Journal of Clinical Epidemiology 133 (2021) 43−52

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Logistic regression and machine learning predicted patient mortality from large sets of diagnosis codes comparably

Thomas E. Cowling[a,b,*], David A. Cromwell[a,b], Alexis Bellot[c,d], Linda D. Sharples[e], Jan van der Meulen[a,b]

[a]*Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK*
[b]*Clinical Effectiveness Unit, Royal College of Surgeons of England, Lincoln's Inn Fields, London WC2A 3PE, UK*
[c]*Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, UK*
[d]*Alan Turing Institute, 96 Euston Road, London NW1 2DB, UK*
[e]*Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK*

Accepted 15 December 2020; Published online xxx

Abstract

Objective: The objective of the study was to compare the performance of logistic regression and boosted trees for predicting patient mortality from large sets of diagnosis codes in electronic healthcare records.

Study Design and Setting: We analyzed national hospital records and official death records for patients with myocardial infarction (n = 200,119), hip fracture (n = 169,646), or colorectal cancer surgery (n = 56,515) in England in 2015−2017. One-year mortality was predicted from patient age, sex, and socioeconomic status, and 202 to 257 International Classification of Diseases 10th Revision codes recorded in the preceding year or not (binary predictors). Performance measures included the c-statistic, scaled Brier score, and several measures of calibration.

Results: One-year mortality was 17.2% (34,520) after myocardial infarction, 27.2% (46,115) after hip fracture, and 9.3% (5,273) after colorectal surgery. Optimism-adjusted c-statistics for the logistic regression models were 0.884 (95% confidence interval [CI]: 0.882, 0.886), 0.798 (0.796, 0.800), and 0.811 (0.805, 0.817). The equivalent c-statistics for the boosted tree models were 0.891 (95% CI: 0.889, 0.892), 0.804 (0.802, 0.806), and 0.803 (0.797, 0.809). Model performance was also similar when measured using scaled Brier scores. All models were well calibrated overall.

Conclusion: In large datasets of electronic healthcare records, logistic regression and boosted tree models of numerous diagnosis codes predicted patient mortality comparably.

Keywords: Machine learning; Regression analysis; Big data; Electronic health records; International Classification of Diseases; Comorbidity; Prognosis

Funding: T.E.C. was supported by the Medical Research Council (grant number MR/S020470/1). The funder had no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Conflicts of interest: None declared.

Data statement: The study was exempt from UK National Research Ethics Service (NRES) approval because it involved the analysis of an existing data set of anonymized data. Hospital Episode Statistics (HES) data were made available by NHS Digital (Copyright 2019, reused with the permission of NHS Digital. All rights reserved.) Approvals for the use of anonymized HES data were obtained as part of the standard NHS Digital data access process. The data governance arrangements for the study do not allow us to redistribute HES data to other parties. Researchers interested in accessing HES data can apply for access through NHS Digital's Data Access Request Service (DARS) https://dataaccessrequest.hscic.gov.uk/.

CRediT authorship contribution statement: T.E.C. contributed Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing − Original Draft, Writing − Review & Editing, Visualization, Funding acquisition. D.A.C. contributed Methodology, Resources, Writing − Review & Editing. A.B. contributed Methodology, Writing − Review & Editing. L.D.S. contributed Methodology, Writing − Review & Editing. J.v.d.M. contributed Methodology, Writing − Review & Editing.

* Corresponding author. Tel.: +44 (0)20 7927 2151.
E-mail address: thomas.cowling@lshtm.ac.uk (T.E. Cowling).

https://doi.org/10.1016/j.jclinepi.2020.12.018
0895-4356/

What is new?

Key findings

- Logistic regression and boosted trees predicted 1-year mortality from large sets of diagnosis codes comparably, in three large and diverse clinical populations.

What this adds to what was known?

- Machine learning approaches have been used to model interactions between many diagnosis codes in large datasets of electronic healthcare records.
- No previous studies have directly compared regression and machine learning approaches for modeling large sets of individual International Classification of Diseases (ICD) codes.

What is the implication and what should change now?

- Our results suggest that there is little or no advantage to using machine learning rather than regression approaches in this particular study context.

1. Introduction

Machine learning has received increasing interest from epidemiologists, clinicians, and health service researchers in recent years [1−3]. Related methods have been applied to various types of data, including gene sequences, medical images, and electronic healthcare records [4−6]. Although some commentators have emphasized the promise of these methods [7,8], others have focused on associated challenges [9,10].

One area where the value of machine learning is particularly unclear is clinical prediction modeling [11−13]. Prediction models can be used to inform clinical decisions and the design of preventive interventions, and they can also contribute to risk adjustment and causal inference methods [14,15]. Predicting future events is a traditional focus of machine learning methods, which typically estimate relationships between variables more flexibly than conventional regression [16]. Although this may reduce bias in predictions, it could also increase the risk of modeling associations in the data that exist only by chance such that a model's predictions do not work well for future patients (overfitting) [11].

Electronic healthcare records offer growing opportunities to develop prediction models using machine learning, as large populations can often be studied using these records and larger samples reduce the risk of model overfitting [11,17]. Several models have been developed with related methods and large datasets of electronic healthcare records [18−22]. These models often include variables for hundreds of diagnosis codes to better capture the complexities of patient morbidity, including potential interactions across many conditions that may be best modeled by flexible methods [23,24]. Regression models with many additive coefficients may be liable to predict some values that are too extreme.

However, it is often unclear how conventional regression methods would have performed if directly compared with the machine learning methods used in these studies. A recent systematic review [25] of prognostic modeling studies that compared logistic regression and machine learning methods was limited by the small sample sizes and few predictor variables used in these studies. The review recommended that future research should examine the specific study contexts in which different approaches are suitable, particularly using large datasets and more predictors [25].

In this study, we compared the performance of logistic regression and boosted tree models for predicting patient outcomes from large sets of diagnosis codes given in electronic healthcare records. Such models have been used to measure patient comorbidity and to adjust measures of healthcare quality for patient case-mix, for example [23,26]. To do this, we analyzed linked national datasets of routinely collected hospital data and official death records from England.

The study populations were patients admitted for acute myocardial infarction, hip fracture, or major surgery for colorectal cancer. We chose these populations partly because they represent many admissions, thus providing relevance to a wide audience and allowing robust internal validation of the models. These populations also vary in terms of clinical specialty, coexisting conditions, and mortality, which helped to assess the consistency of results across diverse groups.

We focused on boosted trees as the machine learning approach because they are often used for prediction modeling in large routinely collected healthcare datasets [6,22,27], they are well-established as a leading approach to tabular data in machine learning competitions [28], and they can be used widely without advanced computing facilities because of quick fitting procedures [29].

2. Methods

2.1. Study populations

We analyzed Hospital Episode Statistics Admitted Patient Care data—administrative data for all inpatient hospital care funded by the National Health Service (NHS) in England [30]. Each record relates to an 'episode' of care under the same senior clinician and contains 20 fields for the International Classification of Diseases 10th Revision (ICD-10) codes [31] relevant to that episode. The first field contains the primary diagnosis—the main condition treated.

Patients with myocardial infarction (I21-22 [32,33]) and hip fracture (S72.0-S72.2 [34,35]) were identified from ICD-10 codes recorded as the primary diagnosis in the first episode of each admission. Patients who underwent colorectal surgery were identified from any episode with both a relevant primary diagnosis (ICD-10: C18-20) and main procedure (OPCS-4: H04-11, H29, H33, X14) [36−39].

We included patients aged 18 years or older or, for hip fracture, only patients aged 60 years or older [35] whose admission was from January 1, 2015 to December 31, 2017. If a patient had two or more admissions for the same index condition in this period (myocardial infarction, hip fracture, or colorectal surgery), only the first was included in the analysis.

2.2. Outcome

The outcome was death up to and including 365 days after the date of admission or, for colorectal surgery, the date of procedure. Mortality is the outcome most often used to assess models of diagnosis codes in hospital settings and to develop prediction models using electronic healthcare records [17,24,40]. We analyzed 365-day mortality, rather than in-hospital or 30-day mortality for example, to increase the effective sample size (which is related to the number of outcome events [41]).

We used dates of death recorded in the Office for National Statistics mortality data [42] up to 31 December 2018, providing complete follow-up for the outcome. These official records were linked to Hospital Episode Statistics based on each patient's unique NHS identifier, date of birth, sex, and postcode [43].

2.3. Predictors

We defined a binary predictor for each ICD-10 code that denoted whether it was recorded or not in each patient's index episode or up to 365 days before. We analyzed the first three characters of these codes (excluding fourth characters) as coding choices at this level will be less variable than with four characters [23]. The first three characters define single conditions or other health-related attributes; fourth characters define sites, subtypes, and causes [44].

In each population, we excluded three-character codes recorded for less than 0.5% of patients in the 365-day 'look-back period' as these codes were so rare that they were unlikely to improve model performance [6,26,45]. We used a 365-day period, rather than only using codes from the index episode, as this improved model performance in some published studies [24].

Patient age, sex, and socioeconomic status were also included as predictors, as is common when examining models of ICD codes [24,40]. Socioeconomic status was measured by the national Index of Multiple Deprivation rank of each residential area (with 1,000 to 3,000 residents in each of 32,482 areas) [46]; we excluded patients with missing data for this variable (1.2%; 5,346/431,626).

2.4. Model estimation

We first estimated associations between the outcome and predictors (age, sex, socioeconomic status, and ICD codes) as the maximum likelihood estimates of a logistic regression model. We did not fit nonlinear associations for age or socioeconomic status or use penalized estimation, as these choices had minimal effects on model performance in our previous analysis of the same data [47].

We used the XGBoost [29] algorithm to develop gradient boosted tree models [48−50], using all predictors as before. This algorithm fits a series of tree models to the data sequentially with each tree attempting to improve on predictions from the previous tree [51]. These models fit many interactions between predictors without these terms having to be prespecified (unlike in conventional regression).

Five boosted tree models were fitted in each population using 100, 200, 300, 400, and 500 boosting iterations. Further tuning parameters were held fixed as various combinations of these parameters gave similar maximum performance across this range of boosting iterations (see Appendix A1). The learning rate, maximum tree depth, minimum node weight, and subsample fraction took the values of 0.1, 5, 100, and 1, respectively (see Appendix A1 for definitions).

2.5. Model performance

Overall model performance was measured using Brier scores [52]. These scores equaled the mean of squared differences between predicted probabilities of death and observed outcomes. We scaled these scores from 0% to 100% (0% for a noninformative model and 100% if perfect) [53].

To assess discrimination, we calculated the *c*-statistic. This equaled the probability that a randomly chosen patient who died had a greater predicted probability of death than a randomly chosen patient who did not [54]. The *c*-statistic equals one for perfect models and 0.5 for predictions made at random.

To assess calibration, we calculated the integrated calibration index (ICI) [55], calibration-in-the-large, and calibration slopes [56]. The ICI and calibration-in-the-large assess the calibration of model predictions across their range and overall, respectively; perfect models have values of zero. Calibration slopes equal one in perfect models, with smaller values indicating overfitting.

For each model in each population, we first calculated the aforementioned measures in the original data used to fit the models (apparent performance). We then repeated all modeling steps in each of 250 bootstrap samples and, for each sample, calculated the performance of the resulting

model in this sample and the original data; the difference in performance values between the bootstrap sample and original data defined the 'optimism.' Finally, an optimism-adjusted value of each performance measure was calculated as the apparent performance value minus the mean optimism [54,57,58]. This is the bootstrap validation approach given in the TRIPOD guidelines [59].

2.6. Secondary analyses

We conducted a secondary analysis using a 1,825-day (5-year) look-back period. This analysis also accounted for the exact number of days since each ICD-10 code was last recorded rather than just whether it was recorded or not in a given time period (see Appendix A2 for details). This analysis, in addition to the main analysis, was prespecified in a published protocol [60]. We have previously reported a separate study that was specified in the same protocol [47].

We conducted two post hoc analyses (also described in Appendix A2). In the first analysis, we examined whether the calibration of the logistic regression models at high predicted probabilities could be improved. We used splines to fit nonlinear associations for age and socioeconomic status and included interactions between three selected predictors. In the second analysis, we assessed the performance of two additional machine learning approaches—random forests and neural networks. Data preparation was performed using Stata (v15). R (v3.5) was used for all analysis; code to implement the different estimation methods is given in Appendix A3.

In response to a peer reviewer's suggestion, we conducted two additional analyses. First, we added 500 extra boosting iterations (1,000 in total) and used other combinations of tuning parameters to see if this improved the boosted trees' performance. Second, we examined the performance of the regression and boosted tree models when only ICD codes with frequencies less than 0.1% (rather than 0.5%) were excluded from the set of predictor variables.

3. Results

The percentage of patients who died within 1 year was 17.2% (34,520/200,119) after myocardial infarction, 27.2% (46,115/169,646) after hip fracture, and 9.3% (5,273/56,515) after colorectal surgery. In each population, between 202 and 257 ICD-10 codes were recorded for at least 0.5% of patients within 1 year before their admission or procedure. This provided 168 (34,520/205; myocardial infarction), 177 (46,115/260; hip fracture), and 25 (5,273/212; colorectal surgery) deaths per predictor variable. Most ICD-10 codes had low frequencies (see Table 1).

The distributions of predicted probabilities were similar between the logistic regression and boosted tree models overall (Fig. 1; see Fig. 2 for distributions by outcome). The most 'important' variables were also similar between models (Appendix A4). Age and metastatic cancer in the respiratory and digestive organs were important predictors of death in each population.

The overall optimism-adjusted performance of the boosted trees was slightly better than that of logistic regression, as measured by Brier scores, in the myocardial infarction and hip fracture populations (Table 2). The absolute differences in scaled Brier scores were 1.9% (95% CI: 1.7% to 2.1%) and 1.2% (95% CI: 1.0% to 1.4%), respectively. Logistic regression had a slightly superior score in the colorectal surgery population (difference = 1.5%; 95% CI: 0.8% to 2.1%). Model discrimination, as measured by the *c*-statistic, followed the same pattern with a minimum value of 0.798 (95% CI: 0.796 to 0.800) across models and populations (see Table 2).

Table 1. Descriptive statistics for outcome and predictor variables, by population

	Acute myocardial infarction	Hip fracture	Major colorectal cancer surgery
Number of patients	200,119	169,646	56,515
Number of patients who died within 1 yr (%)	34,520 (17.2)	46,115 (27.2)	5,273 (9.3)
Patient characteristics			
Median age (IQR)	70 (58 to 80)	84 (77 to 89)	70 (62 to 78)
Male (vs. female) (%)	132,162 (66.0)	48,622 (28.7)	32,004 (56.6)
Median socioeconomic status (IQR)[a]	4.8 (2.4 to 7.3)	5.4 (2.9 to 7.7)	5.7 (3.3 to 7.9)
ICD-10 codes			
Number of codes included[b]	202	257	209
Median frequency (%) of codes (IQR)	1.6 (0.8 to 3.4)	1.8 (0.8 to 4.2)	1.6 (0.9 to 4.5)
Median number of codes per patient (IQR)	6 (4 to 10)	9 (6 to 14)	7 (4 to 11)
Median agreement between codes (IQR)[c]	0.01 (0.00 to 0.02)	0.01 (0.00 to 0.01)	0.01 (0.00 to 0.01)

Abbreviations: IQR, interquartile range; ICD-10, International Classification of Diseases 10th Revision.
[a] Scaled such that the most deprived area of residence nationally had a value of 0 and the least deprived area had a value of 10.
[b] Relative frequency of each three-character code was at least 0.5% in the given population.
[c] Median values of Cohen's kappa coefficient across all unique pairs of codes (1 = perfect agreement, 0 = chance agreement).

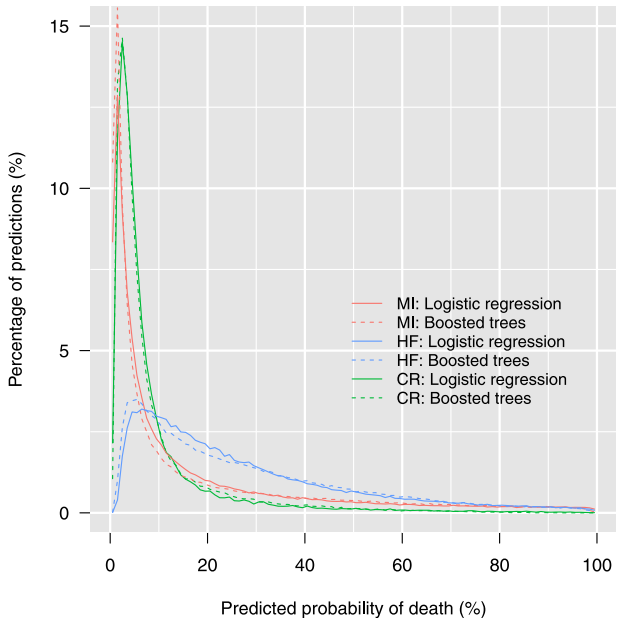

Fig. 1. Frequency distributions of predicted probabilities of death, by population and method. In the MI population, 5% of patients had predicted probabilities equal to or greater than 72.5%. The corresponding values in the HF and CR populations were 73.9% and 35.7%, respectively. MI, myocardial infarction; HF, hip fracture; CR, colorectal surgery.

Both the boosted trees and regression models were well calibrated overall. Values of calibration-in-the-large and calibration slopes were close to their respective ideal values of 0 and 1 (Table 2).

However, logistic regression predictions of very high probabilities of death were too high on average, particularly in the colorectal surgery population (see calibration plots in Fig. 3). In contrast, the predictions of the boosted trees closely agreed with observed outcomes across the range of predicted probabilities. Several ICD-10 codes were frequent among patients with very high predicted risks of death, and these codes were almost identical for the boosted trees and regression models (see Appendix A5 for code frequencies in the top 5% of predicted risks). The inclusion of splines and interactions between selected codes in the logistic regression models did not correct for the worse calibration observed at high predicted risks in each population (Appendix A6).

For the boosted tree models, the maximum scaled Brier scores were attained with 500 boosting iterations in the myocardial infarction and hip fracture populations and 200 iterations in the colorectal surgery population (Appendix A7). These numbers of iterations also provided the models whose calibration slopes were closest to 1 (the ideal value). The differences between apparent and optimism-adjusted performance (optimism) were typically small for the boosted tree models, but the corresponding differences for logistic regression were even smaller (Appendix A7).

The models estimated in the secondary analysis using a 5-year look-back period generally performed similarly to or not as well as those from the main analysis (Appendix A8). The random forest models did not attain scaled Brier scores or *c*-statistics that were greater than those for both the logistic regression and boosted tree models in any of the populations, whereas the neural networks were the worst-performing models in each population (see Appendix A8 for results). Using up to 1,000 boosting iterations for the boosted tree models and other combinations of tuning parameters did not improve prediction performance, neither did using a 0.1% (vs. 0.5%) frequency threshold for including ICD codes as predictors (Appendix A9).

4. Discussion

In large datasets of electronic healthcare records, logistic regression and boosted tree models of numerous diagnosis codes predicted 1-year mortality comparably. This was consistent across the three populations of patients with acute myocardial infarction and hip fracture and those who underwent colorectal surgery. Both the logistic regression and boosted tree models had good discrimination and were well calibrated overall, although the boosted trees were better calibrated at high predicted probabilities of death.

4.1. Interpretation of results

A potential strength of boosted trees is that they include many interactions between predictors by design. Interactions across many conditions were plausible, given relationships between disorders and their management. Several authors have advocated modeling interactions between conditions for this reason [23,24,61]. However, the boosted trees performed comparably to logistic regression models without interactions, suggesting that interactions were unimportant overall in this context.

This finding may be partly explained by the low frequencies of most ICD codes. Two codes may not be recorded together very often, which reduces the potential for their interaction to improve overall model performance, even if the interaction has a large true prognostic effect. It may also be difficult to reliably estimate interactions between codes that are not often recorded together.

Clinical prediction problems have been described as having unfavorable 'signal-to-noise' ratios that question the potential benefits of using more flexible estimation methods that fit many interactions [62]. Misclassification error in the recording of diagnosis codes may add to the 'noise' and result in biased estimates of true interactions. In addition, more flexible methods may be more likely to capture spurious relationships in the data that have arisen by chance. However, the values of optimism for the boosted trees were reasonably small in the present study, which is

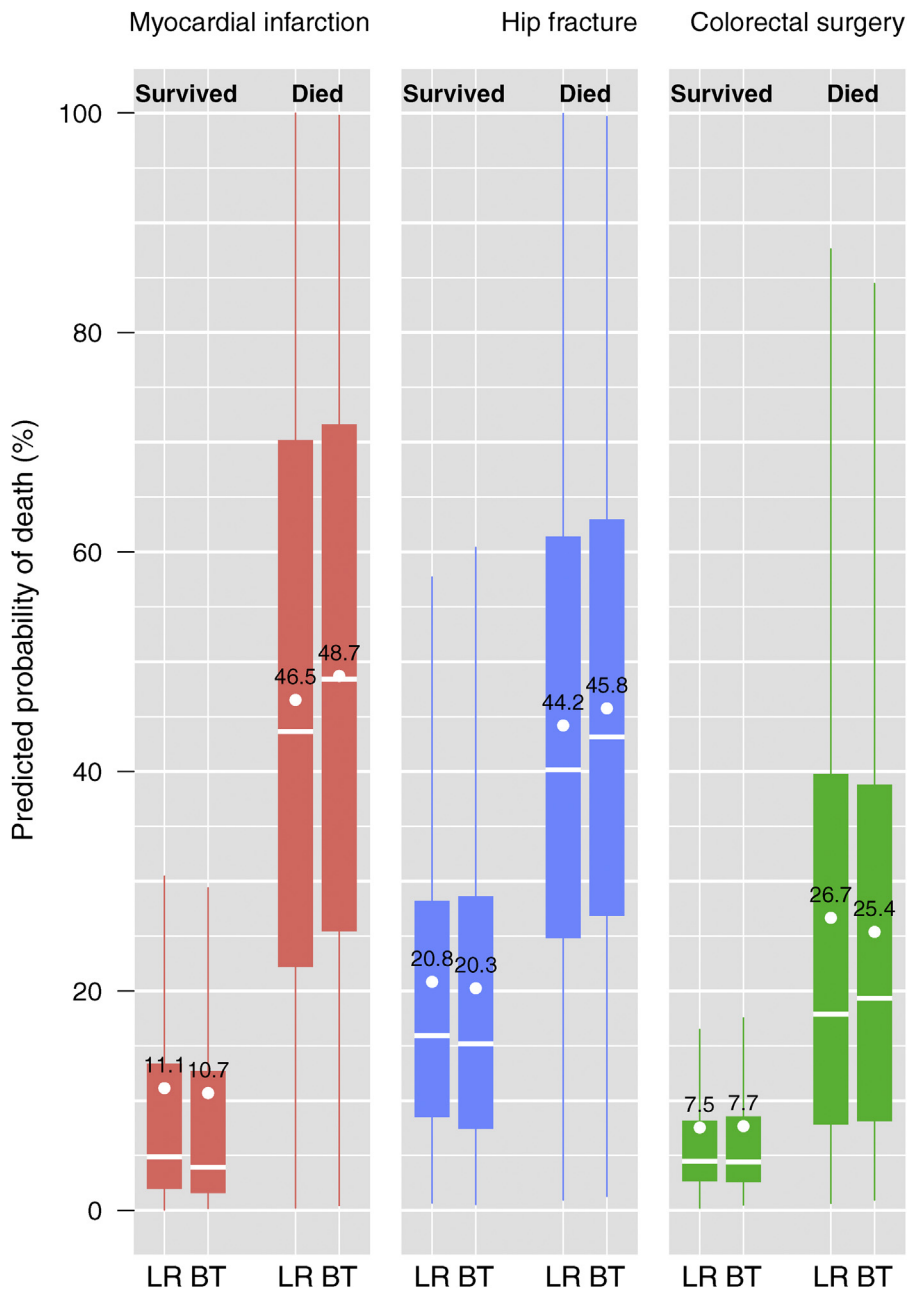

Fig. 2. Frequency distributions of predicted probabilities of death, by population, outcome, and method. Boxes are drawn from the lower to upper quartile of predicted probabilities with a white horizontal line at the median value. Annotated values and white dots correspond to mean values. Whiskers are drawn to the most extreme predicted probabilities that are no more than 1.5 times the interquartile range from the box. LR, logistic regression; BT, boosted trees.

partly explained by the large sample sizes and the shrinkage included in the boosting process to prevent model overfitting.

Larger study populations reduce the potential for overfitting and can thereby improve the performance of more flexible methods [63]. We used 3 years of national data to provide large samples, but many investigators do not have access to such large databases [11]. In smaller populations or when the study outcome occurs less frequently, any benefits of boosted trees over logistic regression in terms of prediction performance are likely to reduce. In addition, important interactions may already be known such that they could be prespecified in regression models.

One benefit of the boosted trees was that very high predicted probabilities were better calibrated than when logistic regression was used. This was not fully explained by the splines or interactions that were added to the regression models, which may be because interactions between many codes needed to be added. Boosted trees fit interactions in each iteration to improve predictions where the existing

Table 2. Prediction performance of the logistic regression and boosted tree models, corrected for optimism using 250 bootstrap samples (with 95% confidence intervals)

	Acute myocardial infarction	Hip fracture	Major colorectal cancer surgery
Scaled Brier score (%)			
Logistic regression	34.6 (34.2 to 35.1)	22.8 (22.4 to 23.2)	17.2 (16.1 to 18.2)
Boosted trees	36.5 (36.1 to 37.0)	24.0 (23.6 to 24.4)	15.7 (14.8 to 16.6)
c-Statistic			
Logistic regression	0.884 (0.882 to 0.886)	0.798 (0.796 to 0.800)	0.811 (0.805 to 0.817)
Boosted trees	0.891 (0.889 to 0.892)	0.804 (0.802 to 0.806)	0.803 (0.797 to 0.809)
Calibration-in-the-large			
Logistic regression	−0.001 (−0.017 to 0.015)	0.000 (−0.013 to 0.013)	0.000 (−0.032 to 0.031)
Boosted trees	0.000 (−0.016 to 0.016)	0.001 (−0.012 to 0.014)	0.002 (−0.028 to 0.033)
Calibration slope			
Logistic regression	0.993 (0.984 to 1.003)	0.989 (0.977 to 1.002)	0.961 (0.936 to 0.987)
Boosted trees	1.003 (0.993 to 1.013)	1.006 (0.993 to 1.018)	0.988 (0.963 to 1.013)
Integrated calibration index			
Logistic regression	0.012 (0.011 to 0.013)	0.015 (0.014 to 0.017)	0.007 (0.006 to 0.009)
Boosted trees	0.002 (0.001 to 0.003)	0.004 (0.002 to 0.006)	0.001 (0.000 to 0.003)

Results for boosted trees correspond to models with 500 boosting iterations in the myocardial infarction and hip fracture populations and 200 iterations in the colorectal surgery population.

model works less well, such as extreme cases. In contrast, logistic regression models may fit well overall but are not designed to capture unusual cases with very high risks of death because the many patients at low risk dominate model estimates. However, interactions fitted by boosted trees may not be generalizable to other datasets, which could reduce this benefit.

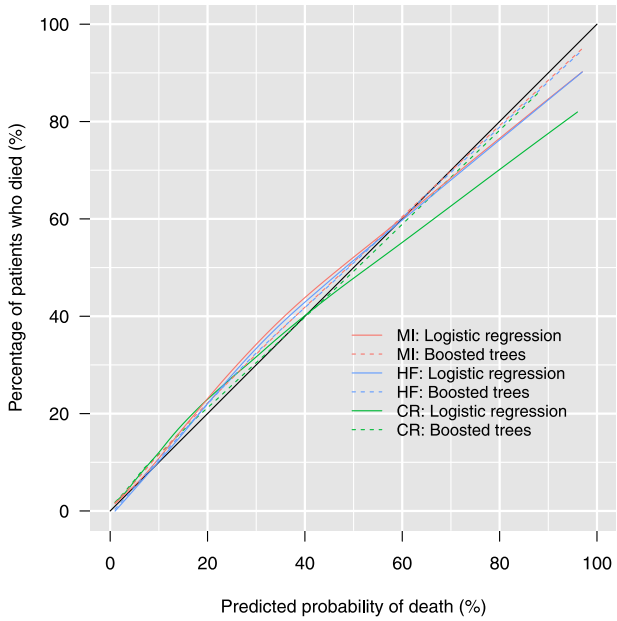

Fig. 3. Calibration plots for the logistic regression and boosted tree models, by population, corrected for optimism using 250 bootstrap samples (shown with the line of perfect calibration). In the MI and HF populations, 3.5% of predicted probabilities were equal to or greater than 80%. In the CR population, 2.8% of predicted probabilities were equal to or greater than 50%. The black 45° line represents perfect calibration. MI, myocardial infarction; HF, hip fracture; CR, colorectal surgery.

4.2. Relation to existing literature

To our knowledge, no previous studies have directly compared regression and machine learning approaches for modeling large sets of individual ICD codes specifically. In a previous study of Hospital Episode Statistics data (up to 2013), logistic regression models had similar discrimination to support vector machines, neural networks, and random forests when predicting in-hospital mortality using small sets of comorbidities [64]. Using the same datasets as in the present study, we have previously found that large sets of individual ICD codes can predict patient outcomes better than traditional sets of comorbidities [47], which is consistent with other studies [23,26,65].

Many analyses have compared logistic regression with boosted trees and other machine learning approaches in various large datasets of electronic healthcare records, with differing results [22,27,62,66,67]. Two studies [22,27] in which boosted trees performed better than regression analyzed large primary care datasets, which may suggest that boosted trees have an advantage in very heterogeneous populations. This contrasts to our analysis which was performed within populations defined by an index condition. It is difficult to draw general conclusions from such studies, as results may be sensitive to the specific prediction problem (such as sample size, predictors, and data quality) and the exact implementation of algorithms. One approach will not work best across all contexts [68,69].

A recent systematic review [25] of studies that compared logistic regression and machine learning for clinical

prediction modeling stated that 'Future research should focus more on delineating the type of predictive problems in which various algorithms have maximal value' (p.18). Our study aligns with this call and suggests that logistic regression and boosted trees predict patient mortality comparably from numerous diagnosis codes in large electronic healthcare datasets.

4.3. Limitations of the study

Our study focused on diagnosis codes, given their central role in analyzing patient morbidity using electronic healthcare records. In addition, the ICD-10 coding system has a standardized core format internationally, which may improve the generalizability of our results to other countries. Future work could include other predictors that are likely to have strong effects but may be recorded variably or not at all in the datasets of different countries, such as the hospitalization pathway. Some variables modeled in other studies using boosted trees, including laboratory test values and prescription information [21,27], are not recorded in Hospital Episode Statistics data.

Future research should conduct similar comparisons for other populations, outcomes, and datasets to see whether our results apply more generally. For example, in study populations without a defined index condition, interactions between primary and secondary diagnosis codes may improve prediction performance. In large datasets with greater frequencies of ICD codes, possibly in older populations, interactions between codes may be estimated with greater precision. The external validity of models produced using regression and machine learning approaches should also be compared when investigators intend to use the models in another data set or context.

4.4. Implications for research

Many studies use diagnosis codes from electronic healthcare records to model patient morbidity [70]. Our results suggest that there is little or no advantage to using machine learning rather than regression approaches in the particular context examined. Investigators may prefer to use regression instead if they require a model that is transparent, easily interpreted, and familiar to a wide audience. We have previously reported a regression-based approach for selecting small sets of ICD codes with high prediction performance [47].

Electronic healthcare records are increasing in volume and scope, presenting growing opportunities to use large sets of predictors and model their relationships with more flexible methods [17]. High-quality comparisons in large datasets are required to determine the contexts in which these methods should be used and when more conventional approaches are sufficient [25]. In the context of the study presented here, our results suggest that regression approaches perform well.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.12.018.

References

[1] Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319:1317−8.
[2] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380:1347−58.
[3] Rose S. Intersections of machine learning and epidemiological methods for health services research. Int J Epidemiol 2020. https://doi.org/10.1093/ije/dyaa035.
[4] Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet 2019;51:12−8.
[5] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402−10.
[6] Einav L, Finkelstein A, Mullainathan S, Obermeyer Z. Predictive modeling of U.S. health care spending in late life. Science 2018;360:1462−5.
[7] Obermeyer Z, Emanuel EJ. Predicting the future - Big data, machine learning, and clinical medicine. N Engl J Med 2016;375:1216−9.
[8] Hinton G. Deep learning-A technology with the potential to transform health care. JAMA 2018;320:1101−2.
[9] Chen JH, Asch SM. Machine learning and prediction in medicine - beyond the peak of inflated expectations. N Engl J Med 2017;376:2507−9.
[10] Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017;318:517−8.
[11] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. 2nd ed. Cham: Springer; 2019.
[12] Van Calster B, Wynants L. Machine learning in medicine. N Engl J Med 2019;380:2588.
[13] Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393:1577−9.
[14] Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10:e1001381.
[15] Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods: when worlds collide-prediction, machine learning and causal inference. Int J Epidemiol 2019. https://doi.org/10.1093/ije/dyz132.
[16] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009.
[17] Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017;24:198−208.
[18] Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. BMC Med Inform Decis Mak 2018;18:122.
[19] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digital Med 2018;1:18.
[20] Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PLoS One 2018;13:e0202344.
[21] Elfiky AA, Pany MJ, Parikh RB, Obermeyer Z. Development and application of a machine learning approach to assess short-term

mortality risk among patients with cancer starting chemotherapy. JAMA Netw Open 2018;1:e180926.
[22] Jung K, Sudat SEK, Kwon N, Stewart WF, Shah NH. Predicting need for advanced illness or palliative care in A primary care population using electronic health record data. J Biomed Inform 2019;92:103115.
[23] Holman CD, Preen DB, Baynham NJ, Finn JC, Semmens JB. A multipurpose comorbidity scoring system performed better than the Charlson index. J Clin Epidemiol 2005;58:1006−14.
[24] Sharabiani MT, Aylin P, Bottle A. Systematic review of comorbidity indices for administrative data. Med Care 2012;50:1109−18.
[25] 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.
[26] Krumholz HM, Coppi AC, Warner F, Triche EW, Li SX, Mahajan S, et al. Comparative effectiveness of new approaches to improve mortality risk models from medicare claims data. JAMA Netw Open 2019;2:e197314.
[27] Rahimian F, Salimi-Khorshidi G, Payberah AH, Tran J, Ayala Solares R, Raimondi F, et al. Predicting the risk of emergency admission with machine learning: development and validation using linked electronic health records. PLoS Med 2018;15:e1002695.
[28] Kaggle. What is XGBoost. Available at https://www.kaggle.com/dansbecker/xgboost. Accessed January 19, 2021.
[29] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. arXiv; 2016. https://arxiv.org/abs/1603.02754v3.
[30] Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data resource profile: hospital episode statistics admitted patient care (HES APC). Int J Epidemiol 2017;46:1093.
[31] World Health Organization. International Statistical Classification of Diseases and Related Health Problems - 10th revision. 5th ed. 2016. Available at https://icd.who.int/browse10/Content/statichtml/ICD10Volume2_en_2016.pdf. Accessed January 19, 2021.
[32] Metcalfe A, Neudam A, Forde S, Liu M, Drosler S, Quan H, et al. Case definitions for acute myocardial infarction in administrative databases and their impact on in-hospital mortality rates. Health Serv Res 2013;48:290−318.
[33] McCormick N, Lacaille D, Bhole V, Avina-Zubieta JA. Validity of myocardial infarction diagnoses in administrative databases: a systematic review. PLoS One 2014;9:e92286.
[34] Toson B, Harvey LA, Close JC. The ICD-10 Charlson Comorbidity Index predicted mortality but not resource utilization following hip fracture. J Clin Epidemiol 2015;68:44−51.
[35] Royal College of Physicians. National Hip Fracture Database (NHFD) annual report 2016 2016. Available at https://www.nhfd.co.uk/report2016. Accessed January 19, 2021.
[36] Burns EM, Bottle A, Aylin P, Darzi A, Nicholls RJ, Faiz O. Variation in reoperation after colorectal surgery in England as an indicator of surgical performance: retrospective analysis of Hospital Episode Statistics. BMJ 2011;343:d4836.
[37] Byrne BE, Mamidanna R, Vincent CA, Faiz O. Population-based cohort study comparing 30- and 90-day institutional mortality rates after colorectal surgery. Br J Surg 2013;100:1810−7.
[38] Morris EJ, Taylor EF, Thomas JD, Quirke P, Finan PJ, Coleman MP, et al. Thirty-day postoperative mortality after colorectal cancer surgery in England. Gut 2011;60:806−13.
[39] Redaniel MT, Martin RM, Blazeby JM, Wade J, Jeffreys M. The association of time between diagnosis and major resection with poorer colorectal cancer survival: a retrospective cohort study. BMC Cancer 2014;14:642.
[40] Yurkovich M, Avina-Zubieta JA, Thomas J, Gorenchtein M, Lacaille D. A systematic review identifies valid comorbidity indices derived from administrative health data. J Clin Epidemiol 2015;68:3−14.
[41] Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 2019;38:1276−96.
[42] Office for National Statistics. Deaths. Available at https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths. Accessed January 19, 2021.
[43] NHS Digital. A guide to linked mortality data from hospital episode statistics and the Office for national statistics. Available at https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/linked-hes-ons-mortality-data. Accessed January 19, 2021.
[44] World Health Organization. Classification of Diseases (ICD). Available at https://www.who.int/classifications/icd/en/. Accessed January 19, 2021.
[45] Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Med Res Methodol 2012;12:82.
[46] Ministry of Housing, Communities & Local Government. English indices of deprivation. Available at https://www.gov.uk/government/collections/english-indices-of-deprivation. Accessed January 19, 2021.
[47] Cowling TE, Cromwell DA, Sharples LD, van der Meulen J. A novel approach selected small sets of diagnosis codes with high prediction performance in large healthcare datasets. J Clin Epidemiol 2020;128:20−8.
[48] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189−232.
[49] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). Ann Stat 2000;28:337−407.
[50] Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal 2002;38:367−78.
[51] Chen T, He T, Benesty M. XGBoost R tutorial. Available at https://cran.r-project.org/web/packages/xgboost/vignettes/xgboostPresentation.html. Accessed January 19, 2021.
[52] Brier GW. Verification of forecasts expressed in terms of probability. Monthly Weather Rev 1950;78:1−3.
[53] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128−38.
[54] Harrell FE Jr. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ed. Cham: Springer; 2015.
[55] Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med 2019;38:4051−65.
[56] Cox DR. Two further applications of a model for binary regression. Biometrika 1958;45:562−5.
[57] Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774−81.
[58] Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
[59] Moons KG, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1−73.
[60] Cowling TE, Cromwell DA, Sharples LD, van der Meulen J. Protocol for an observational study evaluating new approaches to modelling diagnostic information from large administrative hospital datasets. medRxiv 201919011338.
[61] Romano PS, Roos LL, Jollis JG. Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives. J Clin Epidemiol 1993;46:1075−9.
[62] Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the Gusto database. Stat Med 1998;17:2501−8.

[63] 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;14:137.

[64] Bottle A, Gaudoin R, Goudie R, Jones S, Aylin P. Can valid and practical risk-prediction or casemix adjustment models, including adjustment for comorbidity, be generated from English hospital administrative data (Hospital Episode Statistics)? A national observational study. Southampton, UK: NIHR Journals Library; 2014.

[65] Stanley J, Sarfati D. The new measuring multimorbidity index predicted mortality better than Charlson and Elixhauser indices among the general population. J Clin Epidemiol 2017;92:99−110.

[66] Austin PC, Lee DS, Steyerberg EW, Tu JV. Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods? Biom J 2012;54:657−73.

[67] Gravesteijn BY, Nieboer D, Ercole A, Lingsma HF, Nelson D, van Calster B, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. J Clin Epidemiol 2020;122:95−107.

[68] Wolpert DH. The lack of A priori distinctions between learning algorithms. Neural Comput 1996;8:1341−90.

[69] Couronne R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics 2018;19:270.

[70] Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 2005;43:1130−9.