Diagnostic and
Prognostic Research

RESEARCH Open Access

Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults

Anita L. Lynam[1], John M. Dennis[1], Katharine R. Owen[2,3], Richard A. Oram[1], Angus G. Jones[1], Beverley M. Shields[1] and Lauric A. Ferrat[1*]

Abstract

Background: There is much interest in the use of prognostic and diagnostic prediction models in all areas of clinical medicine. The use of machine learning to improve prognostic and diagnostic accuracy in this area has been increasing at the expense of classic statistical models. Previous studies have compared performance between these two approaches but their findings are inconsistent and many have limitations. We aimed to compare the discrimination and calibration of seven models built using logistic regression and optimised machine learning algorithms in a clinical setting, where the number of potential predictors is often limited, and externally validate the models.

Methods: We trained models using logistic regression and six commonly used machine learning algorithms to predict if a patient diagnosed with diabetes has type 1 diabetes (versus type 2 diabetes). We used seven predictor variables (age, BMI, GADA islet-autoantibodies, sex, total cholesterol, HDL cholesterol and triglyceride) using a UK cohort of adult participants (aged 18–50 years) with clinically diagnosed diabetes recruited from primary and secondary care (n = 960, 14% with type 1 diabetes). Discrimination performance (ROC AUC), calibration and decision curve analysis of each approach was compared in a separate external validation dataset (n = 504, 21% with type 1 diabetes).

Results: Average performance obtained in internal validation was similar in all models (ROC AUC ≥ 0.94). In external validation, there were very modest reductions in discrimination with AUC ROC remaining ≥ 0.93 for all methods. Logistic regression had the numerically highest value in external validation (ROC AUC 0.95). Logistic regression had good performance in terms of calibration and decision curve analysis. Neural network and gradient boosting machine had the best calibration performance. Both logistic regression and support vector machine had good decision curve analysis for clinical useful threshold probabilities.

(Continued on next page)

* Correspondence: l.a.ferrat@exeter.ac.uk
[1] Institute of Biomedical and Clinical Science, College of Medicine and Health, University of Exeter, Exeter EX2 5DW, UK
Full list of author information is available at the end of the article

(Continued from previous page)

Conclusion: Logistic regression performed as well as optimised machine algorithms to classify patients with type 1 and type 2 diabetes. This study highlights the utility of comparing traditional regression modelling to machine learning, particularly when using a small number of well understood, strong predictor variables.

Keywords: Machine learning, Logistic regression, Model selection

Background

There is much interest in the use of prognostic and diagnostic prediction models in all areas of clinical medicine including cancers [1, 2], cardiovascular disease [3, 4] and diabetes [5, 6]. These models are increasingly being used as web-calculators [7–9] and medical apps for smartphones [10–12], and many have been incorporated into clinical guidelines [13–17].

There are many different approaches that can be used for developing these models. Classic statistical models such as logistic regression are commonly applied but there is increasing interest in the application of machine learning to improve prognostic and diagnostic accuracy in clinical research ([18–21] with many examples of their use [22]. Machine learning (ML) is a data science field dealing with algorithms in which computers (the machines) adapt and learn from experience (data), these algorithms have the ability to process the vast amounts of data such as medical images, biobank and electronic health care records. Supervised learning is the most widely employed category of machine learning. In supervised learning, the machine predicts the value of an outcome (either binary or continuous) trained on a set of predictor variables.

There are many applied studies comparing the performance of classic models to different machine learning algorithms [23–34] but their findings are inconsistent. Many such comparison studies have limitations; not all use non-default parameter settings (hyperparameter tuning) or have validated performance on external data [35]. Discrimination, as measured by area under the receiver operating characteristic curve, is almost always provided but studies have rarely assessed whether risk predictions are reliable (calibration) [35].

We aimed to use a methodological approach to explore and compare the performance of machine learning and a classic statistical modelling approach using an example of a diabetes classification model. Classification of diabetes offers an interesting case study as it is an area where there is considerable misclassification in clinical practice. Type 1 diabetes and type 2 diabetes can be hard to distinguish between, particularly in adults. Correct classification is really important for the patient, particularly in terms of treatment. People with type 1 require insulin injections to prevent life-threatening diabetic ketoacidosis, whereas people with type 2 diabetes can treat their high blood glucose with diet or tablets.

Methods

We selected a classic model, logistic regression (LR) with linear effects only, and six supervised machine learning algorithms that (1) were appropriate for classification problems and (2) had been used previously in medical applications: gradient boosting machine (GBM), multivariate adaptive regression spline (MARS), neural network (NN), k-nearest neighbours (KNN), random forest (RF) and support vector machine (SVM). We trained models using each algorithm, incorporating hyperparameter tuning, and compared the performance of the optimised models on a separate external validation dataset.

Study population—training dataset

The Exeter cohort includes 1378 participants, with known diabetes (identified from the clinical record and confirmed by the participant on recruitment) from Exeter, UK [36–39]. Participants with gestational diabetes, known secondary or monogenic diabetes or a known disorder of the exocrine pancreas were excluded. Summaries of the cohorts including recruitment and data collection methods are shown in Supplementary Table 1 and Figure S1 (see Additional file 1).

Study population—external validation dataset

Five hundred sixty-six participants were identified from the Young Diabetes in Oxford (YDX) study [40]. Participants were recruited in the Thames Valley region, UK, and diagnosed with diabetes up to the age of 50 years. The same eligibility criteria were applied to this cohort.

All participants included in this study (internal and external validation datasets) were of white European origin. Summaries of the cohort including recruitment and data collection methods are shown in Supplementary Table 1 (see Additional file 1).

Model outcome: type 1 and type 2 diabetes definition

We used a binary outcome of type 1 or type 2 diabetes. Type 1 diabetes was defined as having insulin treatment within ≤ 3 years of diabetes diagnosis and severe insulin deficiency (non-fasting C peptide < 200 pmol/L). Type 2 diabetes was defined as either (1) no insulin requirement for 3 years from diabetes diagnosis or (2) where insulin was started within 3 years of diagnosis, substantial retained endogenous insulin secretion (C-peptide > 600 pmol/L) at ≥ 5 years diabetes duration. Participants not meeting the above criteria or with insufficient information were excluded from analysis, as the type of diabetes and rapid insulin requirement could not be robustly defined (n = 342 in the training dataset). These exclusions are unavoidable and in our opinion are unlikely to introduce systemic bias or affect the main question being addressed which is comparative performance of the different modelling approaches. The major reason for exclusion from analysis was short diabetes duration (223 of 342 excluded), and this is because the outcome (based on that the development of severe insulin deficiency is often absent at diagnosis in T1D) cannot be defined in recent onset disease. A tiny number of participants are excluded due to intermediate C-peptide which means outcome cannot be robustly defined (n = 37). In 87 participants, a saved serum sample for C-peptide measurement was not available, because serum was not stored in the very early stages of the DARE study. C-peptide was measured in all other participants in these cohorts that required measurement for the outcome.

Predictor variables

We used seven pre-specified predictor variables, age at diagnosis, BMI, GADA islet-autoantibodies, sex, total cholesterol, HDL cholesterol and triglycerides. Age at diagnosis and sex were self-reported by the participant. Height and weight were measured at study recruitment by a research nurse to calculate BMI. Total cholesterol, HDL cholesterol and triglycerides were extracted from the closest NHS record. Continuous variables were standardised [41]. GADA islet-autoantibodies were dichotomized into negative or positive based on clinically defined cut-offs, in accordance with clinical guidelines [42].

We removed all observations with missing predictor values (complete-case analysis), respectively: 74 for the training cohort (74 HDL cholesterol and 68 triglycerides values missing) and 61 for the external validation cohort (53 sex value missing, 8 total cholesterol missing). We finally removed any observation with clinically impossible values (z score > 50): 2 for the training cohort and 1 for the external validation cohort. Nine hundred sixty participants met inclusion criteria and were included in the training dataset, of whom 135 (14%) were classified as having type 1 diabetes. Five hundred four participants (type 1 diabetes, n = 105 (21%)) in the YDX cohort met criteria and were included in the external validation dataset. Compared to the participants in Exeter cohort, the participants in the YDX cohort were younger at diagnosis (median 37 years vs 43 years, p < 0.001), had a lower BMI (median 31 kg/m2 vs 33 kg/m^2, p < 0.001), had a higher percentage of GADA (20% versus 13%, p < 0.001) and a higher prevalence of type 1 diabetes (as defined by our model outcome definition in the 'Study population—external validation dataset' section) (21% vs 14%, p < 0.001) (Supplementary Table 2 (see Additional file 1) for participant characteristics).

Model training

All models were trained using the entire training dataset. We evaluated seven classification algorithms: gradient boosting machine (GBM), logistic regression (LR), multivariate adaptive regression spline (MARS), neural network (NN), k-nearest neighbours (KNN), random forest (RF) and support vector machine (SVM). For SVM, we used the radial basis function kernel parameter [41], and for NN, we used the most commonly used single-hidden-layer neural network [41] trained using quasi-Newton back propagation (BFGS) [43] optimisation method. There are no clear guidelines regarding either the choice of algorithms or the advantages and disadvantages of each in specific clinical settings. A brief summary of each algorithm is shown in Table 1.

We used a grid search to tune the model parameters (hyperparameter tuning) [60], i.e. optimise the performance of the machine learning algorithm. The hyperparameter metrics applied in the grid searches are shown in Supplementary Table 3 (see Additional file 1). To fit the models over the whole training dataset, we first estimated the hyperparameters with 5-fold cross-validation and fit the models with the estimated algorithms. Internal validation was performed using nested cross-validation. The nested cross-validation consists of an inner loop cross-validation nested in an outer cross-validation. The inner loop is responsible for model selection/hyperparameter tuning (similar to validation set), while the outer loop is for error estimation (test set). For each loop, we used 5 folds. Nested cross-validation is only used to estimate the performance measures, and the final model is fitted on the whole training dataset.

Optimal models were selected using the maximum mean area under the receiver operating characteristic curve (ROC AUC) calculated in the cross-validation. Supplementary Table 3 (see Additional file 1) includes the final model tuning parameters selected for the optimal models in the cross-validation resampling. We computed the 95% CI by assuming that the variation around the mean is normally distributed and computed a standard normalised interval using the different values estimated on each fold computed by the cross-validation.

Table 1 Algorithm description and references

Algorithm	Description	References
Logistic regression	A classic statistical algorithm for binary outcomes that use maximum likelihood estimation. It is fully parametric. There are no model hyperparameters to be set. Coefficients are adjusted to allow for dependence between the characteristics. It is useful for inference, estimation, interpretation and prediction.	[41, 44–46]
Random forest	An algorithm that grows a large ensemble of classification trees on bootstrapped samples using a random selection of the predictor variables and performs bagging for class selection; after all the trees have been grown, the predicted class is determined from the average estimated class probability calculated over the ensemble of trees.	[41, 47, 48]
Gradient boosting machine	An ensemble learning technique similar to random forest in the sense they average a large number of decision trees to make prediction. The difference between the two is the application of gradient boosting. In gradient boosting, the decision trees are trained sequentially with the weights of each successive model adjusted based on reducing the errors of the previous model. The predicted class is determined from the average estimated class probability (or majority vote of predicted class) calculated over the ensemble of trees.	[41, 49, 50]
Multivariate adaptive regression spline	MARS and logistic regression share similarities. For the logistic regression model, the logarithm of the odds is fitted with a linear combination of the predictors. For the MARS model, the logarithm of the odds is fitted with splines to cover non-linear and interactions terms. The hinge function (sometimes called rectifier) is used to model the splines.	[51]
Neural network	A method using an adaptive and non-sequential approach to learning that mimics a biological neural network. It is a non-parametric technique where signals travel from the first layer (the input layer), to the last layer (the output layer). Each layer is made of a set of neurons. The output of each neuron is computed by some non-linear function of the sum of its weighted inputs from neurons from the previous layer. The weight increases or decreases the strength of the signal at a connection.	[41, 52–55]
K-nearest neighbours	A model-free method; it is a type of instance-based learning or lazy learning in which there is no training phase, instead the algorithm memorises the training data. Based on the principle that observations located close together in n-dimensional space will have the same outcome, the classification process involves a search the entire dataset for the k training points closest in Euclidean distance (k-neighbours), the probability predicted class is determined based on the average vote of the actual class among these k-neighbours.	[41, 53, 56, 57]
Support vector machine	It is a quadratic optimisation problem involving minimising penalties and maximising margin width, and the two classes are separated by constructing nonlinear decision boundaries (hyperplanes) using a kernel trick that maximises the margin between them. The produced posterior estimates are a rescaled version of the original classifiers scores through a logistic transformation.	[41, 58, 59]

Model performance measures

We used ROC AUC [61] as the summary metrics to evaluate model discrimination. The ROC AUC quantifies the probability that the risk scores from a randomly selected pair of individuals with and without this condition are correctly ordered. A value of 1 indicates a perfect test.

We assessed calibration visually using calibration plots, computing calibration performance measure, i.e. calibration slope (the closer to 1 the better) and the calibration in the large (the closer to zeros the better). The slope coefficient beta of the linear predictors reflects the deviations from the ideal slope of 1.

We compared the performance of the model to support decision-making with decision curve analysis [62]. In decision curve analysis, a clinical judgement of the relative value of benefits (treating a true-positive case) and harms (treating a false-positive case) associated with prediction models is made for different threshold probability [63]. The net benefit is computed by subtracting the proportion of all patients who are false-positive from the proportion who are true-positive, weighting by the relative harm of a false-positive and a false-negative result.

External testing

For each optimal model developed in the training dataset, external performance was evaluated in the YDX study cohort and compared to the internal (cross-validation resampling) performance. Calibration was investigated using calibration curves. We also checked for Pearson's correlation in the predictions from each model.

Software

All analyses were performed using R software (version 3.5.2). Model training was performed using the Caret R package [64–68].

Code

In the supplementary material, we share the code to allow reproduction of similar comparisons of machine learning algorithms with any number of predictor variables (see Additional file 2).

Results

The average (mean) performance ROC AUC for the optimal models obtained in the resampling was high in all models (ROC AUC ≥ 0.93) (Table 2) with small difference in performance between models.

There was a decrease in the ROC AUC of all models when they were applied to the external validation dataset (Table 2), but all still showed high levels of performance (ROC AUC ≥ 0.92, Figure S3) for all model. Model predictions were highly correlated across models (Figure S2 (see Additional file 1)). ROC AUC performance was similar when fitting the model with or without resampling.

In the calibration tests performed on the external validation dataset, GBM and NN shows very good calibration performance with a calibration in the large close to 0 and a calibration slope close to 1. Logistic regression and support vector machine have a satisfactory calibration results but the likelihood to predict type 1 diabetes is on average slightly underestimated. All other models have unsatisfactory calibration performance (Fig. 1 and Table 3 (calibration in the large values < 0 indicate overestimating risk)) and there was evidence of visual miscalibration in these models (often due to an underestimation of type 1).

Figure S3 highlights that the most performing machine learning algorithms give similar predictions. In this figure, the prediction for each algorithm is plotted for each observation in XYD. SVM, NN and LR predictions are strongly correlated (LR-NN, 0.992; LR-SVM, 0.99; NN-SVM, 0.983). For all models, the majority of predicted probability is below 0.3 (as expected, 79% of people do not have type 1 diabetes). Excepted for the KNN model, few predictions lie between 0.3 and 0.7.

Figure 2 is the decision curve analysis where the net benefit is plotted against the threshold probability. The LR model is superior or similar to the other models across a wide range of threshold probabilities but becomes worse than the other models for higher threshold probabilities. In practice, it is likely clinicians would be cautious and treat patients with insulin at much lower probabilities threshold, as not receiving correct treatment for a type 1 diabetes can be life-threatening while giving insulin to a patient with type 2 diabetes is inconvenient and expensive but not life-threatening.

Table 2 ROC AUC [95% CI] performance comparison of the seven models applied to the internal and external validation datasets. Internal validation was estimated with 5-fold-nested cross-validation while external validation was performed on the YDX dataset

Model	Internal validation	External validation
Gradient boosting machine	0.96 [0.94, 0.98]	0.93 [0.90, 0.96]
K-nearest neighbours	0.93 [0.90, 0.97]	0.92 [0.89, 0.95]
Logistic regression	0.96 [0.93, 0.98]	0.95 [0.92, 0.97]
MARS	0.96 [0.90, 0.99]	0.94 [0.92, 0.97]
Neural network	0.96 [0.93, 0.99]	0.94 [0.92, 0.97]
Random forest	0.95 [0.92, 0.98]	0.94 [0.91, 0.96]
Support vector machine	0.96 [0.93, 0.98]	0.94 [0.92, 0.97]

The poor performance at high threshold probabilities is due to the fact that the LR model, like the SVM model, tends to overestimate the risk of having type 1 diabetes for people with the highest risk (risk above 85%), see Fig. 1.

Discussion

Summary of main findings

We found similar performance when applying logistic regression and six optimised machine learning algorithms to classify type 1 and type 2 diabetes, in both internal and external validation datasets. Discrimination was high for all models, while logistic regression showed the numerically highest discrimination in external validation differences in discrimination were small. Neural network and gradient boosting machine had the best calibration performance, with logistic regression and support vector machine also showing satisfactory calibration.

Strengths and limitations

Strengths of our study include the use of a systematic approach to model comparison dealing with limitations from previous studies [35, 70] including (1) use of different datasets to train and test models, (2) optimisation of tuning parameters [24, 30], (3) calibration [18] and (4) decision curve analysis. We have used the same dataset to train all our models; since model performance will differ between settings, the use of the same dataset is crucial for valid model comparisons. The choice of tuning parameters will affect the performance of the model [60], and we have optimised our models by applying hyperparameter tuning using a recognised grid search approach. We have increased the validity of our results by using an external validation dataset.

We have compared several machine learning algorithms that have been selected for their suitability to our setting. The use of only seven predictor variables means that we have a very low risk of over-fitting; for machine learning algorithms, it has been suggested that over ten times as many events per variable is required to achieve stable results compared to traditional statistical modelling [69]. The use of only seven predictors may also be considered as a limitation of our study since these machine learning algorithms are designed to deal with larger datasets and more variables. However, working with a few meaningful predictors is common in clinical settings. Knowing the performance of machine learning models using low numbers of predictors is important. It is possible that with more variables or more observations, machine learning approaches may prove more discriminative. However, we have achieved excellent

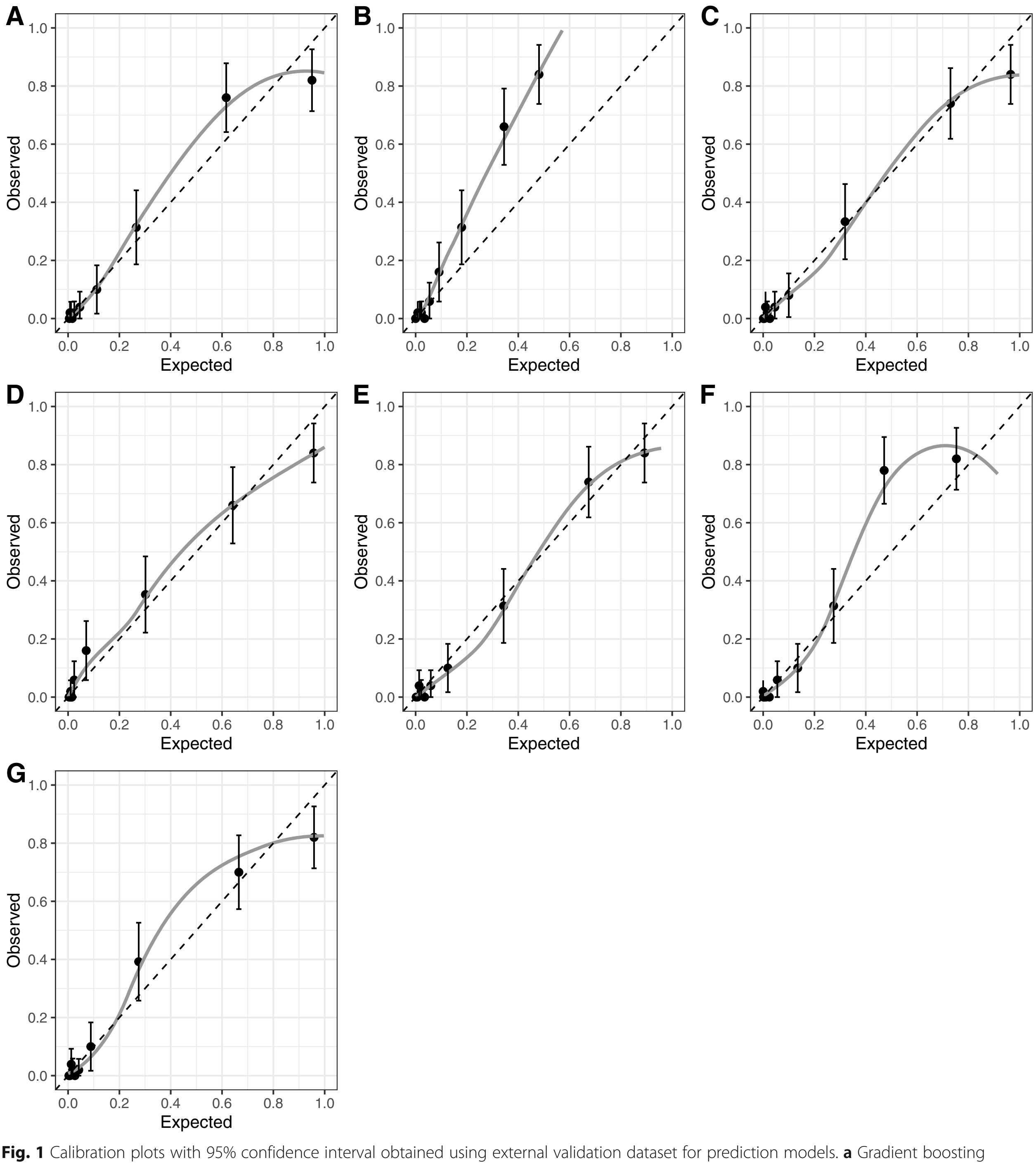

Fig. 1 Calibration plots with 95% confidence interval obtained using external validation dataset for prediction models. **a** Gradient boosting machine. **b** K-nearest neighbours. **c** Logistic regression. **d** MARS. **e** Neural network. **f** Random forest. **g** Support vector machine. Legend: Dashed line = reference line, solid black line = linear model

performance using just these seven predictors. Another limitation of our study is that we judge the model only on its performance. In real practice, we would want to consider ease of implementation and interpretation when selecting the 'best' model.

LR, SVM and NN are the models with the highest ROC AUC. If accuracy of estimated probability were an importance factor, NN, LR, GBM and SVM would be best approaches. Overall, the notion of best model is context-dependent, but in this study, the models

Table 3 Calibration test results on external validation dataset. Calibration-in-the-large indicates whether predicted probabilities are, on average, too high (value below 0) or too low (value above 0). Conversely, the calibration slope quantifies whether predicted risks are, on average, too extreme (value below 1) or too invariant (value above 1)

Model	Calibration slope ($\boldsymbol{b_L}$)	Calibration-in-the-large ($\boldsymbol{a}\|\boldsymbol{b_L} = 1$)
Gradient boosting machine	0.979	− 0.005
K-nearest neighbours	1.495	0.046
Logistic regression	0.903	− 0.039
MARS	0.799	0.081
Neural network	0.995	− 0.031
Random forest	1.412	0.065
Support vector machine	0.914	− 0.028

perform similarly. In terms of clinical utility, LR and SVM appeared to perform slightly better than other models.

The observed decrease in ROC AUC when assessed in the external validation data highlights the importance of external validation to test the transportability of models. Indeed, all of the algorithms slightly underperformed in the external validation set. The model fit on the training data set might be over-fitted and their performance could be overestimated despite a rigorous internal validation (see the difference between internal and external performance in Table 2). However, the most likely reason is that the YDX population has a smaller range in age and BMI, and GADA is less discriminative in YDX compared to the Exeter cohort. This may diminish performance and does not necessarily mean over-fitting.

The performance of LR on both internal and external validation datasets shows that classic algorithms can perform as well as more advanced algorithms even when disadvantaged by assuming linearity in the predictors. LR models are relatively easy to use and understand compared to machine learning algorithms where usage is limited by the difficultly of interpreting the model, often referred to as a 'black boxes'. LR models also have a strong theoretical background which leads to the possibility of using well-defined statistical tests to explore the statistical significance of the variables. There is an increasing number of studies demonstrating that LR can perform as well if not better, in a large number of settings [35]. However, we could not find a study that compared machine learning algorithms with optimised

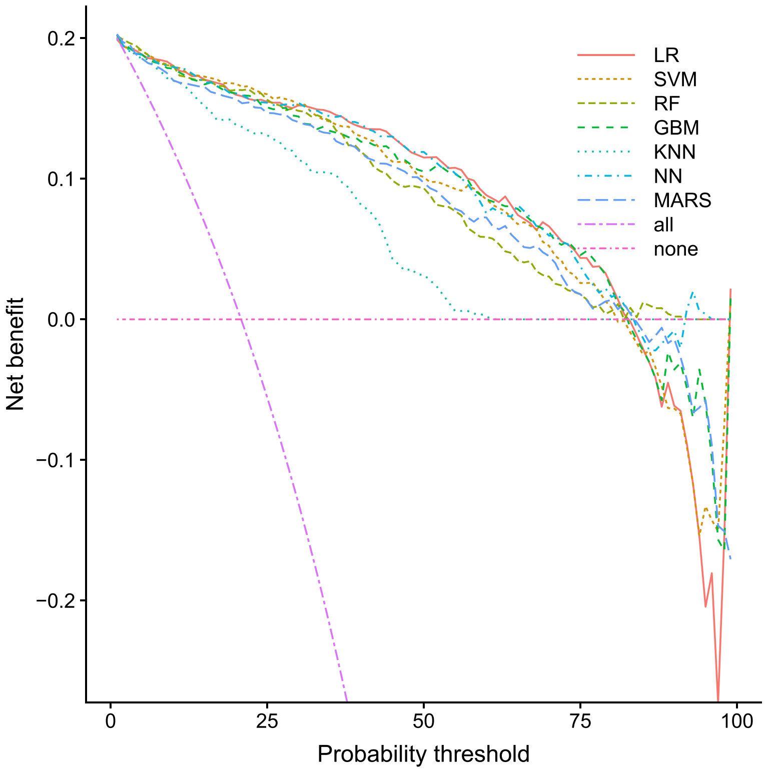

Fig. 2 Decision curve analysis obtained using external validation dataset for prediction models. The graph gives the expected net benefit per patient relative to treat all patients as type 2 diabetes. The unit is the benefit associated with one patient with type 1 diabetes receiving the correct treatment. 'all': assume all patients have type 1 diabetes. 'none': assume no patients have type 1 diabetes

hyperparameters versus LR on an external dataset as we have done in this study which shows again that LR performs as well as more complex approaches.

While real-world data medical applications are likely to be unbalanced, the use of sampling methods such as Synthetic Minority Over-Sampling Technique (SMOTE) might improve model prediction performance [70]. We compared the use of SMOTE to the classic approach without resampling. Nevertheless, we only present the results without SMOTE as similar ROC AUC, but better result calibration and decision curve analysis performance were achieved without it.

We have shown through this study that machine learning performs similarly for this prediction problem; however, some differences subsist. As previously described [71], each database is unique and there is no 'free lunch', i.e. if an algorithm performs well on a certain class of problems, then it necessarily pays for that with degraded performance on the set of other problems [35, 72]. It is thus important to test different algorithms benchmarked against logistic regression to identify if one algorithm outperforms the other; if performance is similar, then the simplest and most interpretable model can be used.

Conclusion

In a diabetes classification setting with three strongly predictive variables, a classic logistic regression algorithm performed as well as more advanced machine algorithms. This study highlights the utility of comparing traditional regression modelling to machine learning, particularly when using a small number of well understood, strong predictor variables. Furthermore, this article highlights once again the need to perform external validation when selecting models as we demonstrate that all algorithms can underperform on external data.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s41512-020-00075-2.

Additional file 1: Figure S1. Flow diagram of participants through the model development stages. T1D: type 1 diabetes, T2D: type 2 diabetes. **Figure S2.** ROC AUC plots obtained using external validation dataset for seven prediction models. Legend: Solid lines: black = Support Vector Machine, dark grey = Logistic Regression, light grey = Random Forest. Dotted lines: black = Neural Network, dark grey = K-Nearest Neighbours, light grey = Gradient Boosting Machine. **Figure S3.** Correlation coefficient matrix and scatter plot of model predictions obtained from external test validation data.

Additional file 2. R script.

Abbreviations

GADA: Glutamic acid decarboxylase antibodies; YDX: Young Diabetes in Oxford; LR: Logistic regression; SVM: Support vector machine; GBM: Gradient boosting machine; NN: Neural network; KNN: K-nearest neighbours; RF: Random forest; SMOTE: Synthetic Minority Over-Sampling Technique; ROC AUC: Area under the receiver operating characteristic curve

Acknowledgements

The authors thank the participants who took part in these studies and the research teams who undertook cohort recruitment. We thank Catherine Angwin of the NIHR Exeter Clinical Research Facility for the assistance with data preparation, and Rachel Nice of the Blood Sciences Department, Royal Devon and Exeter Hospital for the assistance with sample analysis.

Authors' contributions

ALL and LAF conceived the idea and designed the study. ALL, KRO and AGJ researched the data. ALL and LAF analysed the data with assistance from BMS. JMD, KRO, RAO and AGJ discussed and contributed to the study design and provided support for the analysis and interpretation of the results. ALL drafted the manuscript with assistance from LAF. All authors critically revised the manuscript and approved the final version. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding

This work was funded by NIHR Clinician Scientist award CS-2015-15-018. The Young Diabetes in Oxford Study was funded by the European Community FP7 programme CEED3 (HEALTH-F2-2008-223211), the Oxford Hospitals Charitable Fund and by the NIHR Oxford Biomedical Research Centre. The Diabetes Alliance for Research in England (DARE) study was funded by the Wellcome Trust and supported by the Exeter NIHR Clinical Research Facility. The MASTERMIND study was funded by the UK Medical Research Council (MR/N00633X/1) and supported by the NIHR Exeter Clinical Research Facility. The PRIBA study was funded by the National Institute for Health Research (UK) (DRF-2010-03-72) and supported by the NIHR Exeter Clinical Research Facility.

B.M.S is supported by the NIHR Exeter Clinical Research Facility. A.G.J is supported by an NIHR Clinician Scientist award (CS-2015-15-018). R.A.O is supported by a Diabetes UK Harry Keen Fellowship (16/0005529).

The views expressed in this article are those of the authors do not necessarily represent those of the National Institute for Health Research, the National Health Service or the Department of Health.

Availability of data and materials

The data that support the findings of this study are available from University of Exeter Medical School/Oxford University but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of University of Exeter Medical School/Oxford University. R code is made available in supplementary file (see Additional file 2).

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Institute of Biomedical and Clinical Science, College of Medicine and Health, University of Exeter, Exeter EX2 5DW, UK. [2]Oxford Centre for Diabetes Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford OX3 7LE, UK. [3]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals Foundation Trust, John Radcliffe Hospital, Oxford, UK.

Received: 7 August 2019 Accepted: 26 March 2020
Published online: 04 June 2020

References

1. Shariat SF, Karakiewicz PI, Roehrborn CG, Kattan MW. An updated catalog of prostate cancer predictive tools. Cancer. 2008;113(11):3075–99.

2. Amir E, Freedman OC, Seruga B, Evans DG. Assessing Women at High Risk of Breast Cancer: A Review of Risk Assessment Models. J Natl Cancer Inst. 2010;102(10):680–91.
3. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353:i2416.
4. Wessler BS, Lai Yh L, Kramer W, Cangelosi M, Raman G, Lutz JS, et al. Clinical prediction models for cardiovascular disease: tufts predictive analytics and comparative effectiveness clinical prediction model database. Circ Cardiovasc Qual Outcomes. 2015;8(4):368–75.
5. Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. BMJ. 2011;343.
6. Abbasi A, Peelen LM, Corpeleijn E, van der Schouw YT, Stolk RP, Spijkerman AM, et al. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. BMJ. 2012;345:e5900.
7. Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. BMJ Open. 2015;5(3):e007825.
8. Gray LJ, Taub NA, Khunti K, Gardiner E, Hiles S, Webb DR, et al. The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. Diabet Med. 2010;27(8):887–95.
9. Rabin BA, Gaglio B, Sanders T, Nekhlyudov L, Dearing JW, Bull S, et al. Predicting cancer prognosis using interactive online tools: a systematic review and implications for cancer care providers. Cancer Epidemiol Biomarkers Prev. 2013;22(10):1645–56.
10. Watson HA, Carter J, Seed PT, Tribe RM, Shennan AH. The QUiPP App: a safe alternative to a treat-all strategy for threatened preterm labor. Ultrasound Obstet Gynecol. 2017;50(3):342–6.
11. Shields BM, McDonald TJ, Ellard S, Campbell MJ, Hyde C, Hattersley AT. The development and validation of a clinical prediction model to determine the probability of MODY in patients with young-onset diabetes. Diabetologia. 2012;55(5):1265–72.
12. D'Agostino RB Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008;117(6):743–53.
13. Hippisley-Cox J, Coupland C, Robson J, Brindle P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. BMJ. 2010;341:c6624.
14. Fong Y, Evans J, Brook D, Kenkre J, Jarvis P, Gower-Thomas K. The Nottingham Prognostic Index: five- and ten-year data for all-cause survival within a screened population. Ann R Coll Surg Engl. 2015;97(2):137–9.
15. Fox KA, Dabbous OH, Goldberg RJ, Pieper KS, Eagle KA, Van de Werf F, et al. Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (GRACE). BMJ. 2006;333(7578):1091.
16. Johnston SC, Rothwell PM, Nguyen-Huynh MN, Giles MF, Elkins JS, Bernstein AL, et al. Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack. Lancet. 2007;369(9558):283–92.
17. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest. 2010;137(2):263–72.
18. Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics: recalibrating expectations. JAMA. 2018;320(1):27–8.
19. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317–8.
20. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15:104–16.
21. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8–17.
22. Shillan D, Sterne JAC, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Crit Care. 2019;23(1):284.
23. Talaei-Khoei A, Wilson JM. Identifying people at risk of developing type 2 diabetes: A comparison of predictive analytics techniques and predictor variables. Int J Med Inform. 2018;119:22–38.
24. van der Ploeg T, Smits M, Dippel DW, Hunink M, Steyerberg EW. Prediction of intracranial findings on CT-scans by alternative modelling techniques. BMC Med Res Methodol. 2011;11(1):143.
25. Casanova R, Saldana S, Chew EY, Danis RP, Greven CM, Ambrosius WT. Application of random forests methods to diabetic retinopathy classification analyses. PLoS One. 2014;9(6):e98587.
26. Casanova R, Saldana S, Simpson SL, Lacy ME, Subauste AR, Blackshear C, et al. Prediction of incident diabetes in the Jackson Heart Study using high-dimensional machine learning. PloS One. 2016;11(10):e0163942-e.
27. Lo-Ciganic W-H, Huang JL, Zhang HH, Weiss JC, Wu Y, Kwoh CK, et al. Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. JAMA Network Open. 2019;2(3):e190968-e.
28. Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley D. Development and validation of an electronic health record–based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. JAMA Network Open. 2018;1(4): e181018-e.
29. Dreiseitl S, Ohno-Machado L, Kittler H, Vinterbo S, Billhardt H, Binder M. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. J Biomed Inform. 2001;34(1):28–36.
30. Harrison RF, Kennedy RL. Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. Ann Emerg Med. 2005;46(5):431–9.
31. Faisal M, Scally A, Howes R, Beatson K, Richardson D, Mohammed MA. A comparison of logistic regression models with alternative machine learning methods to predict the risk of in-hospital mortality in emergency medical admissions via external validation. Health Inform J. 2018;1460458218813600.
32. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the Gusto database. Stat Med. 1998;17(21): 2501–8.
33. Hsieh MH, Sun L-M, Lin C-L, Hsieh M-J, Hsu C-Y, Kao C-H. Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. Cancer Manag Res. 2018;10:6317–24.
34. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. JAMA Cardiol. 2017;2(2):204–9.
35. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019; 110:12–22.
36. DiabetesGenes.org. Diabetes alliance for research in England (DARE) [Cited 15/11/2018]. Available from: https://www.diabetesgenes.org/current-research/dare/.
37. ClinicalTrials.gov. RetroMASTER - Retrospective Cohort MRC ABPI STratification and Extreme Response Mechanism in Diabetes [Cited 15/11/2018]. Available from: https://www.clinicaltrials.gov/ct2/show/NCT02109978.
38. ClinicalTrials.gov. MASTERMIND - Understanding individual variation in treatment response in type 2 diabetes (Mastermind) [Cited 31/07/2018]. Available from: https://www.clinicaltrials.gov/ct2/show/NCT01847144?term=mastermind.
39. clinicaltrials.gov. PROMASTER - PROspective Cohort MRC ABPI STratification and Extreme Response Mechanism in Diabetes (PROMASTER) [Cited 31/07/2018]. Available from: https://www.clinicaltrials.gov/ct2/show/NCT02105792?term=promaster&rank=1.
40. Thanabalasingham G, Pal A, Selwood MP, Dudley C, Fisher K, Bingley PJ, et al. Systematic assessment of etiology in adults with a clinical diagnosis of young-onset type 2 diabetes is a successful strategy for identifying maturity-onset diabetes of the Young. Diabet Care. 2012;35(6):1206–12.
41. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer New York Inc.; 2001.
42. National Institute for Health and Care Excellence. Type 1 diabetes in adults: diagnosis and management (NICE guideline NG17) 2015 [Cited 14/08/2018]. Available from: https://www.nice.org.uk/guidance/ng17.
43. Setiono R, Hui LCK. Use of a quasi-Newton method in a feedforward neural network construction algorithm. IEEE Trans Neural Netw. 1995;6(1):273–7.
44. Menard SW. Applied logistic regression analysis. Thousand Oaks: Sage Publications; 1995.

45. van Houwelingen JC, le Cessie S. Logistic Regression, a review. Statistica Neerlandica. 1988;42(4):215–32.
46. Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making. 2001;21(1):45–56.
47. Breiman L. Random forests. Machine Learning. 2001;45(1):5–32.
48. Ho TK, editor. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition; 1995 14-16 Aug. New York: IEEE Computer society press; 1995. p. 278–82.
49. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189–232.
50. Ridgeway G. Generalized boosted models: a guide to the gbm package. 2007(21/06/2019).
51. Friedman JH. Multivariate adaptive regression splines. Ann Stat. 1991;19(1): 1–67.
52. Goodfellow I, Bengio Y, Courville A. Deep learning: the MIT press; 2016. p. 800.
53. Ripley BD. Pattern Recognition and Neural Networks. New York: Cambridge University Press; 1996.
54. Hertz J, Krogh A, Palmer R. Introduction to the theory of neural computation. Redwood City: Addison-Wesley; 1991.
55. Bishop C. Neural networks for pattern recognition. New York: Oxford University Press; 1995.
56. Kotsiantis S, Zaharakis I, Pintelas P. Supervised machine learning: a review of classification techniques. Informatica. 2007;31:249–68.
57. Dasarathy B. Nearest neighbor: pattern classification techniques. Los Alamitos: IEEE Computer Society Press; 1991.
58. Vapnik VN. The nature of statistical learning theory: Springer-Verlag; 1995. p. 188.
59. Moguerza JM, Munoz A. Support vector machines with applications. Statist Sci. 2006;21(3):322–36.
60. Claesen M, Moor BD. Hyperparameter search in machine learning: MIC 2015: The XI Metaheuristics International Conference; 2015.
61. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. 1997;30(7):1145–59.
62. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–74.
63. Zhang Z, Rousson V, Lee W-C, Ferdynus C, Chen M, Qian X, et al. Decision curve analysis: a technical note. Ann Transl Med. 2018;6(15).
64. Greenwell B, Boehmke B, Cunningham J, Developers G. gbm: Generalized Boosted Regression Models 2018 [Available from: https://CRAN.R-project.org/package=gbm.
65. Meyer D, Dimitriadou E, Hornik J, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien 2018 [Available from: https://CRAN.R-project.org/package=e1071.
66. Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth ed. New York: Springer; 2002.
67. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.
68. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Software. 2008;28(5):1–26.
69. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. 2014;14(1):137.
70. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer.
71. Wolpert DH, Macready WG. No free lunch theorems for optimization. IEEE Transact Evol Comput. 1997;1(1):67–82.
72. Fernandez-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res. 2014;15(1):3133–81.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.