An approach for selecting small sets of diagnosis codes with high prediction performance in large datasets of electronic medical records

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

London School of Hygiene and Tropical Medicine, Keppel St, London, WC1E 7HT, UK

[a]Department of Health Services Research and Policy

[c]Department of Medical Statistics

Royal College of Surgeons of England, Lincoln's Inn Fields, London, WC2A 3PE, UK

[b]Clinical Effectiveness Unit

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

Abstract

Objective: To examine an approach for selecting small sets of diagnosis codes with high prediction performance in large datasets of electronic medical records.

Study Design and Setting: Modelling study using national hospital and mortality 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 ICD-10 codes recorded for at least 0.5% of patients using logistic regression ('full' models). An approximation method was used to select fewer codes that explained at least 95% of variation in full model predictions ('reduced' models).

Results: One-year mortality was 17.2% (34 520) after myocardial infarction, 27.2% (46 115) after hip fracture, and 9.3% (5273) after colorectal surgery. Full models included 202, 257, and 209 ICD-10 codes in these populations. *C*-statistics for these models were 0.884 (95% CI 0.882, 0.886), 0.798 (0.795, 0.800), 0.810 (0.804, 0.817). Reduced models included 18, 33, and 41 codes and had *c*-statistics of 0.874 (95% CI 0.872, 0.876), 0.791 (0.788, 0.793), 0.807 (0.801, 0.813). Performance was also similar when measured using Brier scores. All models were well calibrated.

Conclusion: Our approach selected small sets of diagnosis codes that predicted patient outcomes comparably to large, comprehensive sets of codes.

Key words: Big data, electronic medical records, International Classification of Diseases, ICD-10, comorbidity, multimorbidity, prognosis, statistical models, variable selection

Running title: Selecting small sets of diagnosis codes with high prediction performance

Word count: 3454

What is new?

Key findings

• Our approach selected small sets of diagnosis codes that predicted one-year mortality comparably to large, comprehensive sets of codes, in three clinical populations

What this adds to what was known

• A relatively small set of diagnosis codes may predict most variation in a given patient outcome in a given study and such a set can be selected using statistical methods

What should change now?

• This approach may be useful to many studies that need to develop a study-specific measure of comorbidity using a large dataset of electronic medical records

1. Introduction

Electronic medical records are increasingly used for clinical, epidemiological, and healthcare research.[1,2] They offer growing opportunities to study large, representative populations over long periods of time and often contain many diagnosis codes representing a wide range of clinical information.[3] For example, the World Health Organization's International Classification of Diseases (ICD), used in many datasets, includes over 10 000 codes for different health attributes.[4]

Many studies use these diagnosis codes to model patients' overall morbidity.[5,6] Such models are applied widely, including in clinical prediction tools,[7] in randomised trials to assess patient characteristics,[8] and in observational studies to reduce confounding between treatment groups or healthcare providers.[9] In the global context of population ageing and greater burdens of non-communicable disease, models of morbidity are likely to be increasingly important.[10,11]

Large numbers of diagnosis codes could be included in these models when the study population is also large, as is common in electronic medical record studies.[12] Larger sets of codes may predict patient outcomes better.[13] However, these sets will also be more difficult to interpret, present, and apply in future studies or clinical practice.[14] Investigators therefore need to select a small set of codes that best balances model size and prediction performance.[15] This is difficult, however, as the number of different sets that can be chosen from n codes equals 2_n ($2_{30} \approx 1$ billion for example).[16]

This highlights the potential value of a data modelling approach that includes large, comprehensive sets of codes but produces a final model that includes far fewer codes and predicts the study outcome to a similar extent.[17] This model may then be small enough to easily interpret but still have close to maximum achievable performance. The existing literature has not investigated how this can be done in the context of electronic medical records and diagnosis codes.

In this study, we aimed to examine such an approach by comparing the prediction performance of models including large, comprehensive sets of ICD codes with models including fewer codes. One-

year mortality of three clinical populations was the modelling context, utilising linked national datasets of routine hospital and mortality data in England.

2. Methods

2.1 Study populations

We analysed Hospital Episode Statistics Admitted Patient Care data—administrative data for all inpatient care funded by the National Health Service (NHS) in England.[18] Each record relates to an 'episode' of care under the same senior clinician and has 20 fields for ICD-10 codes[4] relevant to that episode. The first field contains the primary diagnosis—the main condition treated.

The study populations were patients admitted for acute myocardial infarction (MI), hip fracture (HF), or major surgery for colorectal cancer (CS). MI (I21-22[19,20]) and HF (S72.0-S72.2[21,22]) patients were identified from the ICD-10 codes recorded as the primary diagnosis in the first episode of each admission. CS 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).[23-26]

These populations represent many admissions and vary in terms of clinical specialty, co-existing conditions, and mortality. We included MI and CS patients aged 18 years or older and HF patients aged 60 years or older[22] whose admission was from January 1, 2015, to December 31, 2017. Only a patient's earliest admission of two or more of the same type (MI, HF, CS) was included.

2.2 Outcome

The outcome was death up to and including 365 days after the date of admission (MI and HF) or procedure (CS). We used the official dates of death recorded in Office for National Statistics mortality data[27] up to December 31, 2018. These records were linked to Hospital Episode Statistics based on each patient's NHS number, date of birth, sex, and postcode.[28] Approximately 95% of linked records

matched exactly on at least three of these variables; other records were linked allowing partial matches of dates of birth or using exact matches for two variables only.[28]

Mortality is the outcome most often used to assess models of ICD codes in hospital settings.[5,29] We analysed 365-day mortality as the other outcomes most often used—in-hospital and 30-day mortality—may be more strongly affected by the primary event than other conditions. Also, more deaths over a longer time span increased the effective sample size.[30]

2.3 Predictors

We defined a binary predictor for each ICD 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.[13] The first three characters define single conditions or other health-related attributes; fourth characters define sites, subtypes, and causes.[4] Higher levels of the ICD coding system—the 22 'chapters' and the 'blocks' of three-character codes—were not analysed, as these levels may be too broad to retain the predictive ability of the three-character codes.

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.[31-33] We used a 365-day period, rather than only using codes from the index episode, as this improved model performance in some studies.[5]

Patient age, sex, and socioeconomic status were also predictors, as is common when examining models of ICD codes.[5,29] 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).[34] We excluded patients with missing data (1.2%; 5346/431 626).

2.4 Model estimation

We first estimated associations between the outcome and the full set of predictors as the maximum likelihood estimates from logistic regression ('full' model). We then developed a 'reduced' model that approximated this full model following the proposals of Harrell.[17,35]

First in our approach, the predicted log-odds of the outcome for each patient was calculated from the coefficients of the full model. Second, an ordinary least squares regression model was fitted between these predictions and all predictors; the coefficients of this model were identical to those of the full model and R^2 equalled one, by definition. Third, the ICD code predictor whose omission caused the smallest decrease in R^2 (the 'approximation R^2') was removed; this was repeated until all ICD code predictors had been removed. This process was based on the fast variable elimination methods of Lawless and Singhal,[36] using the full model and Wald statistics for the submodels to select which variables to eliminate. As shown by Ambler, Brady, and Royston,[15] when all predictors are uncorrelated and standardised the decrease in R^2 from omission of a given predictor is proportional to its squared model coefficient. Fourth in our approach, the model with the fewest predictors and an approximation R^2 equal to or greater than 95% was selected as the final model. This reduced model explained at least 95% of variation in predictions from the full model.

We refer to models without any ICD codes as 'baseline' models (including only age, sex, and socioeconomic status). Modelling non-linear associations for age and socioeconomic status using restricted cubic splines did not improve the prediction performance of the baseline or full models, so all final results assume linear associations. We did not examine interactions, partly due to the difficulty in estimating them reliably when most ICD codes are infrequent.

2.5 Model performance

Overall model performance was measured using Brier scores.[37] These scores equalled the mean of squared differences between predicted probabilities of death and observed outcomes. We scaled these scores from 0–100% (0% if non-informative and 100% if perfect).[38]

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 die.[17] *C*-statistics equal one for perfect models and 0.5 for random predictions. To assess calibration, we calculated the integrated calibration index (ICI),[39] calibration-in-the-large, and calibration slopes.[40] ICI and calibration-in-the-large assess the calibration of 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.

We first calculated the above measures in the original data used to fit the regression models ('apparent performance'). We then repeated all modelling steps in each of 500 bootstrap samples and, for each sample, calculated the performance of the resulting models 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.[17,41,42]

To contextualise model performance, we compared the results to models based on the conditions of Charlson *et al.*[43] and Elixhauser *et al.*[44] which are those most often used to measure inpatient morbidity (see Appendix for further details).[5,29] These models included the baseline predictors (age, sex, and socioeconomic status) and 17 or 31 binary predictors for the Charlson and Elixhauser conditions, respectively.

2.6 Sensitivity analyses

Five analyses tested the sensitivity of results to the methods. First, ICD codes recorded for less than 1% of patients were excluded. Second, we defined predictors using the index episode only or a three-year look-back period. Third, we fitted models including both the Charlson and Elixhauser conditions; for a condition included in both sets, the predictors were defined using the broadest ICD code definition from the two sets.[45] Fourth, we grouped all ICD codes into 260 Clinical Classification Software (CCS) groups and replaced the ICD code predictors with binary predictors for these

groups.[46] CCS groups are intended to aggregate ICD codes into a manageable number of clinically meaningful categories.[47] Fifth, we assessed changes in coefficients from the full models when penalised maximum likelihood estimation was used.[17,48,49]

We pre-specified the study methods in a published protocol and performed the main and sensitivity analyses as described in this protocol.[50] Data management was done using Stata (version 15). R (version 3.5) was used for all statistical analysis.

In response to a peer reviewer's suggestion, we conducted an additional analysis that used three alternative approaches for selecting which ICD codes from the full models to include in the final models: backwards elimination using the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) and the least absolute shrinkage and selection operator (lasso).[51,52] The lambda value of the lasso model was tuned using 5-fold cross-validation.[53]

3. Results

The percentage of patients who died within one year was 17.2% (34 520/200 119) after MI, 27.2% (46 115/169 646) after HF, and 9.3% (5273/56 515) after CS.

Overall, 8445 unique four-character ICD codes and 1857 three-character codes were recorded. In each population, 202 to 257 three-character codes were recorded for at least 0.5% of patients and included in further analysis. The numbers of deaths per predictor variable were 168 (34 520/205; MI), 177 (46 115/260; HF), and 25 (5273/212; CS).

Most included ICD codes had low frequencies (Figure 1; Table A1 in the Appendix lists the 20 most frequent). The median number of codes included for each patient ranged from 6 to 9 across the populations. Correlations between codes were generally very low (Table 1). The maximum variance inflation factor for an ICD code predictor in any of the populations was 3.6.

In the original data, the full models (with a predictor for each ICD code) attained scaled Brier scores of 34.9% (MI), 23.1% (HF), and 18.5% (CS). Many codes were removed from these models without the explained variation (R_2) in predictions decreasing below 95% (Figure 2).

The reduced models included 18 (MI), 33 (HF), and 41 (CS) ICD codes. Overall, 61 unique codes were included, with 41 codes included in only one model (see Table A2).

The corresponding scaled Brier scores were 32.2%, 21.9%, and 17.6%, which were only slightly lower than those for the full models. The *c*-statistics were also similar between the full and reduced models and indicated very good discrimination ($c \geq 0.791$ across all models). These measures were much lower when all ICD codes had been removed (Figure 2; Figure A1).

The approximation R_2, scaled Brier scores, and *c*-statistics at different numbers of ICD codes were highly correlated in each population (minimum Pearson's $r = 0.984$), such that the shapes of relationships between these measures and the number of codes were similar.

After adjusting these performance measures for optimism using bootstrapping, the values were similar, indicating minimal overfitting (Table 2, Figure 3). Values of the integrated calibration index and calibration-in-the-large were close to zero, implying that the models accurately predicted risks of death, on average, and the overall log-odds of death in each population (Table 2). All calibration slopes were only slightly less than the perfect value of 1 (minimum=0.961; 95% CI 0.935–0.987), indicating that model predictions were slightly too extreme (Figure A2).

The codes included in the reduced models were reasonably stable across bootstrap samples: 13/18 (MI), 27/33 (HF), and 28/41 (CS) codes were selected in ≥90% of samples (Figure A3).

The full and reduced models consistently performed better than models based on the Charlson or Elixhauser conditions in each population, when all ICD codes were eligible for inclusion or when a restricted set was used (Figure A4). For example, the scaled Brier scores for the reduced, Charlson-

based, and Elixhauser-based models in the MI population were 32.2%, 22.4%, and 23.7% (all codes) and 26.0%, 20.7%, and 22.0% (restricted codes); the score for the baseline model was 15.0%.

3.1 Sensitivity analyses

The sensitivity analyses did not identify an approach that performed better than that used in the main analysis (Table A3). For example, only including ICD codes with frequencies of at least 1% (versus 0.5%) led to full models with slightly worse overall and discrimination performance in each population (maximum decreases in scaled Brier score: 2.2%; *c*-statistic: 0.011); these models included far fewer codes than when a 0.5% frequency threshold was used (130 versus 202 for MI, 177 versus 257 for HF, and 147 versus 209 for CS). When penalisation was used, full model coefficients were very similar to those from the main analysis (Figure A5).

Variable selection using AIC, BIC, and the lasso produced models with greater numbers of ICD codes than were included in the reduced models. The ranges of the number of codes included in the final models by each approach, across populations, were 85 to 169 (AIC), 51 to 99 (BIC) and 121 to 221 (lasso). Model performance was similar to that of the full models (Table A4).

4. Discussion

In a large dataset of electronic medical records, our approach consistently selected small sets of ICD-10 codes that performed comparably to large, comprehensive sets of codes. In each of the three populations, a relatively small set of codes explained at least 95% of variation in predictions from a much larger set of codes. Our approach therefore produced small models that had close to maximum achievable performance, very good discrimination, and were well calibrated.

The ICD codes included in the final models varied between populations, which was expected given the different characteristics of these groups. Overall, 41 of the 61 codes included in the reduced models were only included for one population. This variation could be even greater across different

outcomes, settings, and datasets, which supports the view that the diagnosis codes included in morbidity models should be tailored to the study in which the model will be used.[11,13]

The 95% R^2 threshold used to define the reduced models worked well, though a lower value (such as 90%) may be reasonable if the benefits of including fewer codes outweigh the costs of lower performance.[15] The MI model included the fewest codes, partly because age, sex, and socioeconomic status better predicted the outcome in this population than in the other two groups.

The models developed could be considered as models of patient morbidity, specifically 'morbidity burden' which includes the presence of multiple conditions, socio-demographic characteristics, and other health-related attributes.[54] Some of the included ICD codes may also be relevant to frailty and disability.[55] Our approach provides a general framework for selecting small sets of diagnosis codes to predict a particular outcome in a given study population and dataset. The codes included in the full model can be adapted to suit different morbidity constructs and clinical perspectives.

4.1 Defining and selecting codes

Harrell's proposals regarding model approximation[17,35] do not appear widely in the existing literature and were not mentioned in recent reviews of variable selection.[56,57] This could be partly because related approaches may produce similar final models to more popular methods, based on *P*-values or AIC for example, in most study contexts.[15,17] However, variable selection in large datasets using *P*-values or AIC is unlikely to produce models with relatively few predictors, as even weak predictors will have small *P*-values.[12] This is supported by the results of our analyses using AIC and BIC as the selection criteria and also applied to the lasso models.

A strength of our approach is that users can trade between the number of predictors included in the final model and prediction performance by varying the approximation R^2 threshold (which we set at 95%). Investigators requiring smaller models, to improve the feasibility of use in clinical practice for example, can quantify reductions in performance from removing predictors (as in Figure 2).

This method can be used with binary, continuous, and time-to-event outcomes; incorporate non-linear and interaction terms; and retain penalisation applied to the full models.[17] Variable selection methods that start with full models are preferred as they consider all correlations between predictors.[58]

Subject knowledge should be used to help select the candidate ICD codes.[12,17] Codes that are highly unlikely to have strong prognostic effects, are recorded unreliably, or are inappropriate in the study context should be excluded in advance. This may reduce the potential for model overfitting. Subject knowledge could also be used to pre-specify codes to force into the reduced model, such as those known to be important, and to assess the face validity of models.

Our proposed approach is essential, however, because prior knowledge alone is unlikely to clearly indicate an exact set of codes that best balances the number of included codes and prediction performance, and related decision-making may be non-transparent and unreproducible.

A general limitation of statistical variable selection methods is that the predictors included in the final model may vary between repeat samples of the data.[12,17] This is most problematic when a study's main interest is the estimated association between each individual predictor and the outcome (and its variability).[59] We focused on developing reduced models with high prediction performance, such that low variability in performance of the models overall (as shown in Figure 3) was most relevant.

4.2 Limitations of the study

Our approach should be applied to other populations, outcomes, and datasets to assess whether it performs similarly well. We tested it in three varied inpatient populations, but other populations could have different case-mixes that affect our results. The prognostic effects of codes are also likely to differ between outcomes, such as mental health and physical outcomes.[60] Variation in the recording of codes between datasets may also affect the results of our approach.

Future research should examine how our approach should be used in smaller datasets. Using subject knowledge to exclude more codes from the full models is likely to be particularly important in smaller

samples, to avoid model overfitting.[30] Shrinkage methods are not guaranteed to work well in any given study due to uncertainty in estimating shrinkage or penalty terms. Electronic medical records may often provide very large samples that exceed the minimum sizes required.[3,30]

Given the large set of binary predictors defined by the ICD codes, and the potential for interactions between them, modelling approaches based on random forests or boosted trees may predict outcomes well. However, the low frequencies of most codes may mean that a given combination of codes is not recorded very often such that any interaction is estimated imprecisely. This may be improved by including broader levels of the hierarchical ICD coding system as predictors, but these levels may group codes associated with very different prognoses thus reducing the prediction value. The performance of these approaches could be examined in future research.

4.3 Implications for research

Many studies use diagnosis codes from electronic medical records to model patient morbidity. We have shown how small sets of codes can be selected that predict patient outcomes almost as well as much larger sets of codes. R code to apply our approach is given in the Appendix.

In the global context of population ageing and greater burdens of non-communicable disease, patient morbidity is becoming more complex.[61-63] At the same time, electronic medical records are increasing in volume and scope, presenting growing opportunities to better model this complexity.[2] Further research should investigate how we can utilise these records to improve morbidity measures.

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. **L.D.S.**: Conceptualization, 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. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* 2015;12:e1001885. doi: 10.1371/journal.pmed.1001885.

2. 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.

3. Jordan KP, Moons KG. Electronic healthcare records and prognosis research. In: Riley RD, van der Windt D, Croft P, Moons KG. (eds.) *Prognosis Research in Healthcare: Concepts, Methods, and Impact*. Oxford: Oxford University Press; 2019. p. 298-310.

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

5. Sharabiani MT, Aylin P, Bottle A. Systematic review of comorbidity indices for administrative data. *Med Care* 2012;50:1109-18.

6. Huntley AL, Johnson R, Purdy S, Valderas JM, Salisbury C. Measures of multimorbidity and morbidity burden for use in primary care and community settings: a systematic review and guide. *Ann Fam Med* 2012;10:134-41.

7. Brooks GA, Kansagra AJ, Rao SR, Weitzman JI, Linden EA, Jacobson JO. A Clinical Prediction Model to Assess Risk for Chemotherapy-Related Hospitalization in Patients Initiating Palliative Chemotherapy. *JAMA Oncol* 2015;1:441-7.

8. Yealy DM, Kellum JA, Huang DT, et al. A randomized trial of protocol-based care for early septic shock. *N Engl J Med* 2014;370:1683-93.

9. Werner RM, Bradlow ET. Relationship between Medicare's hospital compare performance measures and mortality rates. *JAMA* 2006;296:2694-702.

10. Stirland LE, González-Saavedra L, Mullin DS, Ritchie CW, Muniz-Terrera G, Russ TC. Measuring multimorbidity beyond counting diseases: systematic review of community and population studies and guide to index choice. *BMJ* 2020;368:m160. doi: 10.1136/bmj.m160.

11. Johnston MC, Crilly M, Black C, Prescott GJ, Mercer SW. Defining and measuring multimorbidity: a systematic review of systematic reviews. *Eur J Public Health* 2019;29:182-9.

12. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 2nd ed. Cham: Springer; 2019.

13. 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.

14. Wyatt JC, Altman DG. Prognostic Models: Clinically Useful or Quickly Forgotten? *BMJ* 1995;311:1539-41.

15. Ambler G, Brady AR, Royston P. Simplifying a prognostic model: a simulation study based on clinical data. *Stat Med* 2002;21:3803-22.

16. Hocking RR, Leslie RN. Selection of the Best Subset in Regression Analysis. *Technometrics* 1967;9:531-40.

17. Harrell FE, Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. Cham: Springer; 2015.

18. 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.

19. 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.

20. 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.

21. 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.

22. Royal College of Physicians. *National Hip Fracture Database (NHFD) annual report 2016*. 2016. Available from: https://www.nhfd.co.uk/report2016.

23. 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.

24. 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.

25. Morris EJ, Taylor EF, Thomas JD, et al. Thirty-day postoperative mortality after colorectal cancer surgery in England. *Gut* 2011;60:806-13.

26. 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.

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

28. 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.

29. 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.

30. 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.

31. 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.

32. Gensheimer MF, Henry AS, Wood DJ, et al. Automated Survival Prediction in Metastatic Cancer Patients Using High-Dimensional Electronic Medical Record Data. *J Natl Cancer Inst* 2018. doi: 10.1093/jnci/djy178.

33. 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.

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

35. Harrell FE, Jr., Margolis PA, Gove S, et al. Development of a clinical prediction model for an ordinal outcome: the World Health Organization Multicentre Study of Clinical Signs and Etiological agents of Pneumonia, Sepsis and Meningitis in Young Infants. WHO/ARI Young Infant Multicentre Study Group. *Stat Med* 1998;17:909-44.

36. Lawless JF, Singhal K. Efficient Screening of Nonnormal Regression-Models. *Biometrics* 1978;34:318-27.

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

38. 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.

39. 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.

40. Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958;45:562-5.

41. 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.

42. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. New York: Chapman & Hall; 1993.

43. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987;40:373-83.

44. Elixhauser A, Steiner C, Harris DR, Coffey RN. Comorbidity measures for use with administrative data. *Medical Care* 1998;36:8-27.

45. Simard M, Sirois C, Candas B. Validation of the Combined Comorbidity Index of Charlson and Elixhauser to Predict 30-Day Mortality Across ICD-9 and ICD-10. *Med Care* 2018;56:441-7.

46. NHS Digital. *Summary Hospital-level Mortality Indicator (SHMI): ICD-10 to SHMI diagnosis group lookup table*. Available from: https://digital.nhs.uk/data-and-information/publications/ci-hub/summary-hospital-level-mortality-indicator-shmi.

47. Healthcare Cost and Utilization Project. *Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses*. Available from: https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp.

48. Verweij PJ, Van Houwelingen HC. Penalized likelihood in Cox regression. *Stat Med* 1994;13:2427-36.

49. Cessie SL, Houwelingen JCV. Ridge Estimators in Logistic Regression. *Applied Statistics* 1992;41:191-201.

50. 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.

51. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974;19:716-23.

52. Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological* 1996;58:267-88.

53. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer; 2009.

54. Valderas JM, Starfield B, Sibbald B, Salisbury C, Roland M. Defining comorbidity: implications for understanding health and health services. *Ann Fam Med* 2009;7:357-63.

55. Fried LP, Ferrucci L, Darer J, Williamson JD, Anderson G. Untangling the concepts of disability, frailty, and comorbidity: implications for improved targeting and care. *J Gerontol A Biol Sci Med Sci* 2004;59:255-63.

56. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. *Biom J* 2018;60:431-49.

57. Sauerbrei W, Perperoglou A, Schmid M, et al. State of the art in selection of variables and functional forms in multivariable analysis-outstanding issues. *Diagn Progn Res* 2020;4:3. doi: 10.1186/s41512-020-00074-3.

58. Mantel N. Why Stepdown Procedures in Variable Selection. *Technometrics* 1970;12:621-5.

59. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Stat Med* 2007;26:5512-28.

60. Fortin M, Lapointe L, Hudon C, Vanasse A, Ntetu AL, Maltais D. Multimorbidity and quality of life in primary care: a systematic review. *Health Qual Life Outcomes* 2004;2:51.

61. Lutz W, Sanderson W, Scherbov S. The coming acceleration of global population ageing. *Nature* 2008;451:716-9.

62. Zhou B, Lu Y, Hajifathalian K, et al. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. *Lancet* 2016;387:1513-30.

63. Global Burden of Disease Cancer Collaboration. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncology* 2017;3:524-48.

Table 1. Descriptive statistics for outcome and predictor variables

	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)*	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†	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)
Median correlation between codes (IQR)‡	0.02 (0.01 to 0.03)	0.01 (0.00 to 0.02)	0.01 (0.00 to 0.02)

IQR=interquartile range. *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. †Relative frequency of each three-character code was at least 0.5% in the given population. ‡Median absolute values of Pearson correlation coefficients across all pairwise comparisons.

Table 2. Optimism-adjusted performance of the full and reduced models, as estimated from 500 bootstrap samples (with 95% confidence intervals)

	Acute myocardial infarction	Hip fracture	Major colorectal cancer surgery
Scaled Brier score (%)			
Full models*	34.6 (34.1 to 35.1)	22.8 (22.4 to 23.2)	17.1 (16.1 to 18.2)
Reduced models†	32.1 (31.6 to 32.6)	21.8 (21.4 to 22.2)	16.8 (15.8 to 17.8)
***c*-statistic**			
Full models	0.884 (0.882 to 0.886)	0.798 (0.795 to 0.800)	0.810 (0.804 to 0.817)
Reduced models	0.874 (0.872 to 0.876)	0.791 (0.788 to 0.793)	0.807 (0.801 to 0.813)
Integrated calibration index			
Full models	0.012 (0.011 to 0.013)	0.015 (0.014 to 0.017)	0.007 (0.005 to 0.009)
Reduced models	0.012 (0.011 to 0.013)	0.015 (0.013 to 0.016)	0.008 (0.006 to 0.009)
Calibration-in-the-large			
Full models	0.000 (-0.015 to 0.015)	0.000 (-0.013 to 0.013)	0.001 (-0.032 to 0.034)
Reduced models	0.032 (0.017 to 0.047)	0.013 (0.000 to 0.025)	0.025 (-0.008 to 0.058)
Calibration slope			
Full models	0.993 (0.982 to 1.004)	0.989 (0.978 to 1.001)	0.961 (0.935 to 0.987)
Reduced models	0.983 (0.972 to 0.994)	0.982 (0.970 to 0.993)	0.971 (0.946 to 0.997)

*Number of ICD-10 codes in full models (in column order): 202, 257, 209. †Number of codes in reduced models: 18, 33, 41.

Figure 1. Relative frequencies of included ICD-10 codes, by population

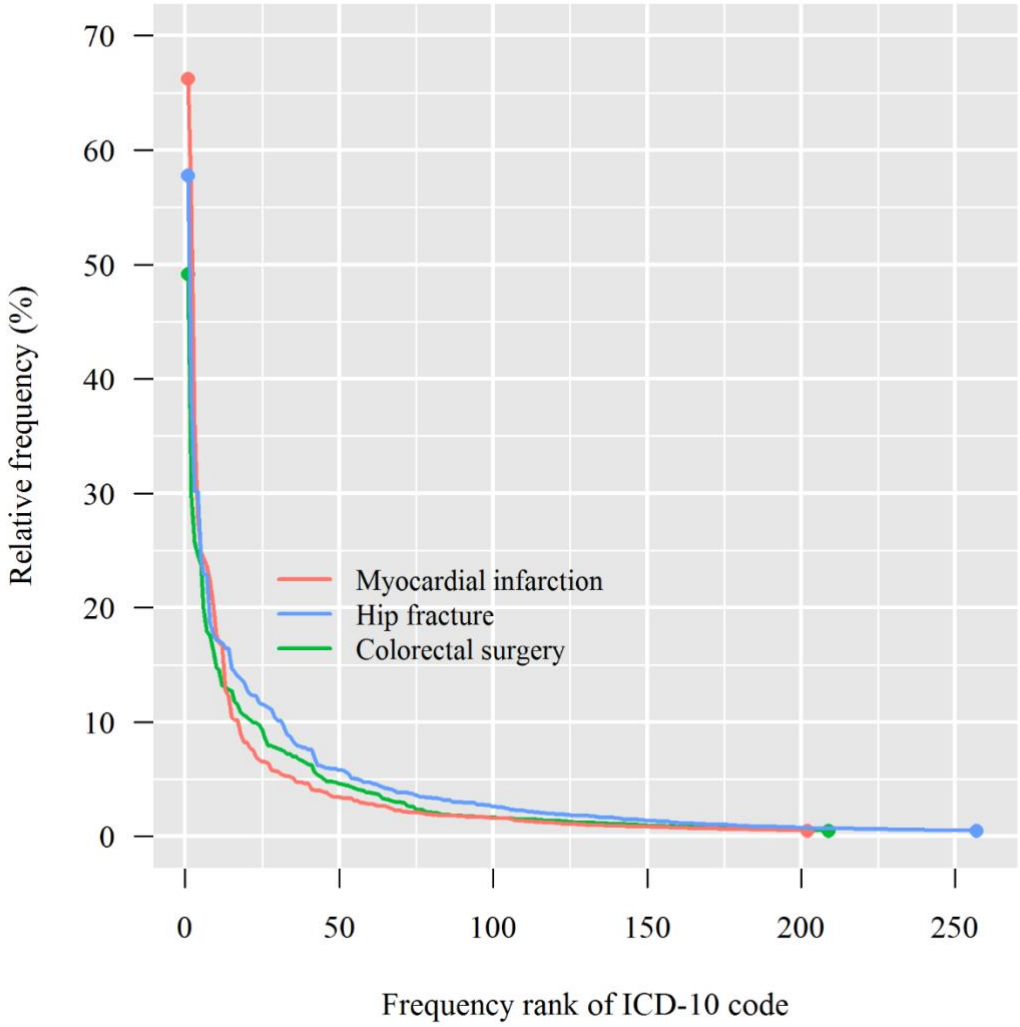
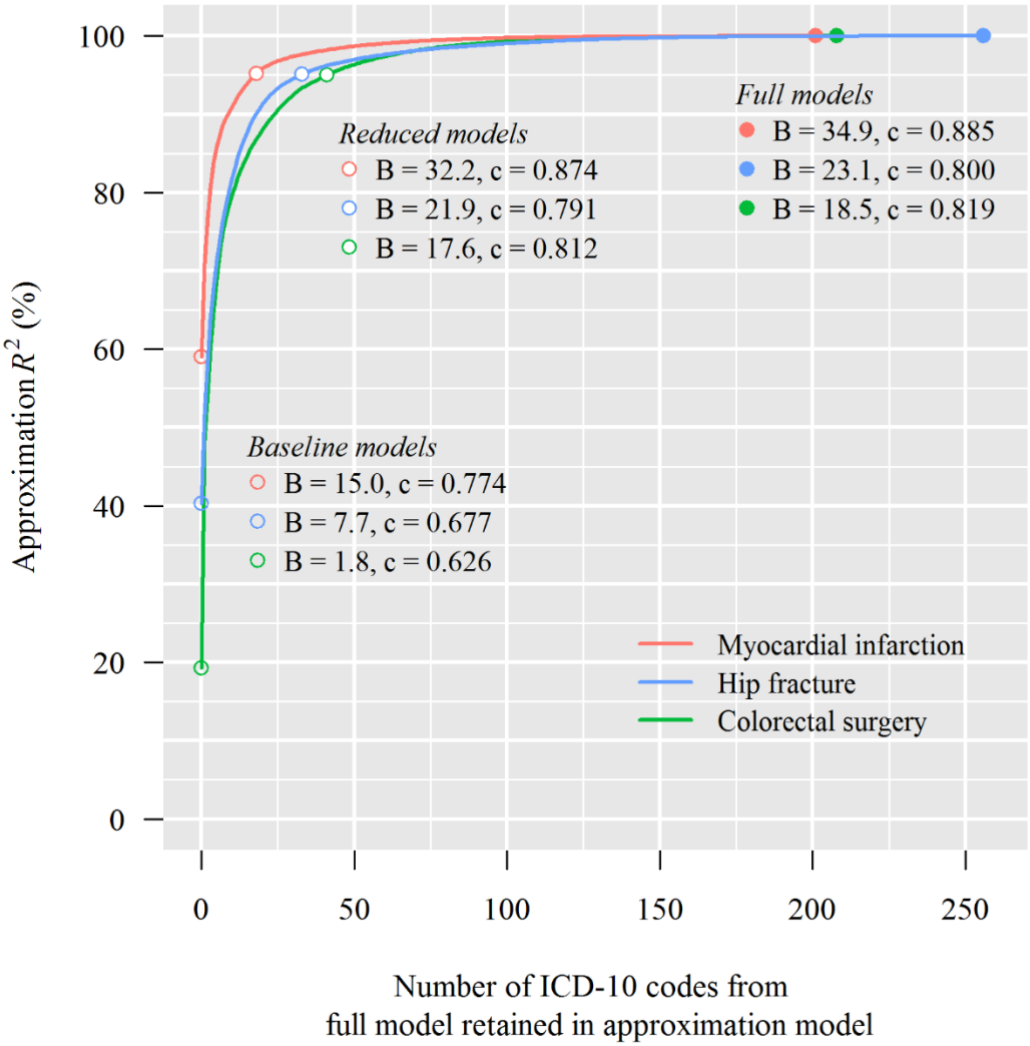
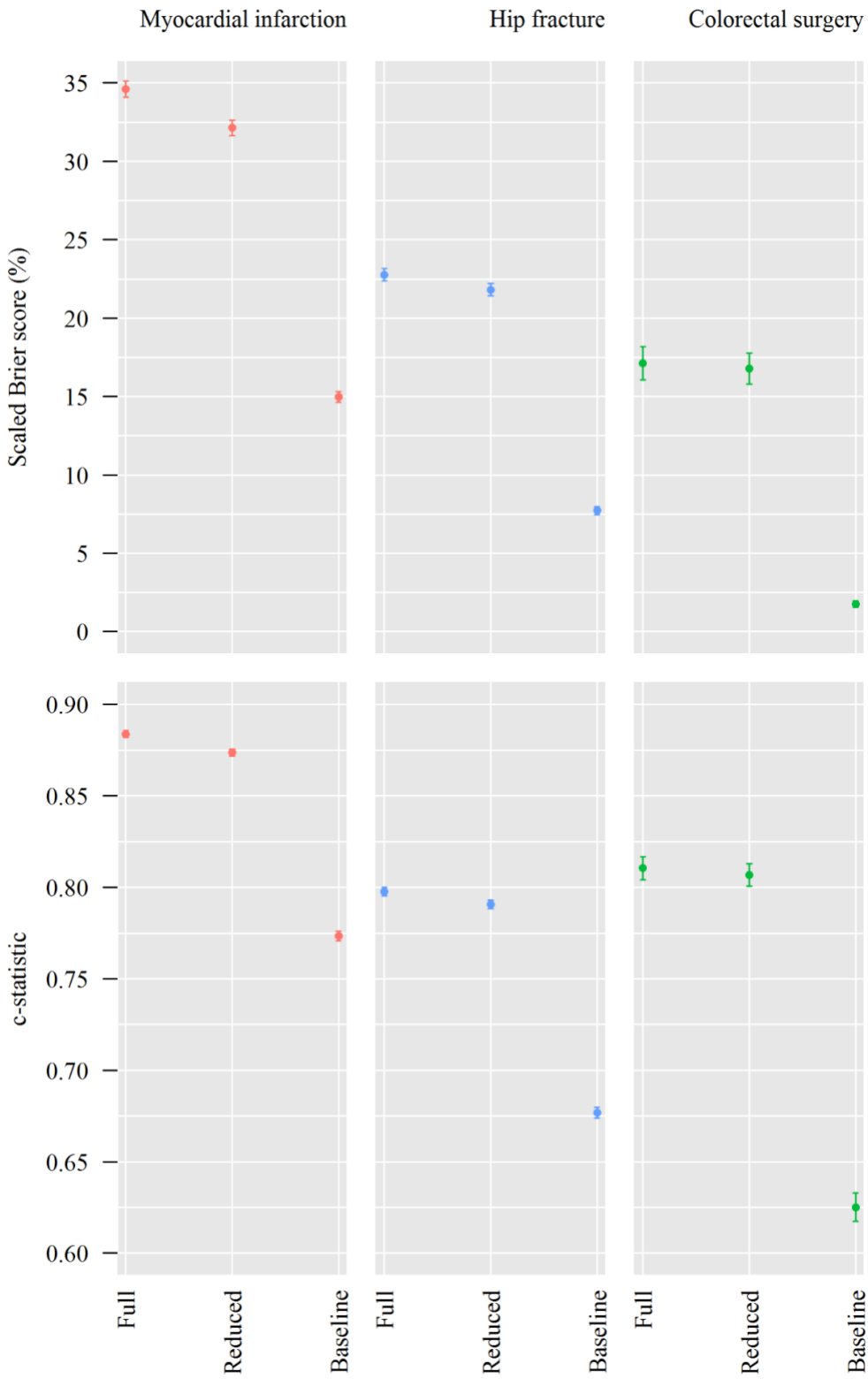


Figure 2. Percentage of variation in full model predictions explained by models with fewer ICD-10 codes (approximation R_2), and related scaled Brier scores (B%) and *c*-statistics (c)



The approximation R_2 equals the percentage of variation explained in the predictions from the full model. In each population, the full model included all ICD-10 codes recorded for at least 0.5% of patients. The 'approximation' models are the set of models with different numbers of codes removed from the full models. From the approximation models with an approximation R_2 of at least 95%, the 'reduced' model is the one with the fewest ICD-10 codes. 'Baseline' models include only age, sex, and socioeconomic status as predictors.

Figure 3. Optimism-adjusted scaled Brier scores and *c*-statistics of the full, reduced, and baseline models, as estimated from 500 bootstrap samples (with 95% confidence intervals)



Number of ICD-10 codes in full models (in column order): 202, 257, 209. Number of codes in reduced models: 18, 33, 41. Baseline models included only age, sex, and socioeconomic status as predictors.

Appendix

	Page
Methods for models based on the Charlson or Elixhauser conditions	2
Table A1. 20 most frequent ICD-10 codes by population	3
Table A2. Reduced models developed in each population	4
Table A3. Apparent performance of models estimated in the sensitivity analyses	7
Table A4. Results from alternative variable selection methods	8
Figure A1. Box plots of predictions from the full, reduced, and baseline models	9
Figure A2. Optimism-adjusted calibration curves for the reduced models	10
Figure A3. Consistency of ICD-10 codes included in the reduced models	11
Figure A4. Comparison with the performance of models based on the Charlson or Elixhauser conditions	12
Figure A5. Scatter plots of standard and penalised maximum likelihood estimates	13
Code to implement the model approximation approach in R	14

Methods for models based on the Charlson or Elixhauser conditions

An established list of ICD-10 codes[1] was used to identify the Charlson and Elixhauser conditions in the one-year look-back period. We estimated associations between the outcome and the Charlson or Elixhauser conditions using logistic regression. A binary predictor was modelled for each condition rather than weighted summary measures, as fitted weights differ across studies. We first compared the performance of the resulting models to the full and reduced models when all ICD-10 codes were eligible for inclusion in the analysis.

We repeated the comparison while excluding selected ICD-10 codes from patients' index episodes. These codes related to (i) acute conditions that could have been complications occurring after admission, and (ii) processes of care. The codes were chosen from the sets of codes recorded for at least 0.5% of patients in each population.

Codes excluded from index episodes	
A04 Other bacterial intestinal infections	A09 Diarrhea and gastroenteritis of infectious origin
A41 Other septicemia	B37 Candidiasis
B95 Streptococcus and staphylococcus as the cause of diseases classified to other chapters	B96 Other specified bacterial agents as the cause of diseases classified to other chapters
B98 Other specified infectious agents as the cause of diseases classified to other chapters	E89 Postprocedural endocrine and metabolic disorders, not elsewhere classified
F05 Delirium, not induced by alcohol and other psychoactive substances	I20 Angina pectoris
I21 Acute myocardial infarction	I26 Pulmonary embolism
I46 Cardiac arrest	I47 Paroxysmal tachycardia
I48 Atrial fibrillation and flutter	I49 Other cardiac arrhythmias
I50 Heart failure	I80 Phlebitis and thrombophlebitis
J18 Pneumonia, organism unspecified	J22 Unspecified acute lower respiratory infection
J69 Pneumonitis due to solids and liquids	J81 Pulmonary edema
J96 Respiratory failure, not elsewhere classified	K29 Gastritis and duodenitis
K65 Peritonitis	K91 Postprocedural disorders of digestive system, not elsewhere classified
L89 Decubitus ulcer	N17 Acute renal failure
N39 Other disorders of urinary system	R41 Other symptoms and signs involving cognitive functions and awareness
R57 Shock, not elsewhere classified	S00 Superficial injury of head
S01 Open wound of head	S06 Intracranial injury
S09 Other and unspecified injuries of head	S22 Fracture of rib(s), sternum, and thoracic spine
S32 Fracture of lumbar spine and pelvis	S40 Superficial injury of shoulder and upper arm
S42 Fracture of shoulder and upper arm	S50 Superficial injury of forearm
S51 Open wound of elbow and forearm	S52 Fracture of forearm
S60 Superficial injury of wrist and hand	S61 Open wound of wrist and hand
S62 Fracture at wrist and hand level	S70 Superficial injury of hip and thigh
S72 Fracture of femur	S80 Superficial injury of lower leg
S81 Open wound of lower leg	S82 Fracture of lower leg, including ankle
T81 Complications of procedures, not elsewhere classified	T82 Complications of cardiac and vascular prosthetic devices, implants, and grafts
T83 Complications of genitourinary prosthetic devices	T84 Complications of internal orthopedic prosthetic devices
W01 Fall on same level from slipping, tripping, and stumbling	W03 Other fall on same level due to collision with another person
W06 Fall involving bed	W07 Fall involving chair
W10 Fall on and from stairs and steps	W18 Other fall on same level
W19 Unspecified fall	W22 Striking against or struck by other objects
Y43 Adverse effects in therapeutic use, primarily systemic agents	Y45 Adverse effects in therapeutic use, analgesics, antipyretics, and anti-inflammatory drugs
Y60 Unintentional cut, puncture, perforation, or hemorrhage during surgical and medical care	Y83 Surgical operation and other surgical procedures as the cause of abnormal reaction of the patient, or of later complication
Y84 Other medical procedures as the cause of abnormal reaction of the patient, or of later complication	Y95 Nosocomial condition
Z02 Examination and encounter for administrative purposes	Z08 Follow-up examination after treatment for malignancies
Z09 Follow-up examination after treatment for conditions other than malignant neoplasms	Z12 Special screening examination for neoplasms
Z13 Special screening examination for other diseases	Z22 Carrier of infectious diseases
Z45 Adjustment and management of implanted device	Z50 Care involving use of rehabilitation procedures
Z53 Persons encountering services for procedures, not carried out	Z75 Problems related to medical facilities and other health care
Z95 Presence of cardiac and vascular implants and grafts	Z96 Presence of other functional implants
Z98 Other postsurgical states	Z99 Dependence on enabling machines and devices

1. 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 A1. 20 most frequent ICD-10 codes by population

		Acute myocardial infarction			Hip fracture			Major colorectal cancer surgery	
	Code	Title	%	Code	Title	%	Code	Title	%
1	I25	Chronic ischemic heart disease	66.2	I10	Essential (primary) hypertension	57.8	I10	Essential (primary) hypertension	49.1
2	I10	Essential (primary) hypertension	55.2	W19	Unspecified fall	37.8	Z86	Personal history of certain other diseases	29.6
3	Z86	Personal history of certain other diseases	36.0	W01	Fall on same level from slipping, tripping, and stumbling	30.2	K57	Diverticular disease of intestine	25.7
4	E78	Disorders of lipoprotein metabolism and other lipidemias	27.8	Z86	Personal history of certain other diseases	30.2	K63	Other diseases of intestine	24.6
5	F17	Mental and behavioural disorders due to use of tobacco	24.9	R29	Other symptoms and signs involving the nervous and musculoskeletal systems	24.3	Z92	Personal history of medical treatment	23.8
6	E11	Non-insulin-dependent diabetes mellitus	24.3	I48	Atrial fibrillation and flutter	23.0	D12	Benign neoplasm of colon, rectum, anus, and anal canal	19.9
7	Z92	Personal history of medical treatment	23.7	Z92	Personal history of medical treatment	22.9	D50	Iron deficiency anaemia	18.0
8	I50	Heart failure	22.3	N39	Other disorders of urinary system	18.7	C77	Secondary and unspecified malignant neoplasm of lymph nodes	17.6
9	Z95	Presence of cardiac and vascular implants and grafts	20.3	I25	Chronic ischemic heart disease	17.7	E11	Non-insulin-dependent diabetes mellitus	16.3
10	Z82	Family history of certain disabilities and chronic diseases leading to disablement	17.6	N18	Chronic renal failure	17.2	K66	Other disorders of peritoneum	14.8
11	I20	Angina pectoris	17.0	M81	Osteoporosis without pathological fracture	17.1	D64	Other anaemias	14.5
12	I48	Atrial fibrillation and flutter	16.7	W18	Other fall on same level	16.9	Z85	Personal history of malignant neoplasm	13.2
13	I51	Complications and ill-defined descriptions of heart disease	12.8	F03	Unspecified dementia	16.5	Y83	Surgical operation and other surgical procedures as cause of abnormal reaction	13.1
14	N18	Chronic renal failure	12.2	E11	Non-insulin-dependent diabetes mellitus	16.4	E78	Disorders of lipoprotein metabolism and other lipidemias	12.8
15	J44	Other chronic obstructive pulmonary disease	10.4	N17	Acute renal failure	14.7	K56	Paralytic ileus and intestinal obstruction without hernia	12.7
16	Z88	Personal history of allergy to drugs, medicaments and biological substances	10.2	J18	Pneumonia, organism unspecified	14.4	I48	Atrial fibrillation and flutter	11.9
17	N17	Acute renal failure	10.2	J44	Other chronic obstructive pulmonary disease	14.0	I25	Chronic ischemic heart disease	11.5
18	I44	Atrioventricular and left bundle-branch block	8.9	Z96	Presence of other functional implants	13.8	T81	Complications of procedures, not elsewhere classified	10.8
19	J45	Asthma	8.3	D64	Other anaemias	13.5	K62	Other diseases of anus and rectum	10.6
20	E66	Obesity	8.2	E78	Disorders of lipoprotein metabolism and other lipidemias	12.8	K44	Diaphragmatic hernia	10.5

Table A2. Reduced models developed in each population

Acute myocardial infarction

	Model coefficient	95% confidence interval
Intercept*	-2.49	-2.53 to -2.45
Age (per 5-year increase)	0.39	0.38 to 0.40
Male (versus female)	0.04	0.01 to 0.08
Socioeconomic status (per decile increase in national rank; less deprived)	-0.03	-0.04 to -0.03
ICD-10 codes†		
I46 Cardiac arrest	1.99	1.94 to 2.05
N17 Acute renal failure	0.67	0.63 to 0.71
Z51 Other medical care (e.g. chemotherapy)	1.61	1.54 to 1.69
I50 Heart failure	0.65	0.62 to 0.68
R57 Shock, not elsewhere classified	1.89	1.81 to 1.98
Z82 Family history of certain disabilities and chronic diseases leading to disablement	-0.47	-0.53 to -0.41
C78 Secondary malignant neoplasm of respiratory and digestive organs	2.46	2.28 to 2.63
E87 Other disorders of fluid, electrolyte, and acid-base balance	0.53	0.48 to 0.58
I25 Chronic ischemic heart disease	-0.31	-0.34 to -0.28
Z99 Dependence on enabling machines and devices, not elsewhere classified	0.92	0.83 to 1.01
G93 Other disorders of brain	1.64	1.51 to 1.76
J44 Other chronic obstructive pulmonary disease	0.42	0.38 to 0.46
F03 Unspecified dementia	0.84	0.78 to 0.91
C34 Malignant neoplasm of bronchus and lung	1.92	1.76 to 2.08
E78 Disorders of lipoprotein metabolism and other lipidemias	-0.33	-0.36 to -0.29
E11 Non-insulin-dependent diabetes mellitus	0.26	0.23 to 0.30
J18 Pneumonia, organism unspecified	0.46	0.42 to 0.51
N18 Chronic renal failure	0.37	0.33 to 0.41

*Predicted log-odds of death for 70-year old females with none of the included ICD-10 codes and who live in the most deprived area nationally. †In order of decreasing importance (top to bottom), based on the order in which codes were removed from the approximation models.

Hip fracture

	Model coefficient	95% confidence interval
Intercept*	-2.82	-2.86 to -2.77
Age (per 5-year increase)	0.35	0.34 to 0.36
Male (versus female)	0.51	0.49 to 0.54
Socioeconomic status (per decile increase in national rank; less deprived)	-0.02	-0.02 to -0.01
ICD-10 codes†		
Z51 Other medical care (e.g. chemotherapy)	1.35	1.29 to 1.42
J18 Pneumonia, organism unspecified	0.38	0.35 to 0.42
F03 Unspecified dementia	0.65	0.62 to 0.68
I50 Heart failure	0.47	0.43 to 0.50
C78 Secondary malignant neoplasm of respiratory and digestive organs	1.99	1.84 to 2.13
I46 Cardiac arrest	2.40	2.23 to 2.56
C34 Malignant neoplasm of bronchus and lung	1.63	1.50 to 1.76
F01 Vascular dementia	0.66	0.61 to 0.71
G30 Alzheimer's disease	0.57	0.53 to 0.61
J44 Other chronic obstructive pulmonary disease	0.41	0.38 to 0.45
W01 Fall on same level from slipping, tripping, and stumbling	-0.25	-0.28 to -0.22
N17 Acute renal failure	0.23	0.19 to 0.26
C79 Secondary malignant neoplasm of other sites	1.48	1.35 to 1.60
I48 Atrial fibrillation and flutter	0.30	0.28 to 0.33
W10 Fall on and from stairs and steps	-0.43	-0.50 to -0.37
J96 Respiratory failure, not elsewhere classified	0.59	0.52 to 0.66
L89 Decubitus ulcer	0.43	0.38 to 0.48
J69 Pneumonitis due to solids and liquids	0.73	0.63 to 0.82
N18 Chronic renal failure	0.28	0.24 to 0.31
G20 Parkinson's disease	0.47	0.41 to 0.53
Z96 Presence of other functional implants	-0.26	-0.30 to -0.23
W19 Unspecified fall	0.21	0.18 to 0.24
E87 Other disorders of fluid, electrolyte, and acid-base balance	0.24	0.20 to 0.28
E78 Disorders of lipoprotein metabolism and other lipidemias	-0.21	-0.25 to -0.17
Z99 Dependence on enabling machines and devices, not elsewhere classified	0.58	0.49 to 0.68
R13 Dysphagia	0.46	0.38 to 0.54
J90 Pleural effusion, not elsewhere classified	0.33	0.27 to 0.39
I10 Essential (primary) hypertension	-0.13	-0.16 to -0.11
E86 Volume depletion	0.26	0.21 to 0.31
J84 Other interstitial pulmonary diseases	0.59	0.48 to 0.70
I21 Acute myocardial infarction	0.43	0.35 to 0.51
Z50 Care involving use of rehabilitation procedures	-0.24	-0.30 to -0.19
A41 Other septicemia	0.33	0.27 to 0.40

*Predicted log-odds of death for 70-year old females with none of the included ICD-10 codes and who live in the most deprived area nationally. †In order of decreasing importance (top to bottom), based on the order in which codes were removed from the approximation models.

Major surgery for colorectal cancer

	Model coefficient	95% confidence interval
Intercept*	-2.92	-3.01 to -2.82
Age (per 5-year increase)	0.20	0.19 to 0.22
Male (versus female)	-0.04	-0.11 to 0.02
Socioeconomic status (per decile increase in national rank; less deprived)	-0.04	-0.05 to -0.02
ICD-10 codes†		
C78 Secondary malignant neoplasm of respiratory and digestive organs	1.70	1.62 to 1.78
C77 Secondary and unspecified malignant neoplasm of lymph nodes	0.67	0.60 to 0.74
I46 Cardiac arrest	3.09	2.80 to 3.37
K65 Peritonitis	1.00	0.86 to 1.14
D12 Benign neoplasm of colon, rectum, anus, and anal canal	-0.49	-0.59 to -0.40
N17 Acute renal failure	0.44	0.34 to 0.54
E87 Other disorders of fluid, electrolyte, and acid-base balance	0.44	0.34 to 0.54
Z51 Other medical care (e.g. chemotherapy)	0.71	0.60 to 0.83
J18 Pneumonia, organism unspecified	0.34	0.24 to 0.45
F17 Mental and behavioral disorders due to use of tobacco	0.41	0.31 to 0.51
K62 Other diseases of anus and rectum	-0.37	-0.49 to -0.24
I48 Atrial fibrillation and flutter	0.39	0.30 to 0.48
C79 Secondary malignant neoplasm of other sites	0.89	0.71 to 1.07
Z12 Special screening examination for neoplasms	-0.81	-1.24 to -0.39
R10 Abdominal and pelvic pain	0.37	0.25 to 0.49
A41 Other septicemia	0.49	0.36 to 0.63
K55 Vascular disorders of intestine	0.85	0.65 to 1.05
R59 Enlarged lymph nodes	0.75	0.55 to 0.96
E66 Obesity	-0.28	-0.41 to -0.15
R18 Ascites	0.76	0.55 to 0.96
K66 Other disorders of peritoneum	-0.23	-0.32 to -0.13
J96 Respiratory failure, not elsewhere classified	0.60	0.42 to 0.77
R63 Symptoms and signs concerning food and fluid intake	0.37	0.24 to 0.51
I50 Heart failure	0.39	0.26 to 0.53
Z92 Personal history of medical treatment	-0.19	-0.26 to -0.11
R29 Other symptoms and signs involving the nervous and musculoskeletal systems	0.49	0.31 to 0.66
L89 Decubitus ulcer	0.72	0.49 to 0.94
J84 Other interstitial pulmonary diseases	0.93	0.64 to 1.23
N18 Chronic renal failure	0.26	0.15 to 0.36
J44 Other chronic obstructive pulmonary disease	0.27	0.16 to 0.37
K57 Diverticular disease of intestine	-0.15	-0.22 to -0.07
Z08 Follow-up examination after treatment for malignant neoplasms	-0.61	-1.00 to -0.22
M13 Other arthritis	-0.28	-0.45 to -0.11
Z99 Dependence of enabling machines and devices, not elsewhere classified	0.72	0.41 to 1.03
K51 Ulcerative colitis	0.48	0.21 to 0.75
G20 Parkinson's disease	0.70	0.39 to 1.02
N13 Obstructive and reflux uropathy	0.52	0.29 to 0.75
Y83 Surgical operation and other surgical procedures as the cause of abnormal reaction	-0.19	-0.28 to -0.09
K56 Paralytic ileus and intestinal obstruction without hernia	0.16	0.08 to 0.25
R41 Other symptoms and signs involving cognitive functions and awareness	0.39	0.21 to 0.57
E78 Disorders of lipoprotein metabolism and other lipidemias	-0.15	-0.25 to -0.06

*Predicted log-odds of death for 70-year old females with none of the included ICD-10 codes and who live in the most deprived area nationally. †In order of decreasing importance (top to bottom), based on the order in which codes were removed from the approximation models.

Table A3. Apparent performance of models estimated in the sensitivity analyses

	Acute myocardial infarction	Hip fracture	Major colorectal cancer surgery
Scaled Brier score (%)			
Frequency threshold of 1%	33.1	21.4	16.3
Clinical Classification Software groups	32.8	20.7	15.1
Look-back using index episode only	33.9	23.7	17.8
Look-back period of 3 years	34.4	22.5	18.4
Charlson and Elixhauser combined	24.5	17.8	8.7
***c*-statistic**			
Frequency threshold of 1%	0.878	0.790	0.808
Clinical Classification Software groups	0.880	0.788	0.804
Look-back using index episode only	0.878	0.799	0.808
Look-back period of 3 years	0.884	0.798	0.820
Charlson and Elixhauser combined	0.843	0.770	0.758
Integrated calibration index			
Frequency threshold of 1%	0.012	0.016	0.007
Clinical Classification Software groups	0.014	0.016	0.006
Look-back using index episode only	0.010	0.012	0.008
Look-back period of 3 years	0.013	0.016	0.008
Charlson and Elixhauser combined	0.015	0.017	0.000

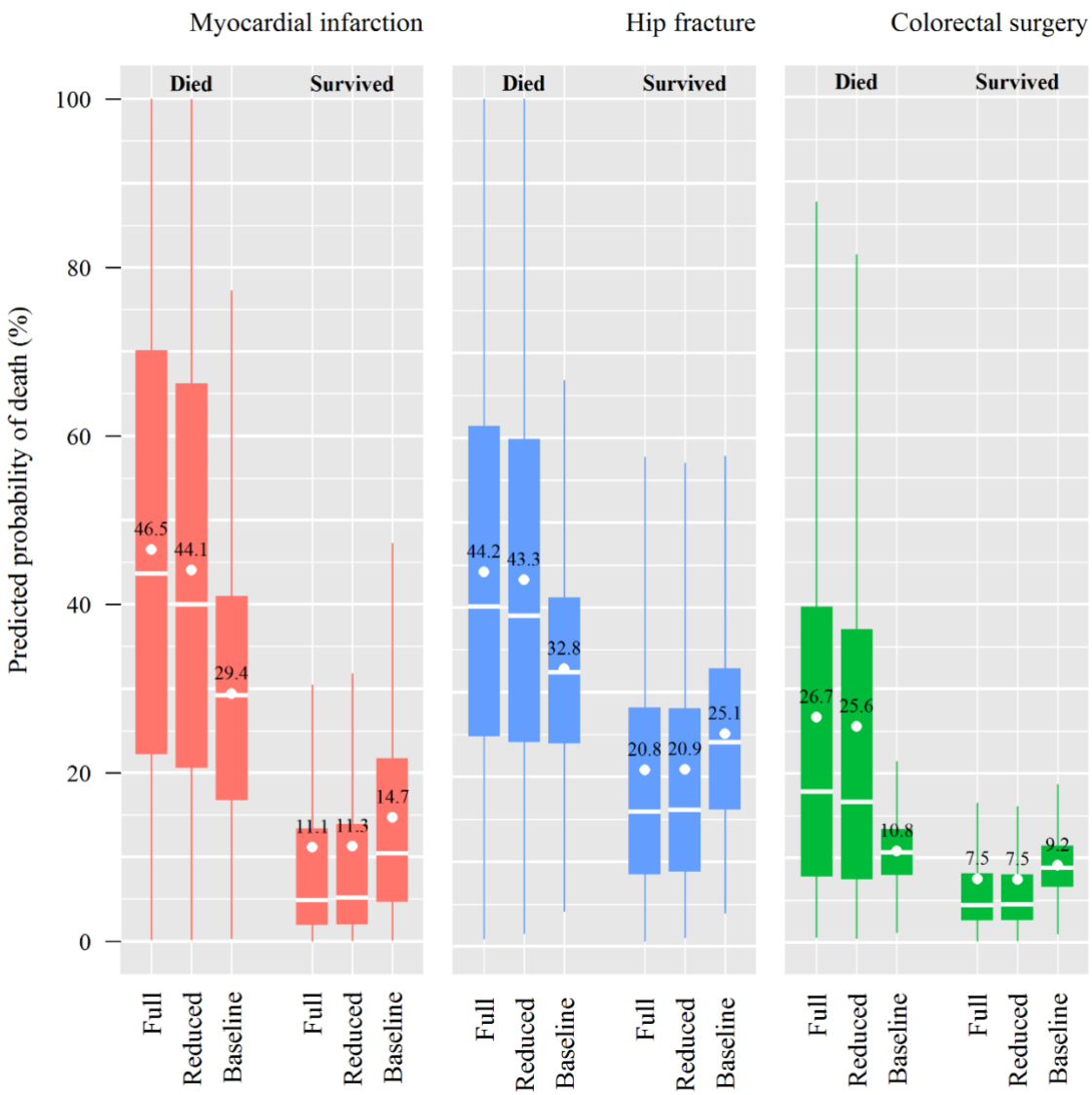
All values of calibration-in-the-large and calibration slope equalled zero and one, respectively, by definition.

Table A4. Number of ICD codes included in the final models and performance measures, by modelling approach and population, in the original data

	Acute myocardial infarction	Hip fracture	Major colorectal cancer surgery
Number of ICD codes			
Full models	202	257	209
Reduced models	18	33	41
Backwards elimination using AIC	137	169	85
Backwards elimination using BIC	91	99	51
Lasso	188	221	121
Scaled Brier score (%)			
Full models	34.9	23.1	18.5
Reduced models	32.2	21.9	17.6
Backwards elimination using AIC	34.9	23.1	18.3
Backwards elimination using BIC	34.8	22.8	17.9
Lasso	34.9	23.1	18.2
***c*-statistic**			
Full models	0.885	0.800	0.819
Reduced models	0.874	0.791	0.812
Backwards elimination using AIC	0.885	0.799	0.818
Backwards elimination using BIC	0.884	0.798	0.815
Lasso	0.885	0.799	0.818

AIC=Akaike Information Criterion. BIC=Bayesian Information Criterion. ICD=International Classification of Diseases.

Figure A1. Box plots of predictions from the full, reduced, and baseline models in the original data, by observed outcome status and population



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 A2. Optimism-adjusted calibration curves for the reduced models, by population, as estimated from loess smoothers in 500 bootstrap samples (shown with line of perfect calibration)

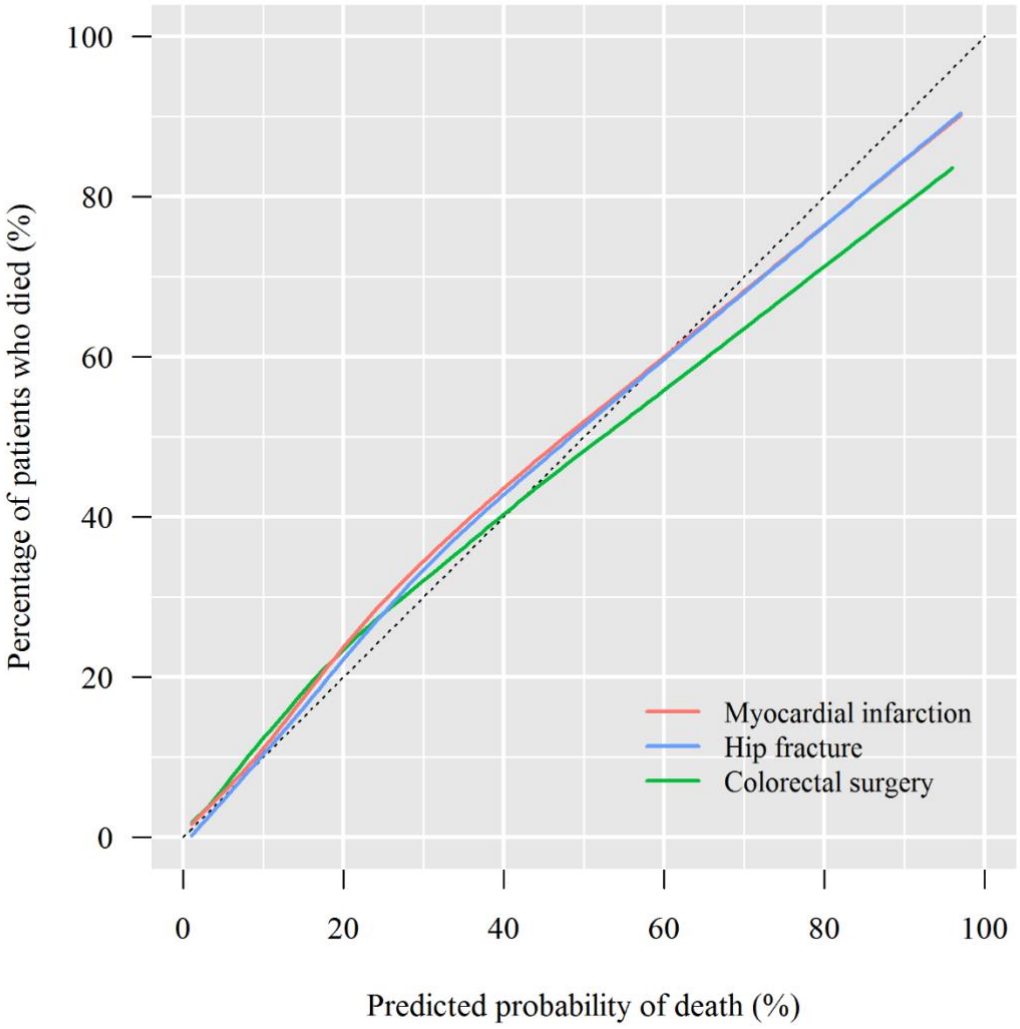
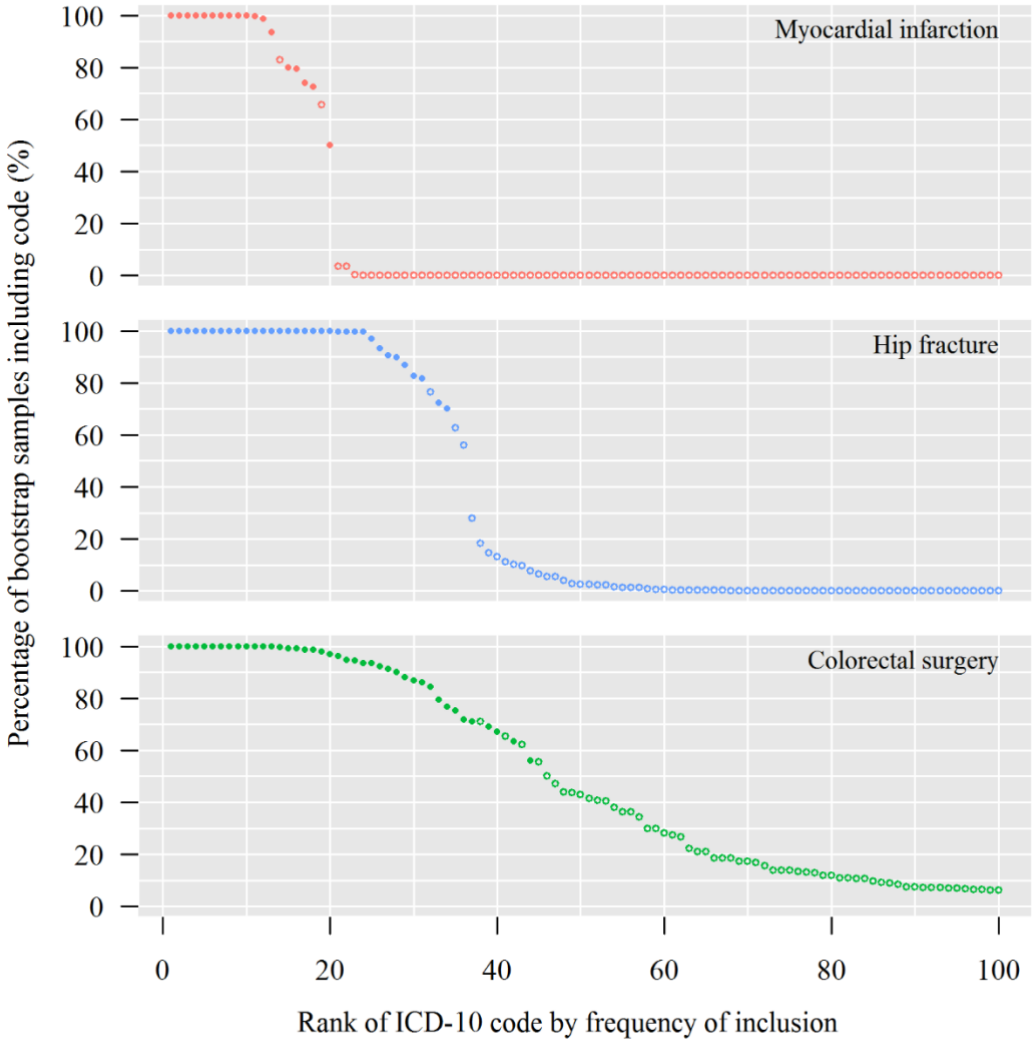


Figure A3. Consistency of ICD-10 codes included in the reduced models across 500 bootstrap samples



Filled points relate to ICD-10 codes that were included in the reduced models in the original data; unfilled points relate to ICD-10 codes that were not included. Only the 100 most frequently included codes shown.

Figure A4. Comparison with the performance of models based on the Charlson or Elixhauser conditions, when all ICD-10 codes were eligible for inclusion (filled points) and when excluding selected codes (unfilled)

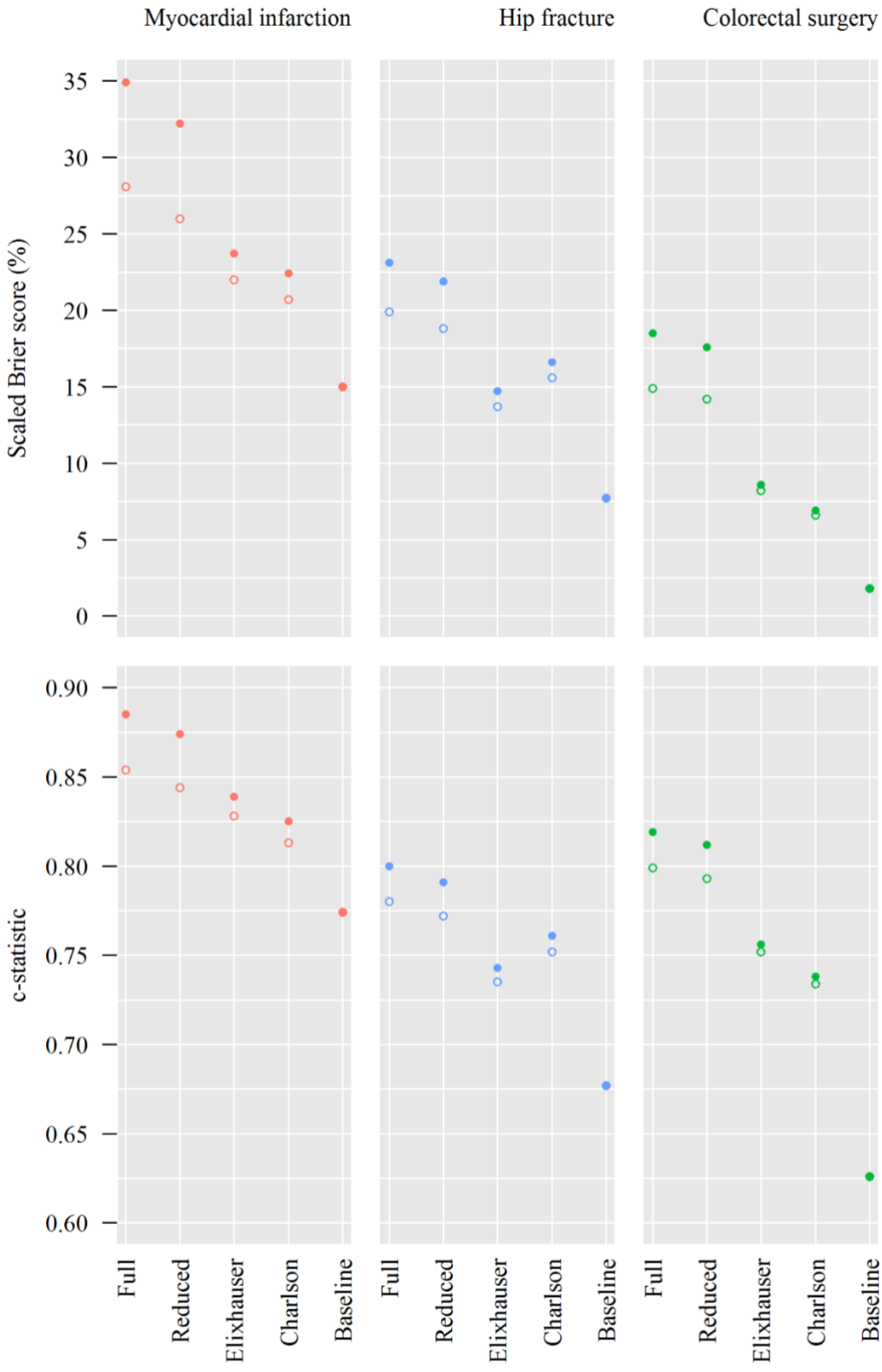
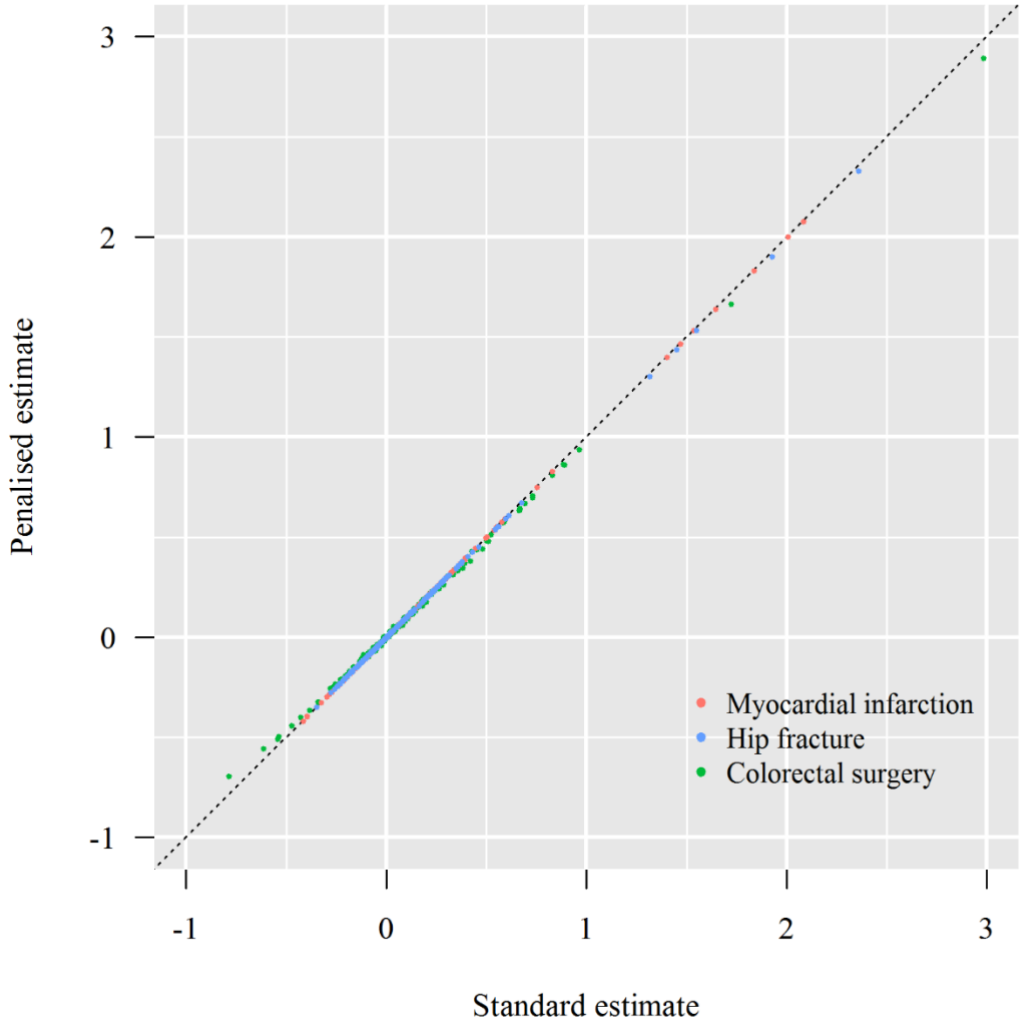


Figure A5. Scatter plots of coefficients estimated for the full models using the standard maximum likelihood (x-axis) and penalised maximum likelihood (y-axis) in each population



The penalty factor was chosen to maximise a modified Akaike's Information Criterion (AIC).

Code to implement the model approximation approach in R

```
# Description of dataset ----------------------------------------------------


# The data are stored in R as 'df_original' with class 'data.frame'.
# The outcome variable is 'mort'. The predictors are 'age', 'sex',
# 'imd', and a binary predictor for each included ICD-10 code.
# The data frame is structured as:

#    mort | age | sex |   imd | A04 | B37 | C34 | D47 | E03 | ...
#   ------------------------------------------------------------------
# 1     1 |  62 |   0 | 17980 |   0 |   1 |   0 |   0 |   0 | ...
# 2     0 |  71 |   1 |  1284 |   0 |   0 |   0 |   0 |   0 | ...
# 3     0 |  84 |   0 |  6421 |   0 |   0 |   1 |   1 |   0 | ...



# Load required packages and functions ---------------------------------------


library(rms)

# The val.prob.ci function can be downloaded from:
# www.clinicalpredictionmodels.org/doku.php?id=rcode_and_data:chapter15

source('C:/val.prob.ci.June09.r')



# Fit models and calculate apparent performance ------------------------------


# Code adapted from: Harrell FE, Jr. Regression Modelling Strategies:
# With Applications to Linear Models, Logistic and Ordinal Regression,
# and Survival Analysis. 2nd ed. Cham: Springer; 2015.


# fit logistic regression model with all predictors (full model)
full_model <- lrm(mort ~ ., data = df_original, x = TRUE, y = TRUE, maxit = 1000)

# store the predicted log-odds of outcome
full_preds <- predict(full_model)

# fit ordinary least squares regression model of predictions
full_ols <- ols(full_preds ~ ., data = df_original[, -1], sigma = 1)

# fit approximation models and calculate R-squared values
approx_models <- fastbw(full_ols, aics = 1000000, type = "individual", force = c(1:3))
# last argument forces age, sex, and imd into all models

# store predictors from smallest model with R-squared >= 0.95
vars <- rev(names(approx_models$result[, "R2"][approx_models$result[, "R2"] < 0.95]))

# fit model with these predictors (reduced model)
formula <- as.formula(paste("full_preds", paste(c("age", "sex", "imd", vars), collapse = " + "), sep = " ~ "))
reduced_model <- ols(formula, data = df_original, x = TRUE)
```

```
# calculate correct standard errors for reduced model
V <- vcov(full_model, regcoef.only = TRUE)
X <- cbind('Intercept' = 1, full_model$x)
x <- cbind('Intercept' = 1, reduced_model$x)
w <- solve(t(x) %*% x, t(x)) %*% X
v <- w %*% V %*% t(w)
reduced_model$var <- v

# create table of reduced model results
results_table <- data.frame('Code' = names(reduced_model$coefficients), 'Coefficient' =
reduced_model$coefficients, 'SE' = sqrt(diag(v)))
results_table$lower_95CI <- results_table$Coefficient - (1.96 * results_table$SE)
results_table$upper_95CI <- results_table$Coefficient + (1.96 * results_table$SE)

# store the predicted log-odds of outcome from reduced model
reduced_preds <- predict(reduced_model)

# calculate apparent performance of full model
full_app_perf <- val.prob.ci(plogis(full_preds), df_original$mort)[c('Brier scaled', 'C (ROC)', 'Eavg', 'Intercept',
'Slope')]

# calculate apparent performance of reduced model
reduced_app_perf <- val.prob.ci(plogis(reduced_preds), df_original$mort)[c('Brier scaled', 'C (ROC)', 'Eavg',
'Intercept', 'Slope')]



# Adjust performance values for optimism ----------------------------------


# create boostrap function
boot_function <- function(data, n_samples) {

  results <- matrix(nrow = n_samples, ncol = 10)

  for (i in 1:n_samples) {

    # create bootstrap sample
    df_sample <- data[sample(1:nrow(data), replace = TRUE), ]

    # fit full and reduced models (as in original data)
    full_model <- lrm(mort ~ ., data = df_sample, x = TRUE, y = TRUE, maxit = 1000)
    full_preds <- predict(full_model)
    full_ols <- ols(full_preds ~ ., data = df_sample[, -1], sigma = 1)
    approx_models <- fastbw(full_ols, aics = 1000000, type = "individual", force = c(1:3))
    vars <- names(approx_models$result[, "R2"][approx_models$result[, "R2"] < 0.95])
    formula <- as.formula(paste("full_preds", paste(c("age", "sex", "imd", vars), collapse = " + "), sep = " ~ "))
    reduced_model <- ols(formula, data = df_sample)
    reduced_preds <- predict(reduced_model)

    # calculate apparent performance of full and reduced models
    full_app_perf <- val.prob.ci(plogis(full_preds), df_sample$mort)
    reduced_app_perf <- val.prob.ci(plogis(reduced_preds), df_sample$mort)

    # calculate performance of full and reduced models in original data
    full_test_perf <- val.prob.ci(predict(full_model, newdata = data, type = "fitted"), data$mort)
    reduced_test_perf <- val.prob.ci(plogis(predict(reduced_model, newdata = data)), data$mort)

    # store values of optimism in matrix
    full_optimism <- full_app_perf - full_test_perf
```

```
    reduced_optimism <- reduced_app_perf - reduced_test_perf

    # store optimism values in matrix
    results[i, 1:5]  <- full_optimism[c('Brier scaled', 'C (ROC)', 'Eavg', 'Intercept', 'Slope')]
    results[i, 6:10] <- reduced_optimism[c('Brier scaled', 'C (ROC)', 'Eavg', 'Intercept', 'Slope')]

    print(i)

  }

  results <- as.data.frame(results)
  colnames(results) <- c('full_Brier', 'full_c', 'full_ICI', 'full_CITL', 'full_slope',
                         'reduced_Brier', 'reduced_c', 'reduced_ICI', 'reduced_CITL', 'reduced_slope')
  return(results)

}

# calculate optimism in performance in each bootstrap sample
optimism <- boot_function(data = df_original, n_samples = 500)

# calculate mean optimism across bootstrap samples
mean_optimism <- colMeans(optimism)

# calculate optimism-adjusted performance values
full_adj_perf <- full_app_perf - mean_optimism[1:5]
reduced_adj_perf <- reduced_app_perf - mean_optimism[6:10]
```