Psychiatry 47 (2002), pp. 262–266. doi: [10.1177/070674370204700307](https://doi.org/10.1177/070674370204700307)
- [27] S. Suissa, *Binary methods for continuous outcomes: A parametric alternative*, J. Clin. Epidemiol. 44 (1991), pp. 241–248. doi: [10.1016/0895-4356\(91\)90035-8](https://doi.org/10.1016/0895-4356(91)90035-8)
- [28] S. Suissa and L. Blais, *Binary regression with continuous outcomes*, Stat. Med. 14 (1995), pp. 247–255. doi: [10.1002/sim.4780140303](https://doi.org/10.1002/sim.4780140303)
- [29] R. Ulrich and M. Wirtz, *On the correlation of a naturally and an artificially dichotomized variable*, Br. J. Stat. Psychol. 57 (2004), pp. 235–251. doi: [10.1348/0007110042307203](https://doi.org/10.1348/0007110042307203)
- [30] M.A. van de Wiel, J. Berkhof, and W.N. van Wieringen, *Testing the prediction error difference between 2 predictors*, Biostat. 10 (2009), pp. 550–560. doi: [10.1093/biostatistics/kxp011](https://doi.org/10.1093/biostatistics/kxp011)
- [31] B. Zadrozny and C. Elkan, *Transforming classifier scores into accurate multiclass probability estimates*, in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 694–699. doi: [10.1145/775047.775151](https://doi.org/10.1145/775047.775151)
- [32] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Stat. Soc., B: Stat. Methodol. 67 (2005), pp. 301–320. doi: [10.1111/j.1467-9868.2005.00503.x](https://doi.org/10.1111/j.1467-9868.2005.00503.x)