Journal of Clinical Epidemiology 128 (2020) 20–28

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

A novel approach selected small sets of diagnosis codes with high prediction performance in large healthcare datasets

Thomas E. Cowling[a,b,*], David A. Cromwell[a,b], Linda D. Sharples[c], 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 Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel St, London WC1E 7HT, UK

Accepted 5 August 2020; Published online 8 August 2020

Abstract

Objectives: The objective of the study was 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: This was a modeling 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 2015–2017. 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% (5,273) 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% confidence interval (CI) 0.882, 0.886), 0.798 (0.795, 0.800), and 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), and 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. © 2020 Elsevier Inc. All rights reserved.

Keywords: Big data; Electronic medical records; International Classification of Diseases; ICD-10; Comorbidity; Multimorbidity; Prognosis; Statistical models; Variable selection

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 randomized 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 aging and greater burdens of noncommunicable disease, models of morbidity are likely to be increasingly important [10,11].

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

Conflicts of interest statement: None declared.

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

* Corresponding author. Tel.: +44 020 7927 2151.
E-mail address: thomas.cowling@lshtm.ac.uk (T.E. Cowling).

https://doi.org/10.1016/j.jclinepi.2020.08.001
0895-4356/© 2020 Elsevier Inc. All rights reserved.

What is new?

Key findings

- Our approach selected small sets of diagnosis codes that predicted 1-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.

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 modeling 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 carried out 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 modeling context, using linked national datasets of routine hospital and mortality data in England.

2. Methods

2.1. Study populations

We analyzed 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). Patients with MI (I21-22 [19,20]) and HF (S72.0-S72.2 [21,22]) were identified from the ICD-10 codes recorded as the primary diagnosis in the first episode of each admission. Patients with CS 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, coexisting conditions, and mortality. We included patients with MI and CS aged 18 years or older and patients with HF 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 analyzed 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. In addition, 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 analyzed the first three characters of these codes (excluding fourth characters) as coding choices at this level will be less variable than with four characters [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 analyzed, 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 1,000 to 3,000 residents in each of 32,482 areas) [34]. We excluded patients with missing data (1.2%; 5,346/431,626).

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 equaled 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 et al. [15], when all predictors are uncorrelated and standardized, 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). Modeling nonlinear 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 equaled the mean of squared differences between predicted probabilities of death and observed outcomes. We scaled these scores from 0% to 100% (0% if noninformative and 100% if perfect) [38].

To assess discrimination, we calculated the *c*-statistic. This equaled the probability that a randomly chosen patient who died had a greater predicted probability of death than a randomly chosen patient who did not 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 modeling 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 contextualize model performance, we compared the results with 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 3-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 penalized maximum likelihood estimation was used [17,48,49].

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

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: backward 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 1 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, 8,445 unique four-character ICD codes and 1,857 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 (5,273/212; CS).

Most included ICD codes had low frequencies (Fig. 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% (Fig. 2).

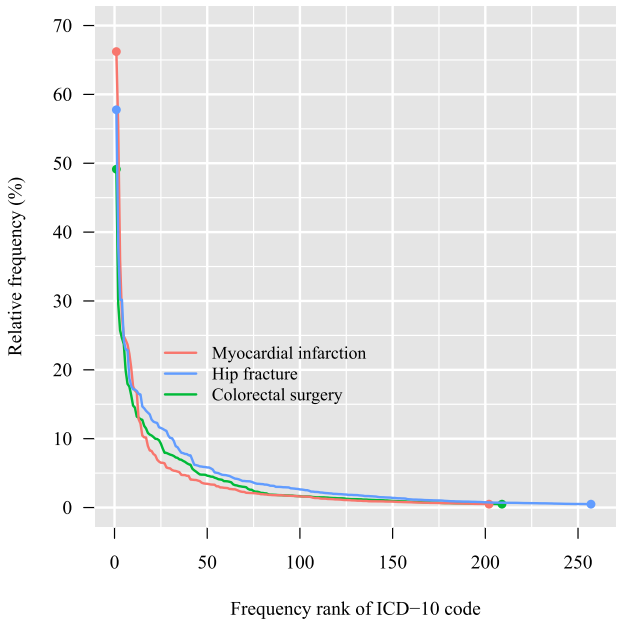

Fig. 1. Relative frequencies of included ICD-10 codes, by population.

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 (Fig. 2; Fig. 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, Fig. 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% confidence interval 0.935–0.987), indicating that model predictions were slightly too extreme (Fig. A2).

The codes included in the reduced models were reasonably stable across bootstrap samples: 13 of 18 (MI), 27 of 33 (HF), and 28 of 41 (CS) codes were selected in ≥90% of samples (Fig. 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 (Fig. 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%, respectively (all codes) and 26.0%, 20.7%, and 22.0%, respectively (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% (vs. 0.5%) led to full models with slightly worse overall and discrimination performance in each population (maximum decreases in the 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 vs. 202 for MI, 177 vs. 257 for HF, and 147 vs. 209 for CS). When penalization was used, full model coefficients were very similar to those from the main analysis (Fig. A5).

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 yr (%)	34,520 (17.2)	46,115 (27.2)	5,273 (9.3)
Patient characteristics			
Median age (IQR)	70 (58 to 80)	84 (77 to 89)	70 (62 to 78)
Male (vs. female) (%)	132,162 (66.0)	48,622 (28.7)	32,004 (56.6)
Median socioeconomic status (IQR)[a]	4.8 (2.4 to 7.3)	5.4 (2.9 to 7.7)	5.7 (3.3 to 7.9)
ICD-10 codes			
Number of codes included[b]	202	257	209
Median 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)[c]	0.02 (0.01 to 0.03)	0.01 (0.00 to 0.02)	0.01 (0.00 to 0.02)

Abbreviation: 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 absolute values of Pearson correlation coefficients across all pairwise comparisons.

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).

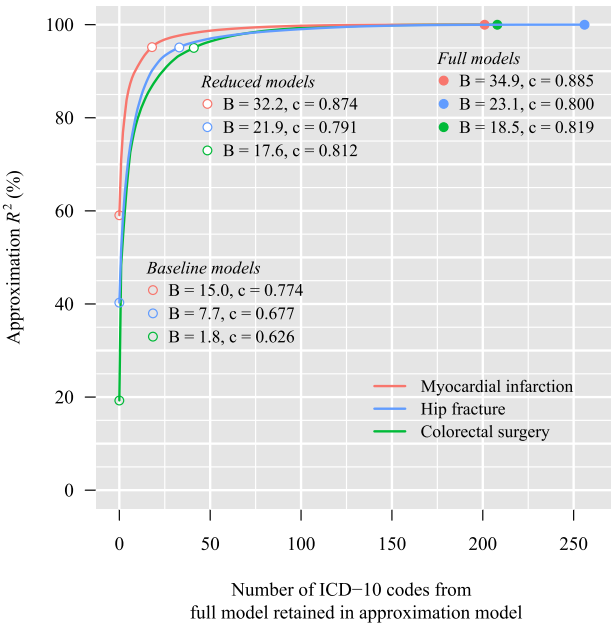

Fig. 2. Percentage of variation in full model predictions explained by models with fewer ICD-10 codes (approximation R^2), and related scaled Brier scores (B%) and *c*-statistics (c). The approximation R^2 equals the percentage of variation explained in the predictions from the full model. In each population, 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.

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, had 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, although 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, sociodemographic characteristics, and other health-related attributes

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[a]	34.6 (34.1 to 35.1)	22.8 (22.4 to 23.2)	17.1 (16.1 to 18.2)
Reduced models[b]	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)

[a] Number of ICD-10 codes in full models (in column order): 202, 257, 209.
[b] Number of codes in reduced models: 18, 33, 41.

[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 Fig. 2).

This method can be used with binary, continuous, and time-to-event outcomes; incorporate nonlinear and interaction terms; and retain penalization 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 prespecify 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 nontransparent 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 Fig. 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

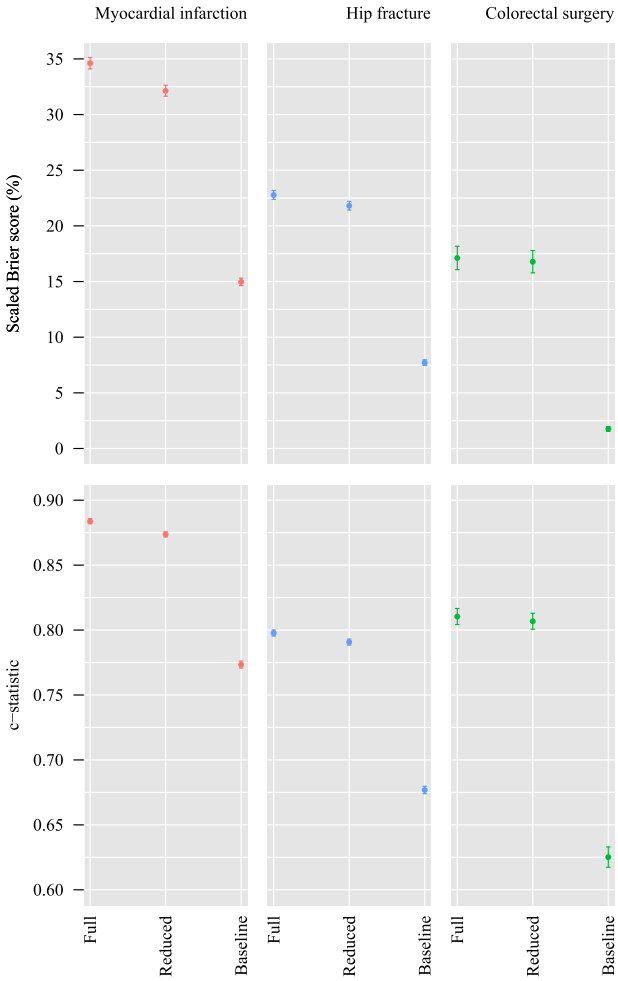

Fig. 3. Optimism-adjusted scaled Brier scores and *c*-statistics of the full, reduced, and baseline models, as estimated from 500 bootstrap samples (with 95% confidence intervals). Number of ICD-10 codes in full models (in column order): 202, 257, 209. Number of codes in reduced models: 18, 33, 41. Baseline models included only age, sex, and socioeconomic status as predictors.

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 because of 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, modeling 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 aging and greater burdens of noncommunicable 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 use these records to improve morbidity measures.

CRediT authorship contribution statement

Thomas E. Cowling: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Funding acquisition. **David A. Cromwell:** Methodology, Resources, Writing - review & editing. **Linda D. Sharples:** Conceptualization, Methodology, Writing - review & editing. **Jan van der Meulen:** Methodology, Writing - review & editing.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.08.001.

References

[1] Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med 2015;12(10):e1001885.
[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(1):198–208.
[3] Jordan KP, Moons KG. Electronic healthcare records and prognosis research. In: Riley RD, van der Windt D, Croft P, Moons KG, editors. Prognosis research in healthcare: concepts, methods, and impact. Oxford: Oxford University Press; 2019:298–310.
[4] World Health Organization. Classification of diseases (ICD). Available at https://www.who.int/classifications/icd/en/. Accessed August 8, 2020.
[5] Sharabiani MT, Aylin P, Bottle A. Systematic review of comorbidity indices for administrative data. Med Care 2012;50(12):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(2):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(4):441−7.
[8] Yealy DM, Yealy DM, Kellum JA, Huang DT, Barnato AE, Weissfeld LA, et al. A randomized trial of protocol-based care for early septic shock. N Engl J Med 2014;370(18):1683−93.
[9] Werner RM, Bradlow ET. Relationship between Medicare's hospital compare performance measures and mortality rates. JAMA 2006; 296(22):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.
[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(1):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(10):1006−14.
[14] Wyatt JC, Altman DG. Commentary: prognostic models: clinically useful or quickly forgotten? BMJ 1995;311(7019):1539−41.
[15] Ambler G, Brady AR, Royston P. Simplifying a prognostic model: a simulation study based on clinical data. Stat Med 2002;21(24): 3803−22.
[16] Hocking RR, Leslie RN. Selection of the best subset in regression analysis. Technometrics 1967;9(4):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(4):1093.
[19] Metcalfe A, Neudam A, Forde S, Liu M, Drosler S, Quan H, et al. Case definitions for acute myocardial infarction in administrative databases and their impact on in-hospital mortality rates. Health Serv Res 2013;48(1):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(3):e92286.
[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(1):44−51.
[22] Royal College of Physicians. National Hip Fracture Database (NHFD) annual report 2016. 2016. Available at https://www.nhfd.co.uk/report2016. Accessed August 8, 2020.
[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.
[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(13):1810−7.
[25] Morris EJ, Taylor EF, Thomas JD, Quirke P, Finan PJ, Coleman MP, Rachet B, Forman D. Thirty-day postoperative mortality after colorectal cancer surgery in England. Gut 2011;60(6):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(1):642.
[27] Office for National Statistics. Deaths. Available at https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths. Accessed August 8, 2020.
[28] NHS Digital. A guide to linked mortality data from hospital episode statistics and the Office for national statistics. Available at https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/linked-hes-ons-mortality-data. Accessed August 8, 2020.
[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(1):3−14.
[30] Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, Collins GS. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 2018;38(7):1276.
[31] Krumholz HM, Coppi AC, Warner F, Triche EW, Li S-X, Mahajan S, Li Y, Bernheim SM, Grady J, Dorsey K, Lin Z, Normand S-LT. Comparative effectiveness of new approaches to improve mortality risk models from medicare claims data. JAMA Netw Open 2019; 2(7):e197314.
[32] Gensheimer MF, Henry AS, Wood DJ, Hastie TJ, Aggarwal S, Dudley SA, et al. Automated survival prediction in metastatic cancer patients using high-dimensional electronic medical record data. J Natl Cancer Inst 2018;111(6):568.
[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 Med Res Methodol 2012;12:82.
[34] Ministry of Housing. Communities & local government, English indices of deprivation. Available at https://www.gov.uk/government/collections/english-indices-of-deprivation. Accessed August 8, 2020.
[35] Harrell FE Jr, Margolis PA, Gove S, Mason KE, Mulholland EK, Lehmann D, 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(8):909−44.
[36] Lawless JF, Singhal K. Efficient screening of Nonnormal regression models. Biometrics 1978;34(2):318−27.
[37] Brier GW. Verification of forecasts expressed in terms of probability. Mon Wea Rev 1950;78(1):1−3.
[38] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21(1):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;38(21):4051.
[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(8):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(5):373−83.
[44] Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998;36(1):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(5):441−7.
[46] NHS Digital. Summary Hospital-level Mortality Indicator (SHMI): ICD-10 to SHMI diagnosis group lookup table. Available at https://digital.nhs.uk/data-and-information/publications/ci-hub/summary-hospital-level-mortality-indicator-shmi. Accessed August 8, 2020.

[47] Healthcare Cost and Utilization Project. Clinical classifications software refined (CCSR) for ICD-10-CM diagnoses. Available at https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp. Accessed August 8, 2020.
[48] Verweij PJ, Van Houwelingen HC. Penalized likelihood in Cox regression. Stat Med 1994;13(23-24):2427−36.
[49] Cessie SL, Houwelingen JCV. Ridge estimators in logistic regression. Appl Stat 1992;41(1):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. https://doi.org/10.1101/19011338.
[51] Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr 1974;19(6):716−23.
[52] Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc 1996;58(1):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(4):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(3):255−63.
[56] Heinze G, Wallisch C, Dunkler D. Variable selection - a review and recommendations for the practicing statistician. Biom J 2018;60(3):431−49.
[57] Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, et al. State of the art in selection of variables and functional forms in multivariable analysis-outstanding issues. Diagn Progn Res 2020;4. https://doi.org/10.1186/s41512-020-00074-3.
[58] Mantel N. Why stepdown procedures in variable selection. Technometrics 1970;12(3):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(30):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(7179):716−9.
[62] Zhou B, Lu Y, Hajifathalian K, Bentham J, Di Cesare M, Danaei G, et al. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. Lancet 2016;387(10027):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 Oncol 2017;3:524−48.