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 St, 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 St, London, WC1E 7HT, UK

*Corresponding author. London School of Hygiene and Tropical Medicine, Keppel St, London, WC1E 7HT, UK. Tel: +44 (0)20 7927 2151. E-mail: thomas.cowling@lshtm.ac.uk

Abstract

Objective: 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 analysed 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 2015-17. 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% 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.

Key words: Machine learning, regression analysis, big data, electronic health records, International Classification of Diseases, comorbidity, prognosis

Running title: Logistic regression, machine learning, and large sets of diagnosis codes

Word count: 3391

What is new?

Key findings

- Logistic regression and boosted trees predicted one-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 modelling large sets of individual International Classification of Diseases (ICD) codes

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 services 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] While some commentators have emphasised 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 modelling.[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] While this may reduce bias in predictions, it could also increase the risk of modelling 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 modelled 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 to the machine learning methods used in these studies. A recent systematic review[25] of prognostic modelling 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 analysed 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, co-existing 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 modelling 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 due to quick fitting procedures.[29]

2. Methods

2.1 Study populations

We analysed 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 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.

Myocardial infarction (I21-22[32,33]) and hip fracture (S72.0-S72.2[34,35]) patients were identified from ICD-10 codes recorded as the primary diagnosis in the first episode of each admission. Colorectal surgery patients 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 1 January 2015 to 31 December 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 analysed 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 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 analysed 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 1000 to 3000 residents in each of 32 482 areas)[46]; we excluded patients with missing data for this variable (1.2%; 5346/431 626).

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 non-linear associations for age or socioeconomic status or use penalised 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 pre-specified (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 equalled the mean of squared differences between predicted probabilities of death and observed outcomes. We scaled these scores from 0–100% (0% for a non-informative model and 100% if perfect).[53]

To assess discrimination, we calculated the *c*-statistic. This equalled 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] 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 above measures in the original data used to fit the models ('apparent performance'). We then repeated all modelling 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 1825-day (five-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 pre-specified 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 non-linear 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 done 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 (1000 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 one year was 17.2% (34 520/200 119) after myocardial infarction, 27.2% (46 115/169 646) after hip fracture, and 9.3% (5273/56 515) after colorectal surgery. In each population, between 202 and 257 ICD-10 codes were recorded for at least 0.5% of patients within one year before their admission or procedure. This provided 168 (34 520/205; myocardial infarction), 177 (46 115/260; hip fracture), and 25 (5273/212; 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 (Figure 1; see Figure 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).

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 Figure 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 amongst 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 five-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, while the neural networks were the worst-performing models in each population (see Appendix A8 for results). Using up to 1000 boosting iterations for the boosted tree models and other combinations of tuning parameters did not improve prediction performance, neither did using a 0.1% (versus 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 one-year mortality comparably. This was consistent across the three populations of acute myocardial infarction, hip fracture, and colorectal surgery patients. Both the logistic regression and boosted tree models had good discrimination and were well calibrated overall, though 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 modelling 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 unfavourable ‘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 current study which is 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 three 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 pre-specified 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 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 generalisable to other datasets which could reduce this benefit.

4.2 Relation to existing literature

To our knowledge, no previous studies have directly compared regression and machine learning approaches for modelling 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 current 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 (for example[22,27,62,66,67]). Two studies[22,27] in which boosted trees performed better than regression analysed large primary care datasets, which may suggest that boosted trees have an advantage in very heterogeneous populations. This contrasts to our analysis which was done 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 modelling 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 analysing patient morbidity using electronic healthcare records. In addition, the ICD-10 coding system has a standardised core format internationally which may improve the generalisability 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 hospitalisation pathway. Some variables modelled 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 dataset 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.

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.

CRediT authorship contribution statement

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

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 dataset of anonymised data. Hospital Episode Statistics (HES) data were made available by NHS Digital (Copyright 2019, re-used with the permission of NHS Digital. All rights reserved.) Approvals for the use of anonymised 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/.

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. doi: 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, 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, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10:e1001381. doi: 10.1371/journal.pmed.1001381.

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. doi: 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. doi: 10.1186/s12911-018-0677-8.

19. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *Npj Digital Medicine* 2018;1:18. doi: 10.1038/s41746-018-0029-1.

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. doi: 10.1371/journal.pone.0202344.

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. doi: 10.1001/jamanetworkopen.2018.0926.

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* 2019:103115. doi: 10.1016/j.jbi.2019.103115.

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, et al. Comparative Effectiveness of New Approaches to Improve Mortality Risk Models From Medicare Claims Data. *JAMA Netw Open* 2019;2:e197314. doi: 10.1001/jamanetworkopen.2019.7314.

27. Rahimian F, Salimi-Khorshidi G, Payberah AH, et al. Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLOS Medicine* 2018;15:e1002695. doi: 10.1371/journal.pmed.1002695.

28. Kaggle. *What is XGBoost*. Available from: https://www.kaggle.com/dansbecker/xgboost.

29. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *arXiv* 2016. doi: 10.1145/2939672.2939785.

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-i. doi: 10.1093/ije/dyx015.

31. World Health Organization. *International Statistical Classification of Diseases and Related Health Problems - 10th revision (5th edition)*. 2016. Available from: https://icd.who.int/browse10/Content/statichtml/ICD10Volume2_en_2016.pdf.

32. Metcalfe A, Neudam A, Forde S, 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. doi: 10.1371/journal.pone.0092286.

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 from: https://www.nhfd.co.uk/report2016.

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. doi: 10.1136/bmj.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, 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. doi: 10.1186/1471-2407-14-642.

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, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* 2018. doi: 10.1002/sim.7992.

42. Office for National Statistics. *Deaths*. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths.

43. NHS Digital. *A Guide to Linked Mortality Data from Hospital Episode Statistics and the Office for National Statistics*. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/linked-hes-ons-mortality-data.

44. World Health Organization. *Classification of Diseases (ICD)*. Available from: https://www.who.int/classifications/icd/en/.

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 Medical Research Methodology* 2012;12:82. doi: 10.1186/1471-2288-12-82.

46. Ministry of Housing, Communities & Local Government,. *English indices of deprivation*. Available from: https://www.gov.uk/government/collections/english-indices-of-deprivation.

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. doi: 10.1016/j.jclinepi.2020.08.001.

48. Friedman JH. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 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). *The Annals of Statistics* 2000;28:337-407.

50. Friedman JH. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 2002;38:367-78.

51. Chen T, He T, Benesty M. *XGBoost R Tutorial*. Available from: https://cran.r-project.org/web/packages/xgboost/vignettes/xgboostPresentation.html.

52. Brier GW. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* 1950;78:1-3.

53. Steyerberg EW, Vickers AJ, Cook NR, 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. doi: 10.1002/sim.8281.

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, 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. doi: 10.7326/M14-0698.

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* 2019:19011338. doi: 10.1101/19011338.

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. doi: 10.1186/1471-2288-14-137.

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, 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 Computation* 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. doi: 10.1186/s12859-018-2264-5.

70. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005;43:1130-9.

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 who died within 1 year (%)	34 520 (17.2)	46 115 (27.2)	5273 (9.3)
Patient characteristics			
Median age (IQR)	70 (58 to 80)	84 (77 to 89)	70 (62 to 78)
Male (versus 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)

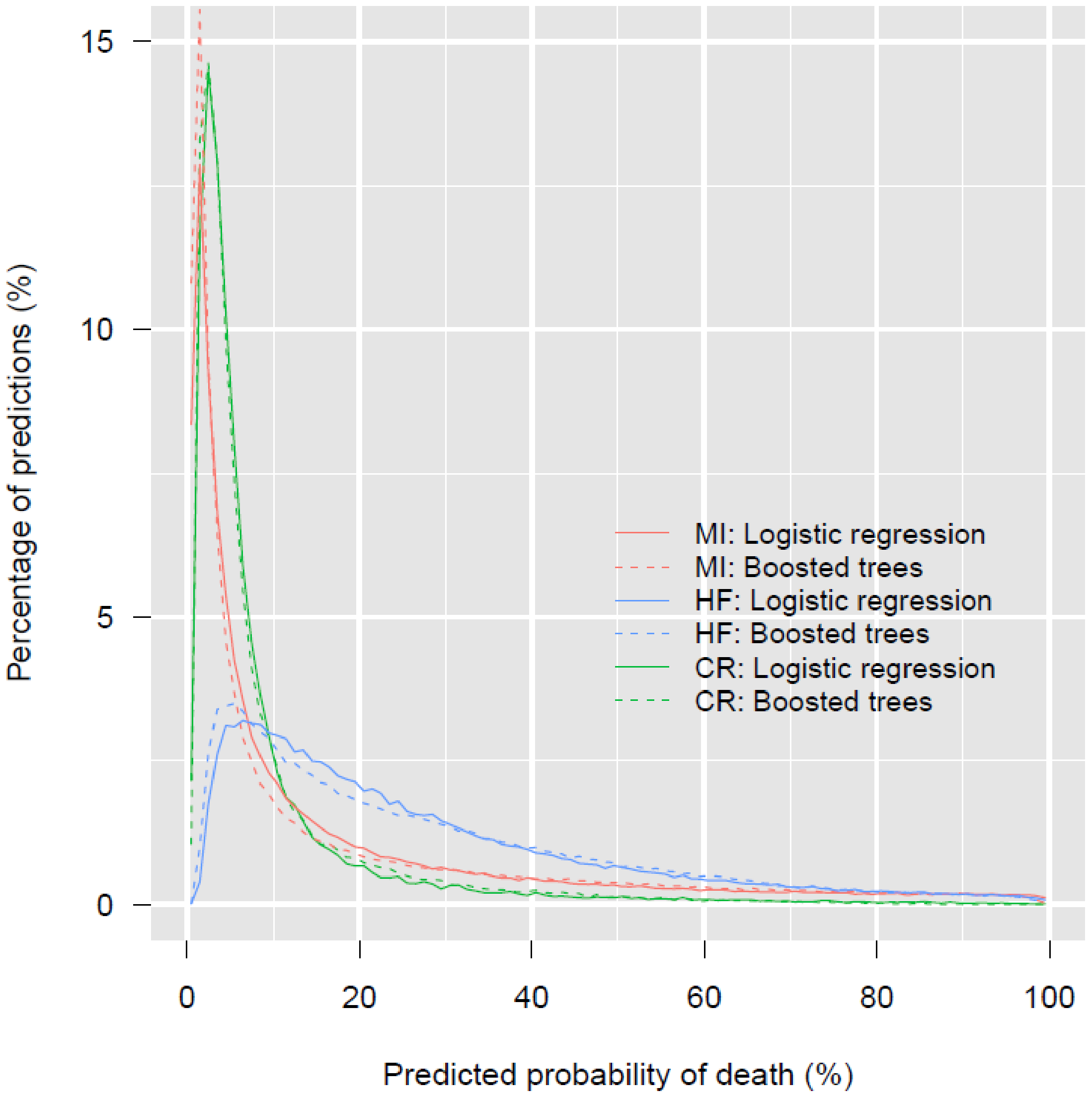
IQR=interquartile range. [a]Scaled such that the most deprived area of residence nationally had a value of 0 and the least deprived area had a value of 10. [b]Relative frequency of each three-character code was at least 0.5% in the given population. [c]Median values of Cohen's kappa coefficient across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

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)

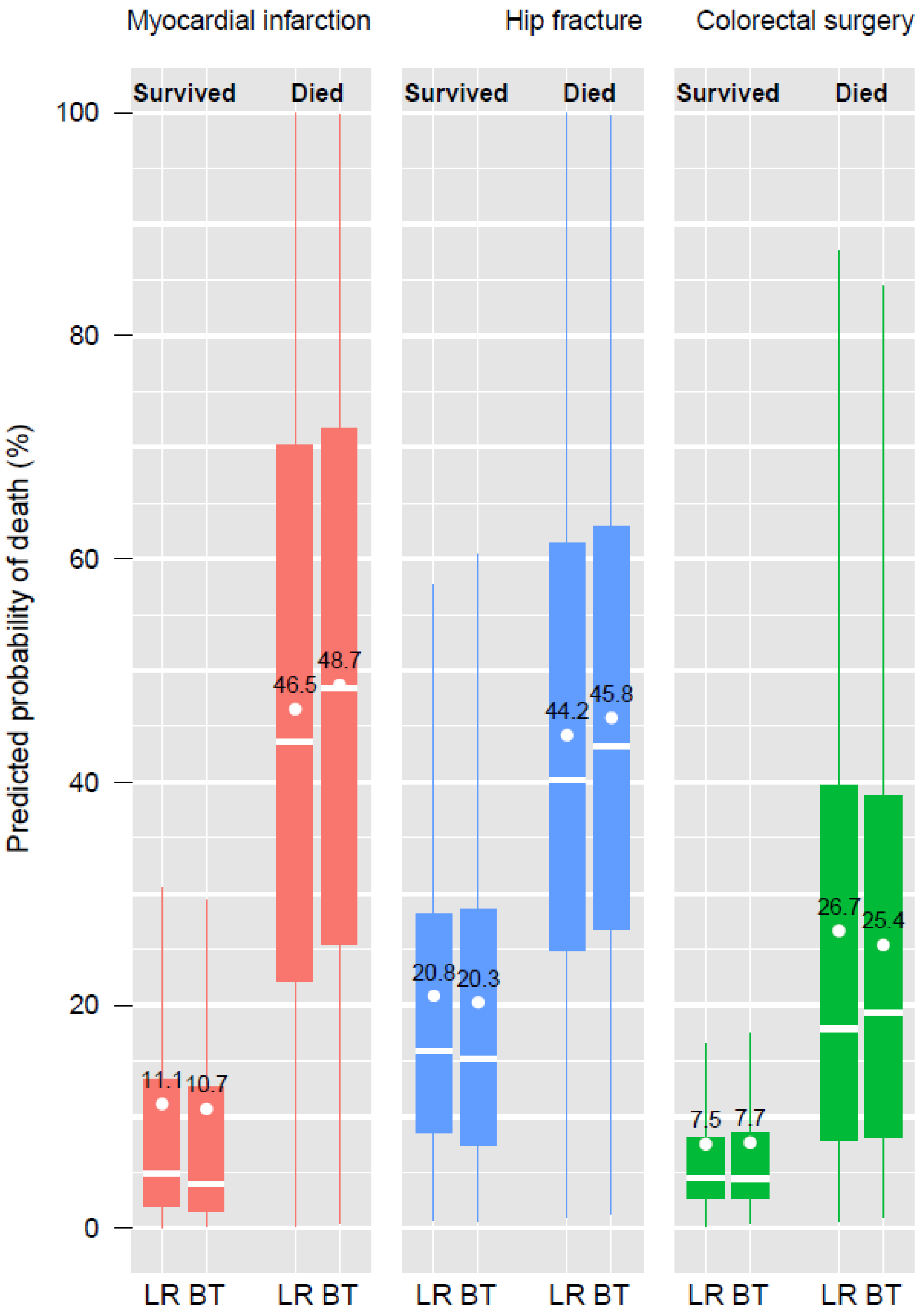
Results for boosted trees correspond to models with 500 boosting iterations in the myocardial infarction and hip fracture populations and 200 iterations in the colorectal surgery population.

Figure 1. Frequency distributions of predicted probabilities of death, by population and method



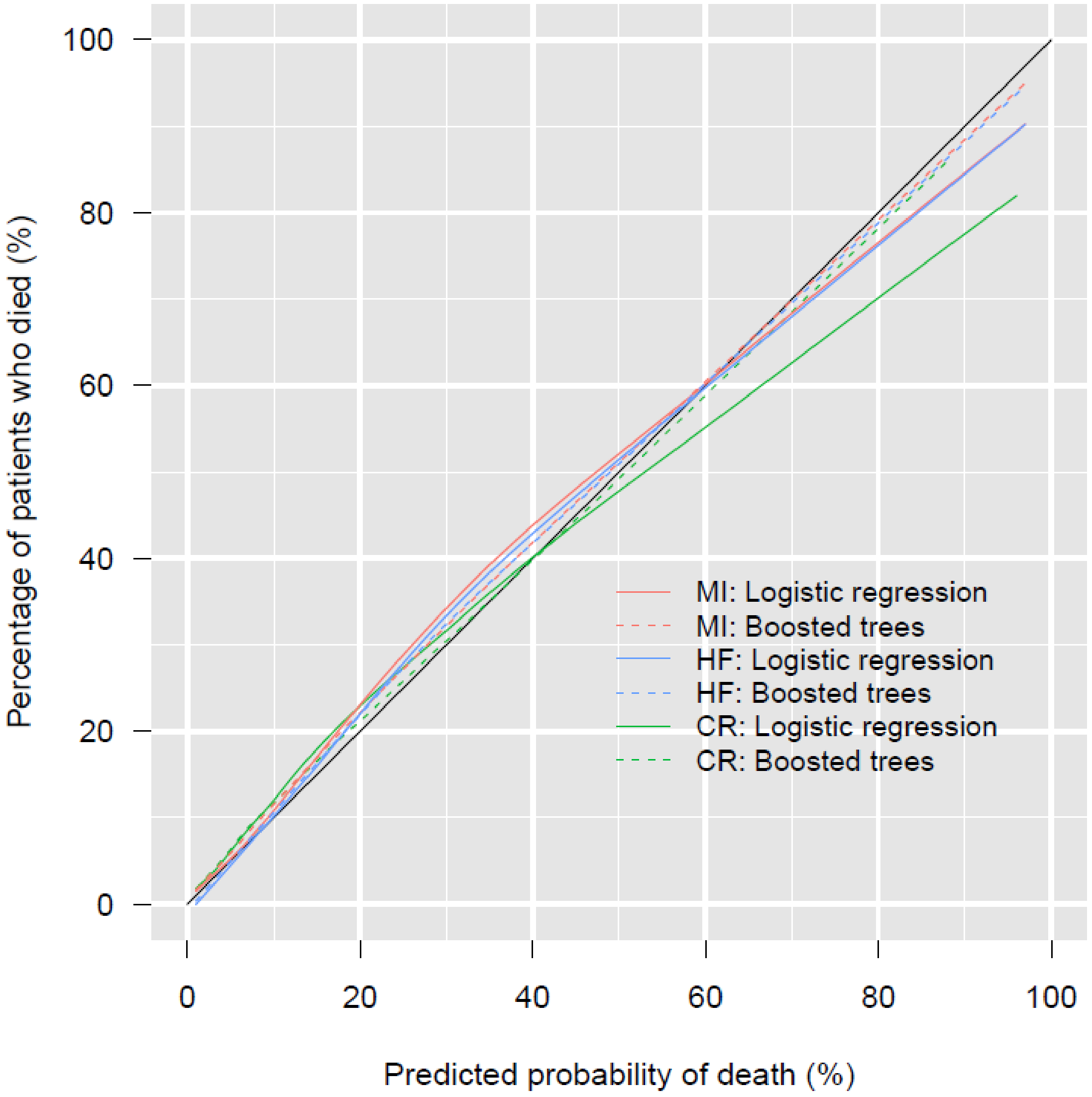
MI: myocardial infarction; HF: hip fracture; CR: colorectal surgery. In the MI population, 5% of patients had predicted probabilities equal to or greater than 72.5%. The corresponding values in the HF and CR populations were 73.9% and 35.7%, respectively.

Figure 2. Frequency distributions of predicted probabilities of death, by population, outcome, and method



LR: logistic regression; BT: boosted trees. 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.

Figure 3. Calibration plots for the logistic regression and boosted tree models, by population, corrected for optimism using 250 bootstrap samples (shown with line of perfect calibration)



MI: myocardial infarction; HF: hip fracture; CR: colorectal surgery. 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.

Appendix

Appendix A1. Tuning parameters of boosted regression trees

Boosted trees are an ensemble of individual trees whose predictions are passed to the next tree in the sequence to improve upon. The number of trees included in the ensemble is a parameter that can be tuned to improve performance. Each tree is developed by splitting a predictor at the best cut-point at each level of the tree. The maximum number of levels, or 'depth', of each tree is another tuning parameter, as are the minimum number of observations allowed to be at the end of one part of the tree and the observations used to fit each tree. The contribution of each tree to the overall ensemble is shrunk to reduce model overfitting; this shrinkage factor or 'learning rate' must also be tuned.

The study protocol[1] stated that the boosted trees would be tuned by varying these five parameters (see table below) and selecting the model with the smallest negative log-likelihood across cross-validation folds. These results showed that different combinations of parameters generally gave similar minimum values of the negative log-likelihood across the range of boosting iterations (1 to 500). The final results were therefore obtained varying the number of boosting iterations only (see far right column below), using 250 bootstrap samples to calculate optimism and 95% confidence intervals.

Parameter	Description	Values tested in cross-validation	Values used in final models
Number of iterations	Maximum number of boosting iterations	1 and 25 to 500 (in steps of 25)	100 to 500 (in steps of 100)
Learning rate	Scales the contribution of each tree's predictions by this value when added to the existing model	0.05 and 0.1	0.1
Maximum tree depth	Highest level of predictor interactions allowed in a tree	3 and 5	5
Minimum node weight	Stops tree splitting if the sum of observation weights in a node is less than this parameter	10 and 100	100
Subsample fraction	Fraction of observations in the training dataset randomly chosen to fit the next tree	0.5 and 1	1

The scaled Brier scores and *c*-statistics obtained from the models tuned using cross-validation are shown below and were similar to those obtained using bootstrapping (given in the main text).

	Acute myocardial infarction	Hip fracture	Major colorectal cancer surgery
Scaled Brier score (%)			
Logistic regression	34.4	22.8	17.0
Boosted trees	35.6	23.6	15.2
***c*-statistic**			
Logistic regression	0.883	0.798	0.809
Boosted trees	0.888	0.802	0.799

[1] 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* 2019:19011338. doi: 10.1101/19011338.

Appendix A2. Methods of the secondary analyses

Five-year look-back period

The main analysis defined a binary predictor for each ICD-10 code based on whether it was recorded or not within one year before the index date. A secondary analysis extended this ‘look-back period’ to five years (1825 days) to account for more diagnostic information. The frequency threshold for including ICD-10 codes was set at 1% so that the numbers of ICD-10 code predictors were similar to in the main analysis (which had a 0.5% threshold but shorter look-back period).

	Acute myocardial infarction	Hip fracture	Major colorectal cancer surgery
ICD-10 codes in 1-year look-back period			
Number of codes included*	202	257	209
Median frequencies (%) 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)
ICD-10 codes in 5-year look-back period			
Number of ICD-10 codes included*	206	251	182
Median frequencies (%) of codes (IQR)	2.7 (1.5 to 5.4)	3.3 (1.7 to 7.1)	2.7 (1.6 to 6.6)
Median number of codes per patient (IQR)	9 (5 to 15)	13 (8 to 21)	8 (5 to 13)
Median number of days to last record of a given code (IQR)	0 (0 to 506)	2 (0 to 548)	4 (0 to 136)

IQR=interquartile range. *Relative frequency of each three-character code was at least 0.5% (one-year look-back) or 1% (five-year look-back) in the given population.

We first estimated associations between the outcome and predictors (now defined using a five-year look-back period) using logistic regression. We then supplemented this model with an extra variable for each ICD-10 code which recorded the number of days before the index date that each code was last recorded (if at all). The resulting models assumed linear associations for these timing variables and were again estimated using logistic regression (see protocol[1] for model equation). To model non-linear associations for the timing variables, we also fitted generalised additive models with smoothing splines (with three degrees of freedom) for the continuous variables.

For the gradient boosted trees, a single predictor for each ICD-10 code was used which recorded the number of days before the index date that the code was last recorded (from 0 to 1825 days). If a patient did not have a given code recorded, the value of the variable for that code was set to a number (2000) that was arbitrarily larger than the maximum recorded value of 1825; the exact larger number chosen was arbitrary as trees dichotomise variables at optimal cut points. The trees were again fitted using the XGBoost algorithm and the tuning parameter values in Appendix A1.

Restricted cubic splines and interaction terms in logistic regression models

In each population, we added restricted cubic splines with three knots for the age and socioeconomic status variables in the logistic regression models. In the same models, we also included two-way interactions between three predictors. In the myocardial infarction population, these predictors were age, heart failure (I50), and cardiac arrest (I46). In the hip fracture population, the relevant predictors were age, sex, and other medical care (Z51). In the colorectal surgery population, the relevant predictors were age, secondary cancer of the lymph nodes (C77), and secondary cancer of the respiratory and digestive organs (C78). These variables were chosen as they were important predictors (Appendix A4) and were relatively frequent among patients whose predicted risks of death were in the top 5% (Appendix A5); this part of the predicted risk distribution was poorly calibrated in the logistic regression models that did not include interactions. To assess the effects on calibration, we plotted calibration curves for these models without and with splines and interactions (Appendix A6).

Random forests and neural networks

In addition to boosted trees, random forests and neural networks are two of the most popular machine learning approaches for analysing structured healthcare data. Like boosted trees, they model interactions between predictors by design but differ in how the model is constructed.

Random forests are an ensemble of individual trees in which predictions are averaged over all trees. A key difference to boosted trees is that random forest algorithms do not pass the predictions of one tree to the next tree in a sequence of trees. A defining characteristic of random forests is that the predictors used to split the tree are chosen at random at each split. This decorrelates trees to reduce the variance of their averaged predictions. The number of predictors to be randomly sampled at each split is a tuning parameter often set as the square root of the total number of predictors; we also tested twice the square root as an alternative value of this parameter. The other tuning parameter we varied was the minimum number of observations allowed to be in an end 'node' of the tree, which we set at 10 and 100. We fitted 500 trees in each random forest model as default in the ranger package in R.

Neural networks model an outcome using an intermediate set of unobserved, or 'hidden', variables which themselves are linear combinations of the original predictors. The number of intermediate layers of hidden variables can be varied, as can the number of variables in each layer, to improve performance. We fitted models with a single intermediate layer and two or four hidden variables in this layer. As neural networks are highly flexible, they tend to over-fit the relationship between predictors and the outcome. To address this, a penalisation term, or 'weight decay', can be used. We tested weight decays of 0 (no decay) and 0.1 in each model. All predictors were mean-standardised before fitting the models using the nnet package in R.

The values of tuning parameters that minimised the negative log-likelihood across five repeats of 5-fold cross-validation were used to develop the final random forest and neural network models.

1. 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* 2019:19011338. doi: 10.1101/19011338.

Appendix A3. R code used to implement the different estimation methods

The data were stored in the data frame ‘dfModel’ with the outcome status (‘mort’) recorded in the first column. The predictor variables (age, sex, socioeconomic status and ICD-10 code predictors) were recorded in the other columns of the data frame.

Logistic regression

```
logreg <- glm(mort ~ ., family = "binomial", data = dfModel)
```

Boosted trees

```
library(xgboost)

tune <- list(eta = 0.1, max_depth = 5, min_child_weight = 100,
             objective = "binary:logistic", eval_metric = "logloss")

trees <- xgboost(data = as.matrix(dfModel[, -1]), label = dfModel$mort,
                 params = tune, nrounds = 500)
```

Random forests

```
library(ranger)
library(caret)

ctrl <- trainControl(summaryFunction = mnLogLoss,
                     method = "repeatedcv",
                     number = 5,
                     repeats = 5,
                     classProbs = TRUE)

vars <- floor(sqrt(ncol(dfModel) - 1))

rangerGrid <- expand.grid(mtry = c(vars, 2 * vars), splitrule = "gini",
                          min.node.size = c(10, 100))

set.seed(145134)
rangerTune <- train(mort ~ .,
                    method = "ranger",
                    data = dfModel,
                    tuneGrid = rangerGrid,
                    trControl = ctrl,
                    metric = "logLoss",
                    maximize = FALSE)
```

Neural networks

```
library(nnet)
library(caret)

ctrl <- trainControl(summaryFunction = mnLogLoss,
                     method = "repeatedcv",
                     number = 5,
                     repeats = 5,
                     classProbs = TRUE)

nnetGrid <- expand.grid(size = c(2, 4), decay = c(0, 0.1))

set.seed(145134)
nnetTune <- train(mort ~ .,
                  method = "nnet",
                  data = dfModel,
                  tuneGrid = nnetGrid,
                  trControl = ctrl,
                  metric = "logLoss",
                  maximize = FALSE,
                  preProc = c('center', 'scale'),
                  MaxNWts = 2000)
```

Appendix A4. Important predictors in the logistic regression and boosted tree models

Logistic regression

Myocardial infarction		χ^2	Hip fracture		χ^2	Colorectal surgery		χ^2
Age		21.2	Age		18.3	C78	Secondary malignant neoplasm of respiratory and digestive organs	27.9
I46	Cardiac arrest	14.1	Z51	Other medical care (e.g. chemotherapy)	5.9	Age		7.9
R57	Shock, not elsewhere classified	4.9	F03	Unspecified dementia	5.0	I46	Cardiac arrest	7.0
Z51	Other medical care (e.g. chemotherapy)	3.9	Sex		3.0	C77	Secondary and unspecified malignant neoplasm of lymph nodes	5.4
I50	Heart failure	2.9	I46	Cardiac arrest	2.9	K65	Peritonitis	3.0
G93	Other disorders of brain	1.7	C78	Secondary malignant neoplasm of respiratory and digestive organs	2.4	Z51	Other medical care (e.g. chemotherapy)	2.5
N17	Acute renal failure	1.5	F01	Vascular dementia	2.4	D12	Benign neoplasm of colon, rectum, anus, and a nal canal	1.6
C78	Secondary malignant neoplasm of respiratory and digestive organs	1.4	C34	Malignant neoplasm of bronchus and lung	1.9	C79	Secondary malignant neoplasm of other sites	1.5
C34	Malignant neoplasm of bronchus and lung	0.9	C79	Secondary malignant neoplasm of other sites	1.8	K55	Vascular disorders of intestine	1.1
F03	Unspecified dementia	0.8	I50	Heart failure	1.7	I48	Atrial fibrillation and flutter	0.9

The variable importance measure, χ^2, is the partial Wald chi-square statistic for a given variable as a percentage of the statistic for the overall model.

Boosted trees

Myocardial infarction		Gain	Hip fracture		Gain	Colorectal surgery		Gain
Age		38.9	Age		21.4	C78	Secondary malignant neoplasm of respiratory and digestive organs	27.2
I46	Cardiac arrest	11.6	Z51	Other medical care (e.g. chemotherapy)	8.1	Age		17.1
N17	Acute renal failure	6.3	J18	Pneumonia, organism unspecified	6.8	K65	Peritonitis	5.1
I50	Heart failure	5.6	F03	Unspecified dementia	6.7	N17	Acute renal failure	5.1
Z51	Other medical care (e.g. chemotherapy)	3.9	I50	Heart failure	4.3	C77	Secondary and unspecified malignant neoplasm of lymph nodes	4.8
R57	Shock, not elsewhere classified	3.9	Sex		3.9	E87	Other disorders of fluid, electrolyte, and acid-base balance	3.8
E87	Other disorders of fluid, electrolyte, and acid-base balance	3.0	W19	Unspecified fall	2.9	Socio. status		3.5
J18	Pneumonia, organism unspecified	2.2	I46	Cardiac arrest	2.8	Z51	Other medical care (e.g. chemotherapy)	2.8
C78	Secondary malignant neoplasm of respiratory and digestive organs	1.5	N17	Acute renal failure	2.6	A41	Other septicemia	2.8
N18	Chronic renal failure	1.4	G30	Alzheimer's disease	2.5	J96	Respiratory failure, not elsewhere classified	2.3

The variable importance measure, ‘Gain’, is the percentage contribution of each variable to the model based on the total gain from this variable’s splits in the trees in minimising the negative log-likelihood.

Appendix A5. Frequent ICD-10 codes recorded for patients whose predicted risks of death were in the top 5% of predictions, by method

Myocardial infarction (n=10 006)

Logistic regression - Prob(death) ≥ 0.725				**Boosted trees** - Prob(death) ≥ 0.721			
Code		**n**	**%**	**Code**		**n**	**%**
I10	Essential (primary) hypertension	6527	65.2	I10	Essential (primary) hypertension	6016	60.1
I50	Heart failure	6280	62.8	I50	Heart failure	5506	55.0
I25	Chronic ischemic heart disease	5691	56.9	I25	Chronic ischemic heart disease	5320	53.2
N17	Acute renal failure	5033	50.3	N17	Acute renal failure	4536	45.3
Z86	Personal history of certain other diseases	4364	43.6	Z86	Personal history of certain other diseases	4040	40.4
I48	Atrial fibrillation and flutter	4180	41.8	I46	Cardiac arrest	3802	38.0
N18	Chronic renal failure	3834	38.3	I48	Atrial fibrillation and flutter	3747	37.4
I46	Cardiac arrest	3647	36.4	N18	Chronic renal failure	3381	33.8
E11	Non-insulin-dependent diabetes mellitus	3537	35.3	E11	Non-insulin-dependent diabetes mellitus	3212	32.1
E87	Other disorders of fluid, electrolyte, and acid-base balance	3495	34.9	Z92	Personal history of medical treatment	3076	30.7

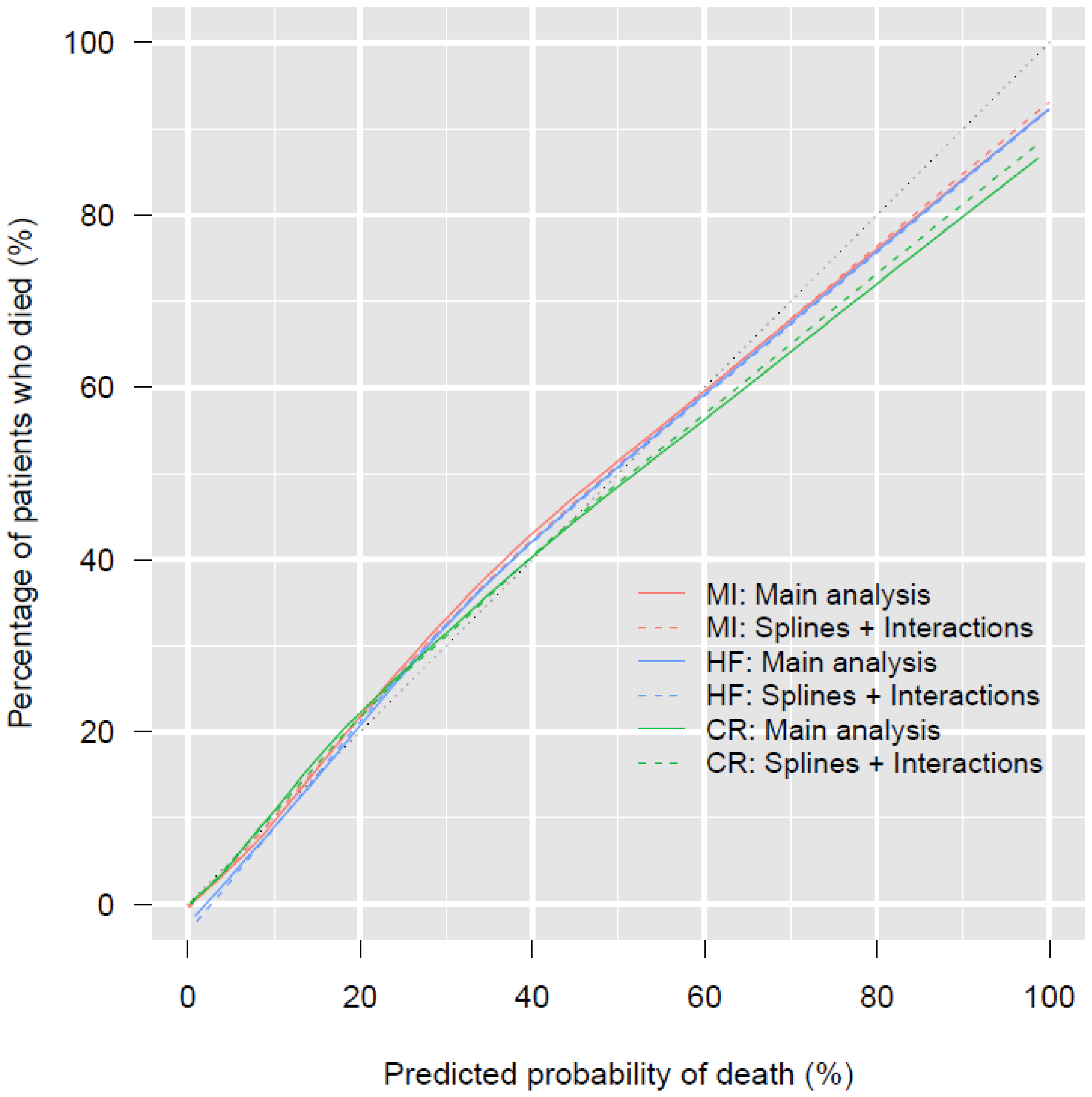
Hip fracture (n=8483)

Logistic regression - Prob(death) ≥ 0.739				**Boosted trees** - Prob(death) ≥ 0.731			
Code		**n**	**%**	**Code**		**n**	**%**
I10	Essential (primary) hypertension	5069	59.8	I10	Essential (primary) hypertension	4686	55.2
W19	Unspecified fall	5035	59.4	W19	Unspecified fall	4668	55.0
J18	Pneumonia, organism unspecified	4368	51.5	J18	Pneumonia, organism unspecified	4129	48.7
I48	Atrial fibrillation and flutter	4216	49.7	I48	Atrial fibrillation and flutter	3953	46.6
Z86	Personal history of certain other diseases	3855	45.4	Z51	Other medical care (e.g. chemotherapy)	3586	42.3
N17	Acute renal failure	3748	44.2	N17	Acute renal failure	3506	41.3
R29	Other symptoms and signs involving the nervous and musculoskeletal systems	3537	41.7	Z86	Personal history of certain other diseases	3429	40.4
I50	Heart failure	3483	41.1	I50	Heart failure	3172	37.4
N18	Chronic renal failure	3381	39.9	N18	Chronic renal failure	3134	36.9
Z51	Other medical care (e.g. chemotherapy)	3257	38.4	R29	Other symptoms and signs involving the nervous and musculoskeletal systems	2997	35.3

Colorectal surgery (n=2826)

Logistic regression - Prob(death) ≥ 0.357				**Boosted trees** - Prob(death) ≥ 0.354			
Code		**n**	**%**	**Code**		**n**	**%**
C78	Secondary malignant neoplasm of respiratory and digestive organs	1679	59.4	C78	Secondary malignant neoplasm of respiratory and digestive organs	1720	60.9
I10	Essential (primary) hypertension	1594	56.4	I10	Essential (primary) hypertension	1596	56.5
C77	Secondary and unspecified malignant neoplasm of lymph nodes	1220	43.2	C77	Secondary and unspecified malignant neoplasm of lymph nodes	1171	41.4
Z86	Personal history of certain other diseases	996	35.2	N17	Acute renal failure	958	33.9
N17	Acute renal failure	858	30.4	Z86	Personal history of certain other diseases	953	33.7
E87	Other disorders of fluid, electrolyte, and acid-base balance	812	28.7	E87	Other disorders of fluid, electrolyte, and acid-base balance	874	30.9
Z92	Personal history of medical treatment	809	28.6	I48	Atrial fibrillation and flutter	804	28.5
I48	Atrial fibrillation and flutter	791	28.0	K56	Paralytic ileus and intestinal obstruction without hernia	778	27.5
K56	Paralytic ileus and intestinal obstruction without hernia	784	27.7	J18	Pneumonia, organism unspecified	749	26.5
K63	Other diseases of intestine	723	25.6	K63	Other diseases of intestine	719	25.4

Appendix A6. Optimism-adjusted calibration plots and performance measures for the logistic regression models of the main analysis and with splines and selected interactions



	Myocardial infarction	Hip fracture	Colorectal surgery
Scaled Brier score (%)			
No splines or interactions	34.6	22.8	17.2
With splines and interactions	34.9	22.9	17.9
***c*-statistic**			
No splines or interactions	0.884	0.798	0.811
With splines and interactions	0.885	0.799	0.810

MI: myocardial infarction; HF: hip fracture; CR: colorectal surgery. Two-way interactions were included between: age, cardiac arrest (I46), and heart failure (I50) in the MI population; age, sex, and other medical care (Z51) in the HF population; and age, nodal metastases (C77), and respiratory/digestive metastases (C78) in the colorectal surgery population. Age and socioeconomic status were modelled using restricted cubic splines with three knots.

Appendix A7. Apparent and optimism-adjusted prediction performance of the logistic regression and boosted tree models as estimated using 250 bootstrap samples

	Myocardial infarction			**Hip fracture**			**Colorectal surgery**		
	App.	**Opt.**	**Adj.**	**App.**	**Opt.**	**Adj.**	**App.**	**Opt.**	**Adj.**
Scaled Brier score (%)									
Logistic regression	34.9	0.3	34.6	23.1	0.3	22.8	18.5	1.3	17.2
Boosted trees:									
100 iterations	36.0	0.8	35.1	23.9	0.9	22.9	17.5	1.9	15.6
200 iterations	37.4	1.3	36.0	25.1	1.5	23.6	18.7	3.0	15.7
300 iterations	38.0	1.7	36.3	25.8	2.0	23.8	19.3	3.7	15.6
400 iterations	38.5	2.1	36.4	26.3	2.4	23.9	19.8	4.3	15.5
500 iterations	38.9	2.4	36.5	26.7	2.7	24.0	20.3	4.8	15.5
***c*-statistic**									
Logistic regression	0.885	0.001	0.884	0.800	0.002	0.798	0.819	0.008	0.811
Boosted trees:									
100 iterations	0.888	0.003	0.886	0.803	0.005	0.798	0.813	0.011	0.802
200 iterations	0.893	0.004	0.889	0.809	0.008	0.801	0.820	0.017	0.803
300 iterations	0.895	0.005	0.890	0.813	0.010	0.803	0.824	0.021	0.803
400 iterations	0.897	0.006	0.890	0.815	0.012	0.803	0.827	0.024	0.802
500 iterations	0.898	0.007	0.891	0.818	0.014	0.804	0.829	0.027	0.803
Calibration-in-the-large									
Logistic regression	0	0.001	-0.001	0	0.000	0.000	0	0.000	0.000
Boosted trees:									
100 iterations	0.000	0.000	0.000	0.000	-0.001	0.001	0.000	-0.002	0.001
200 iterations	0.000	0.000	0.000	0.000	-0.001	0.001	0.000	-0.002	0.002
300 iterations	0.000	0.000	0.000	0.000	-0.001	0.001	0.000	-0.003	0.003
400 iterations	0.000	0.000	0.000	0.000	-0.001	0.001	0.000	-0.003	0.003
500 iterations	0.000	0.000	0.000	0.000	-0.001	0.001	0.000	-0.004	0.004
Calibration slope									
Logistic regression	1	0.007	0.993	1	0.011	0.989	1	0.039	0.961
Boosted trees:									
100 iterations	1.120	0.020	1.100	1.173	0.036	1.138	1.109	0.057	1.052
200 iterations	1.079	0.031	1.048	1.120	0.052	1.068	1.073	0.086	0.988
300 iterations	1.066	0.040	1.026	1.103	0.066	1.037	1.078	0.107	0.970
400 iterations	1.059	0.048	1.011	1.096	0.078	1.018	1.085	0.124	0.961
500 iterations	1.058	0.055	1.003	1.094	0.089	1.006	1.096	0.139	0.957
Integrated calibration index									
Logistic regression	0.012	0.000	0.012	0.015	0.000	0.015	0.007	0.000	0.007
Boosted trees:									
100 iterations	0.011	0.001	0.010	0.019	0.002	0.017	0.007	0.002	0.005
200 iterations	0.008	0.002	0.006	0.014	0.004	0.011	0.005	0.004	0.001
300 iterations	0.006	0.002	0.004	0.012	0.005	0.007	0.005	0.005	0.000
400 iterations	0.006	0.003	0.003	0.012	0.006	0.005	0.005	0.006	-0.001
500 iterations	0.005	0.003	0.002	0.011	0.007	0.004	0.006	0.007	-0.001

Appendix A8. Results of the secondary analyses

Five-year look-back period

The boosted trees attained the largest scaled Brier scores and *c*-statistics in each population, while the logistic regression models that did not account for the timings had the lowest scores (see table below). However, the boosted trees performed comparably to those from the main analysis (which only used a one-year look-back period). This may be partly because most ICD-10 codes were last recorded within a few days of the index dates (see Appendix A2). The approach may work better for non-hospitalised populations with more variation in the times since diagnosis codes were last recorded.

	Acute myocardial infarction	**Hip fracture**	**Major colorectal cancer surgery**
Scaled Brier score (%)			
Boosted trees	36.2	24.1	17.2
Generalised additive models	34.4	23.1	15.1
Logistic regression with time effects	33.7	22.2	14.9
Logistic regression	32.2	20.8	14.8
***c*-statistic**			
Boosted trees	0.889	0.804	0.807
Generalised additive models	0.883	0.799	0.796
Logistic regression with time effects	0.880	0.794	0.797
Logistic regression	0.876	0.789	0.800

Random forests and neural networks

The random forest models did not perform better than both the logistic regression and boosted tree models in any of the populations. Neural networks consistently performed worse than other models.

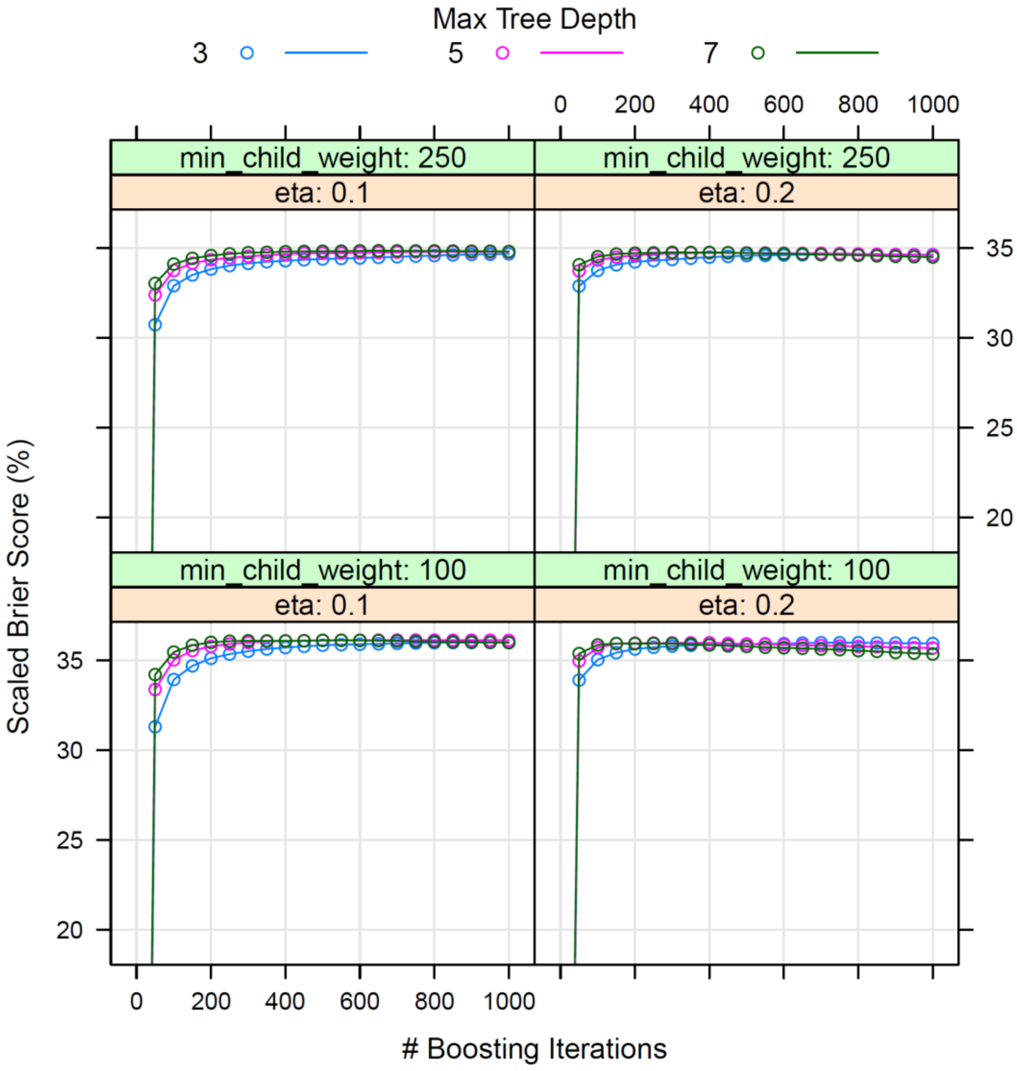
	Acute myocardial infarction	**Hip fracture**	**Major colorectal cancer surgery**
Scaled Brier score (%)			
Random forests	35.1	22.1	16.1
Neural networks	32.2	17.4	11.1
***c*-statistic**			
Random forests	0.887	0.795	0.808
Neural networks	0.878	0.774	0.784

Appendix A9. Results of additional analyses in response to peer reviewers

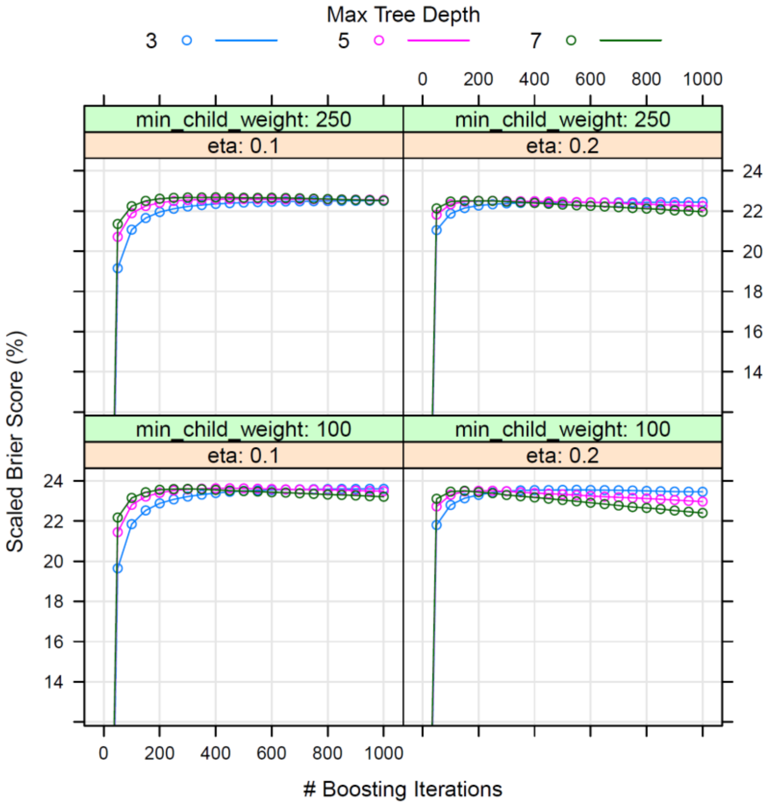
Further combinations of tuning parameters

The main analysis fitted boosted tree models with a maximum tree depth (‘Max Tree Depth’) of 5, learning rate (‘eta’) of 0.1, minimum node weight (‘min_child_weight’) of 100, and up to 500 boosting iterations (Appendix A1). Results for this combination can be seen as the pink line in the bottom left-hand panel of each figure below. The figures present the scaled Brier scores (estimated using five-fold cross-validation) when these tuning parameters were combined with different values. In each population, the maximum performance values were relatively insensitive to the choice of different combinations provided that the number of boosting iterations was tuned appropriately.

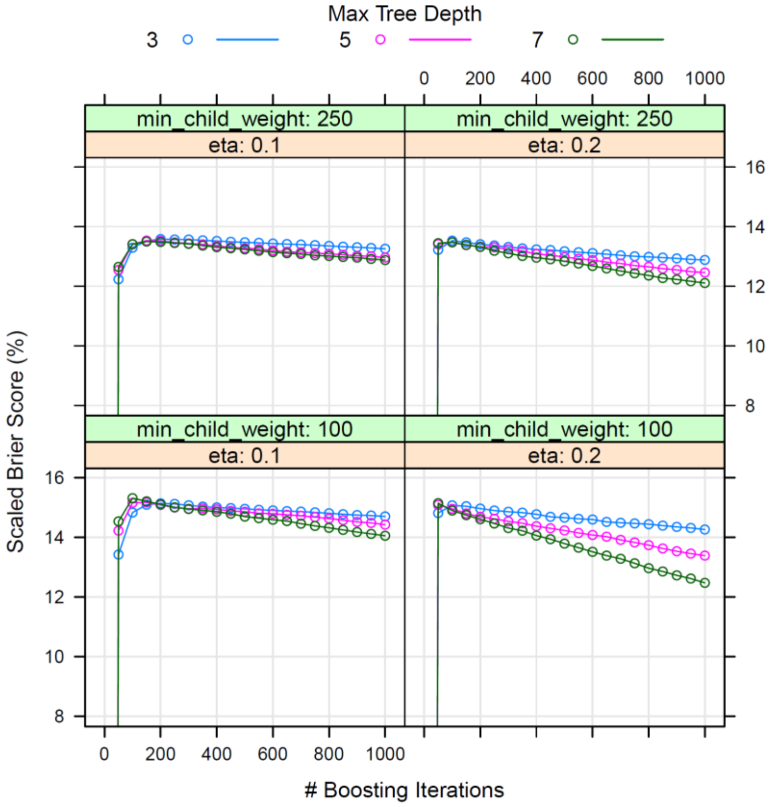
Myocardial infarction:



Hip fracture:



Colorectal surgery:



Frequency threshold of 0.1% for ICD codes

When only ICD codes with frequencies less than 0.1% (rather than 0.5%) were excluded from the set of predictor variables, the number of included ICD codes approximately doubled in each population. However, the performance of the resulting models in the original data (‘apparent performance’) hardly changed, as shown below by population and frequency threshold (0.5% or 0.1%).

	Myocardial infarction		**Hip fracture**		**Colorectal surgery**	
	0.5%	**0.1%**	**0.5%**	**0.1%**	**0.5%**	**0.1%**
Number of ICD codes	202	440	257	522	209	434
Scaled Brier score (%):						
Logistic regression	34.9	35.7	23.1	24.0	18.5	20.0
Boosted trees:						
100 iterations	36.0	36.0	23.9	23.8	17.5	17.5
200 iterations	37.4	37.3	25.1	25.1	18.7	18.7
300 iterations	38.0	38.0	25.8	25.8	19.3	19.3
400 iterations	38.5	38.4	26.3	26.2	19.8	19.8
500 iterations	38.9	38.8	26.7	26.7	20.3	20.3
***c*-statistic:**						
Logistic regression	0.885	0.887	0.800	0.805	0.819	0.827
Boosted trees:						
100 iterations	0.888	0.888	0.803	0.803	0.813	0.813
200 iterations	0.893	0.893	0.809	0.810	0.820	0.820
300 iterations	0.895	0.895	0.813	0.813	0.824	0.824
400 iterations	0.897	0.896	0.815	0.816	0.827	0.827
500 iterations	0.898	0.898	0.818	0.818	0.829	0.829