Performance of models for predicting one to three year mortality in older adults

Performance of models for predicting 1-year to 3-year mortality in older adults: a systematic review of externally validated models

Leonard Ho, Carys Pugh, Sohan Seth, Stella Arakelyan, Nazir I Lone, Marcus J Lyall, Atul Anand, Jacques D Fleuriot, Paola Galdi, Bruce Guthrie

Mortality prediction models support identifying older adults with short life expectancy for whom clinical care might need modifications. We systematically reviewed external validations of mortality prediction models in older adults (ie, aged 65 years and older) with up to 3 years of follow-up. In March, 2023, we conducted a literature search resulting in 36 studies reporting 74 validations of 64 unique models. Model applicability was fair but validation risk of bias was mostly high, with 50 (68%) of 74 validations not reporting calibration. Morbidities (most commonly cardiovascular diseases) were used as predictors by 45 (70%) of 64 of models. For 1-year prediction, 31 (67%) of 46 models had acceptable discrimination, but only one had excellent performance. Models with more than 20 predictors were more likely to have acceptable discrimination (risk ratio [RR] *vs* <10 predictors 1·68, 95% CI 1·06–2·66), as were models including sex (RR 1·75, 95% CI 1·12–2·73) or predicting risk during comprehensive geriatric assessment (RR 1·86, 95% CI 1·12–3·07). Development and validation of better-performing mortality prediction models in older people are needed.

Lancet Healthy Longev 2024

Published Online
February 5, 2024
https://doi.org/10.1016/S2666-7568(23)00264-7

Advanced Care Research Centre, Usher Institute, University of Edinburgh, Edinburgh, UK (L Ho PhD, C Pugh PhD, S Seth PhD, S Arakelyan PhD, Prof J D Fleuriot PhD, Prof B Guthrie PhD)**; Royal Infirmary of Edinburgh, National Health Service Lothian, Edinburgh, UK** (N I Lone PhD, M J Lyall PhD)**; Centre for Population Health Sciences, Usher Institute, University of Edinburgh, Edinburgh, UK** (N I Lone, A Anand PhD)**; School of Informatics, University of Edinburgh, Edinburgh, UK** (S Seth, Prof J D Fleuriot, P Galdi PhD)

Correspondence to:
Prof Bruce Guthrie, Advanced Care Research Centre, Usher Institute, University of Edinburgh, Edinburgh EH16 4UX, UK
bruce.guthrie@ed.ac.uk

Introduction

Clinical decision making is complicated by the presence of multimorbidity, frailty, and reduced life expectancy. Clinical guidelines often recommend that treatment decisions should consider reduced life expectancy, particularly when making decisions about preventive treatment where benefits accrue over long periods but harm might happen at any time.[1,2] Such considerations particularly apply to older adults (commonly defined as people aged 65 years and older) in whom multimorbidity, frailty, and high risk of competing mortality (ie, death from conditions other than the one being treated) are common.[3,4] Populations are rapidly ageing across the world, which means that accounting for life expectancy is increasingly salient when making treatment decisions in routine care;[5] however, accurately estimating life expectancy or mortality is difficult for clinicians. This difficulty has driven interest in using prediction models for short-term and medium-term mortality risk in older adults (ie, age ≥65 years) to support clinical decision making to optimise treatment.[6]

The Charlson Comorbidity Index (CCI) was originally devised in the 1980s and was shown to be strongly associated with mortality over 10 years' follow-up.[7] Although not originally devised and validated as a formal prediction model, CCI and subsequent variations (eg, Deyo and Romano indices)[8,9] based on morbidity coding are commonly used to predict mortality alone or in combination with other variables. Many other prediction models have subsequently been developed using a variety of predictors, including age and sex, the presence of various morbidities, functional status, socioeconomic status, and laboratory results; however, all prediction models require external validation before they can be recommended for clinical use. External validation means evaluating the performance of the models in a different dataset, target population, or setting than the one used to develop them.[10,11]

Previous systematic reviews have synthesised and appraised the models developed for predicting mortality of older adults who underwent colorectal cancer surgery,[12] older adults with dementia,[13] and older adults living in nursing homes.[14] The systematic reviews found that many of these prediction models did not have acceptable predictive performance and validation was often methodologically unsatisfactory; however, these previous reviews often included internal validation results (which might be optimistic compared with external validation) and did not evaluate the performance of mortality prediction models in the wider population of older adults. The aim of this systematic review was, therefore, to evaluate external validations of prediction models for short-term to medium-term mortality (<3 years) in older adults.

Methods

We conducted this review based on the TRIPOD-SRMA checklist[10] and PRISMA guidelines (appendix pp 7–8).[15] The review protocol was registered in PROSPERO (CRD42023410747).

See Online for appendix

Eligibility criteria

Studies were eligible if they were prospective or retrospective cohort studies examining the external validation of mortality prediction models, with the full text written in English. We included studies of well established measures such as CCI,[7] Elixhauser Comorbidity Index (ECI),[16] and Rx-Risk Comorbidity Index,[17] either used as the only predictor or where the authors examined their performance with the addition of covariates (eg, age and sex) not included in the core morbidity measure. We excluded conference abstracts, systematic and umbrella reviews, and clinical guidelines.

Studies were eligible if they involved community-dwelling adults with a mean or median age of 65 years or older. We excluded studies in which all or most participants were residents in long-term care facilities, but included studies where less than 10% of participants lived in care or nursing homes. We excluded studies focusing only on specific populations (eg, people who have had a stroke or people with dementia).

Studies were eligible if they externally validated models for predicting all-cause mortality over a period of 3 years or less, chosen because we (and others[1]) considered this period relevant for varying treatment recommendations, such as those for long-term preventive medication. Validation could be for prediction for everyone living in the community, or for prediction done during hospital admission or emergency department attendance. Model predictors could be derived from electronic health record data, survey or trial data, data from questionnaires, other self-report assessment data, data from structured clinical assessment (eg, comprehensive geriatric assessment [CGA]), or a combination of these data.

Search strategy and selection criteria

We searched MEDLINE, Embase, and Cochrane Library from database inception to March 6, 2023 (appendix pp 2–3), with additional hand-searching of reference lists of included studies and excluded conference abstracts. We imported all records into Covidence (Veritas Health Innovation, Melbourne, VIC, Australia) with title and abstract screening done by two reviewers (LH and BG), and full-text screening completed by one reviewer (LH) and then validated by another reviewer (BG).

For more on **Covidence** see https://www.covidence.org

Data extraction and risk of bias and applicability assessment

Based on CHARMS,[18] we extracted the characteristics of included studies and their prediction models. Extracted study characteristics included first author, publication year, study location, funding source, study design, time period over which predictions were made, source of data, measurement of mortality, use of collected data, number of participants, participant selection criteria, and age, sex, and race or ethnicity of participants. Extracted prediction model characteristics were period of prediction, number of predictors, type of predictors, and reported performance measures. Extracted performance measures (appendix p 9) included measures of discrimination (eg, area under the receiver operating characteristic curve [AUC] and Harrell's C statistic), calibration (eg, calibration plot, calibration-in-the-large, and Hosmer-Lemeshow test), and measures of overall performance, reclassification, and clinical usefulness (eg, Brier score, pseudo R^2, net reclassification index, and decision curve).[19] Discrimination is a measure of how well the model can distinguish between people who die and people who survive. Calibration reflects how accurately the model predicts the outcome, and is a crucial performance feature for clinical use—a model can have good discrimination in terms of predicted risk being higher in those who die versus survivors but still produce inaccurate (poorly calibrated) mortality risk estimates.

We conducted risk of bias and applicability assessment for the validations of prediction models using PROBAST.[20] The procedures described here were performed by one reviewer (LH) and then independently validated by another (BG). Disagreements were resolved by discussion between the two reviewers.

Data synthesis

Few models were externally validated more than once, with high between-study heterogeneity, and meta-analysis to estimate pooled discrimination was not appropriate. Instead, we narratively synthesised findings using descriptive statistics and tables. We adopted commonly used cutoff points for discrimination to aid interpretation, by considering a prediction model with AUC or C statistic of 0·50–0·69 as having poor discrimination, 0·70–0·79 as having acceptable discrimination, 0·80–0·89 as having excellent discrimination, and 0·90 or higher as having outstanding discrimination (for these measures, a value of 0·50 means the model performs no better than chance, and 1·00 means that discrimination is perfect).[21,22] The results were reported according to the period over which predictions were made, which varied between 1 week and 3 years, with some studies reporting model performance over multiple time periods (in these instances we included performance for all reported periods of 3 years or less). Where authors calculated discrimination using two or more sources of data (eg, inpatient data only *vs* inpatient and outpatient data), we used better results to summarise model performance. Calibration is harder to formally assess because it involves more judgement.[23] Where authors applied the Hosmer-Lemeshow test, we considered a prediction model with a p value of at least 0·05 as having adequate calibration.[24] Otherwise, we extracted the authors' summary interpretation of model calibration.

Associations between model characteristics and model discrimination

For studies reporting discrimination using either AUC or C statistic, we examined model characteristics associated with discrimination being acceptable or better (AUC or C statistic ≥0·70) when predicting mortality at 1 month and 1 year (numbers were too small to analyse for prediction over different time periods). The model characteristics examined were the use of morbidities, age, sex, clinical assessment data, and survey or trial data as a predictor, the number of predictors used, and the timepoint for prediction. The associations were examined by calculating univariate relative risks (RRs) with 95% CIs (small sample sizes meant multivariate analyses were infeasible).

Results

Study selection

The literature search yielded 43 807 records. After deduplication, we performed title and abstract screening on 29 215 records, of which 153 full-text records were screened (figure). 34 papers reporting 36 studies were eligible and reported a total of 74 validations of 64 unique prediction models (appendix pp 4–5).

Study characteristics

The 36 included studies were published between 2001 and 2023, with the largest number (n=10; 28%) conducted in the USA (table 1). 19 (53%) studies were fully funded by governments, universities, or other public bodies (or a combination thereof), and seven (19%) were fully or partly supported by private research foundations or medical centres. Half (n=18) of included studies were prospective cohort studies. For 30 (83%) included studies, the measurement of mortality was based on data retrieved from government registries, insurance databases, or electronic health records, with or without additional information provided by family members. 19 (53%) studies focused on model development and validation, compared with 16 (44%) that examined model validation and one that examined model recalibration (3%). Prediction models were validated in a total of 8 492 960 participants, but most validation studies involved fewer than 2100 participants (median 2045, IQR 9542), with mean or median age ranging from 68·9 years to 92·1 years (appendix pp 10–22). Only eight (22%) studies reported participants' race or ethnicity. Seven (19%) studies involved some participants younger than 65 years,[25–30] two (6%) involved only military veterans,[31,32] and one (3%) involved only men.[33]

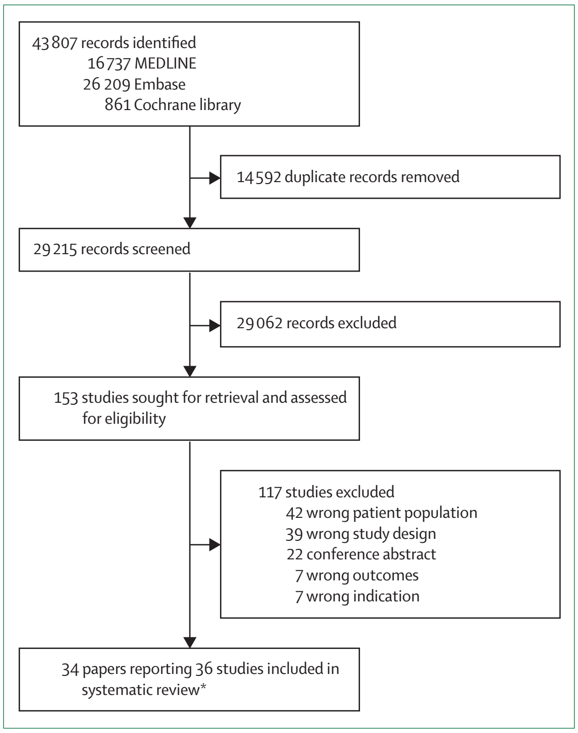
43 807 records identified
16 737 MEDLINE
26 209 Embase
861 Cochrane library
14 592 duplicate records removed
29 215 records screened
29 062 records excluded
153 studies sought for retrieval and assessed for eligibility
117 studies excluded
42 wrong patient population
39 wrong study design
22 conference abstract
7 wrong outcomes
7 wrong indication
34 papers reporting 36 studies included in systematic review*

Figure: **Study selection**

*Studies included 74 external validations and 64 unique prediction models.

Prediction model characteristics

The time period over which prediction was examined ranged from 1 week to 3 years. 44 (59%) of the 74 validations used data recorded in electronic health records, 29 (39%) used questionnaire data, 28 (38%) used clinical assessment data, 11 (15%) used previous survey or trial research data, and three (4%) used participant self-report assessment data (table 2). 28 (38%) mortality predictions were made at a non-specific timepoint (ie, any time), 23 (31%) during hospital admission, and 19 (26%) during emergency department attendance.

Deyo CCI was evaluated in five different populations (appendix pp 10–14),[30,32–34] but the actual models evaluated varied (eg, two Deyo CCI models examined predictive performance of Deyo CCI alone, whereas the other three evaluated models included Deyo CCI and a range of other covariates). Identification of Seniors at Risk,[35–38] Multidimensional Prognostic Index,[39–42] and Quan CCI plus covariates[31,33,43,44] were also validated in four populations.

The median number of predictors included in the models was 12 (range 1–109, IQR 16). The most frequently included types of predictors (table 3, appendix pp 23–41) were morbidities (45 models, 70%), age

	Number of studies (%; N=36)
Study location	
USA	10 (28%)
Italy	6 (17%)
Australia	3 (8%)
China	3 (8%)
Other	14 (39%)
Funding source	
Governments, universities, or other public bodies (or a combination of these)	19 (53%)
Private research foundations or medical centres	7 (19%)
Not reported	10 (28%)
Study design	
Prospective cohort	18 (50%)
Retrospective cohort	18 (50%)
Measurement of mortality	
Based on data retrieved from government registries, insurance databases, or electronic health records	30 (83%)
Based on data obtained from follow-up with participants or family members	2 (6%)
Not reported	4 (11%)
Use of collected data	
Model development and validation	19 (53%)
Model validation only	16 (44%)
Model recalibration	1 (3%)
Number of participants	
<1000	11 (31%)
1000–9999	16 (44%)
≥10 000	9 (25%)

Table 1: **Characteristics of the 36 included studies**

	Number of model validations (%; N=74)
Time period over which prediction is made*	
≤3 months	33 (45%)
6 months	13 (18%)
1 year	46 (62%)
>1 year	13 (18%)
Source of data†	
Electronic health records	44 (59%)
Questionnaires administered by staff	29 (39%)
Clinical assessments	28 (38%)
Previous survey or trial data	11 (15%)
Self-report assessments	3 (4%)
Time of prediction	
Any time	28 (38%)
During emergency department attendance	19 (26%)
During inpatient admission	23 (31%)
During CGA	4 (5%)
Evaluation of discrimination*	
AUC	46 (62%)
Harrell's C statistic	21 (28%)
Other‡	10 (14%)
Not reported	7 (9%)
Rating of discrimination (where reported as AUC or Harrell's C statistic)*	
Excellent (≥0·80)	12 (16%)
Acceptable (0·70–0·79)	60 (81%)
Poor (0·50–0·69)	33 (45%)
Not reported	7 (9%)
Evaluation of calibration*	
Calibration plot analysis with explicit interpretation	12 (16%)
Calibration plot analysis without explicit interpretation	11 (15%)
Hosmer-Lemeshow test results	4 (5%)
Calibration-in-the-large	1 (1%)
Other§	4 (5%)
Not reported	50 (68%)
Rating of Calibration (based on authors' interpretations or Hosmer-Lemeshow test)*¶	
Adequate	20 (25%)
Poor	1 (1%)
Not reported	60 (74%)

AUC=area under the receiver operating characteristic curve. CGA=comprehensive geriatric assessment. *As validation could be done for multiple time periods or report multiple measures of discrimination or calibration, the sum of the counts exceeds the number of model validations (n=74). †As validation could be done in more than one source of data, the sum of the counts might exceed the number of included models (n=74). ‡Other measures of discrimination included sensitivity, specificity, positive and negative predictive values, and positive and negative likelihood ratios. §Other measures of calibration included calibration intercept, calibration slope, calibration error, observed-to-expected ratio, and correct classification. ¶A p value of at least 0·05 in the Hosmer-Lemeshow test was taken to indicate adequate calibration.

Table 2: **Characteristics of the 74 model validations done in included studies**

(33 models, 52%) and sex or gender (30 models, 47%). The included morbidities varied considerably. Cardiovascular diseases, such as hypertension, myocardial infarction, and arrhythmia contributed to 39 (61%) prediction models, with less frequent contribution for cancer (n=37, 58%), neurological or psychiatric diseases (n=35, 55%), respiratory diseases (n=34, 53%), and renal diseases (n=33, 52%). Other uncategorised conditions, such as falls, osteoporosis, musculoskeletal issues, and visual impairment, were adopted by some prediction models. A wide range of other variables was used by small proportions of models (including socioeconomic status or index, care requirements, nutritional status, professional judgements and recommendations, social aspects and support, quality of life, education attainment, electrocardiogram results, general health, insurance enrolment, and palliative care referral).

	Number of models using each type of predictor (%; N=64)
Any morbidity	45 (70%)
Cardiovascular diseases	39 (61%)
Cancer	37 (58%)
Neurological or psychiatric diseases	35 (55%)
Respiratory diseases	34 (53%)
Renal diseases	33 (52%)
Metabolic diseases	30 (47%)
Gastrointestinal diseases	28 (44%)
Liver diseases	27 (42%)
Urogenital or sexually transmitted diseases	26 (41%)
Rheumatological diseases	25 (39%)
Haematological diseases	14 (22%)
Age	33 (52%)
Sex or gender	30 (47%)
Blood test and urinalysis	17 (27%)
Physical status	14 (22%)
Previous admissions or length of stay	13 (20%)
Mental or cognitive status	13 (20%)
Activities of daily living and instrumental activities of daily living	12 (19%)
Medications	11 (17%)
Race and ethnicity	9 (14%)
Alcohol intake or alcoholism	9 (14%)
Weight	8 (13%)
Residency status	8 (13%)
Vital signs	8 (13%)
Number of comorbidities	7 (11%)
Marital status	6 (9%)
Smoking	6 (9%)

28 models used 1–10 predictors, 17 models used 11–20 predictors, 18 models used more than 20 predictors, and one model did not report the number of predictors used. Only predictors used in more than five studies shown; predictors used in all models are shown in appendix (pp 23–41).

Table 3: **Predictors used by the 64 unique prediction models examined**

Prediction model performance

Model discrimination

Overall, 67 (91%) of the 74 validations reported model discrimination (table 2; appendix pp 29–41), either as AUC (n=46) or C statistic (n=21). 46 (62%) reported discrimination at 1 year, compared with 27 (36%) at 1 month and 13 (18%) at 6 months. For 1-year prediction, 31 (67%) of 46 validations reporting discrimination had acceptable model discrimination (AUC or C statistic 0·70–0·79), and discrimination was poor for 14 validations (AUC or C statistic <0·70). Only one validation (Combined Comorbidity Score plus additional covariates) reported excellent discrimination (AUC 0·81, 95% CI 0·81–0·81).[32]

For prediction at 1 month, 11 had acceptable model discrimination and 12 had poor discrimination. Four validations reported excellent discrimination: Combined Comorbidity Score plus covariates (C statistic 0·86, 95% CI 0·85–0·87),[45] Romano CCI plus covariates (C statistic 0·86, 0·85–0·87),[45] van Walraven ECI plus covariates (C statistic 0·84, 0·83–0·85),[45] and the RISE UP Score (AUC 0·83, 0·77–0·90).[38]

For prediction at 3 months, the Combined Comorbidity Score plus covariates (C statistic 0·82, 95% CI 0·82–0·83),[45] Risk Stratification Index 3.0 (AUC 0·82, 0·82–0·82),[26] Romano CCI plus covariates (C statistic 0·81, 0·81–0·81),[45] and van Walraven ECI plus covariates (C statistic 0·81, 0·80–0·81)[45] had excellent discrimination. Combined Comorbidity Score plus covariates had excellent prediction discrimination at 6 months (C statistic 0·81, 0·80–0·81)[45] and 3 years (AUC 0·81, 0·80–0·81).[32] Smolin Model had excellent prediction discrimination at 6 months (AUC 0·85, 0·83–0·86).[46]

Associations between model characteristics and model discrimination

There were no statistically significant associations between model characteristics and discrimination at 1-month prediction; however, for prediction at 1 year, models with more than 20 predictors (RR 1·68, 95% CI 1·06–2·66) were statistically significantly more likely to have acceptable discrimination versus 1–10 predictors, as were models including sex as a predictor (RR 1·75, 1·12–2·73), and models used during CGA (RR 1·86, 1·12–3·07) (table 4). Although marginally statistically non-significant, models that included age as a predictor were possibly more likely to have acceptable discrimination at 1 year (RR 1·91, 95% CI 0·96–3·80, p=0·067).

Model calibration

Only 24 (32%) of the 74 validations reported calibration in any way (table 2). 15 only reported calibration plots, eight reported both calibration plots and at least one calibration statistic (of which four used the Hosmer-Lemeshow test), and one only reported calibration-in-the-large. Only 12 studies reporting calibration plots or Hosmer-Lemeshow tests explicitly interpreted the meaning of those plots. Based on authors' interpretations, four models had adequate calibration for prediction at 1 month, two at 6 months, seven at 1 year, three at 2 years, and four at 3 years.

Other model performance measures

15 studies examined other model performance measures (appendix pp 29–41). Seven calculated pseudo R^2 (but only one explicitly interpreted the results, concluding that Electronic Frailty Index had poor overall performance at 1 year and 3 years)[47] and six calculated Brier scores (but none explicitly interpreted the scores). Five used net reclassification index to compare the performance of different models, but only one explicitly interpreted the results, concluding that the Combined Comorbidity Score was superior to Romano CCI and van Walraven ECI at 1-month, 3-month, 6-month, and 1-year predictions.[45] One used decision curve analysis and concluded that, compared with the original version, the Patient-Reported Outcome Mortality Prediction Tool (Recalibrated) showed a positive net benefit when population mortality was between 0% and 25·0%.[48]

Risk of bias and applicability of the validations

Overall, only 11 (15%) of 74 validations were at low risk of bias, nine (12%) at unclear risk of bias, and 54 (73%) at high risk of bias (appendix pp 6, 42–43). All had satisfactory performance in the participants, predictors, and outcome domains of PROBAST; however, in the analysis domain, those with high risk of bias did not report both model discrimination and model calibration. The nine validations with unclear risk of bias did not have ≥100 or more deaths (events) by the end of follow-up or only narratively described a calibration plot without explicit interpretation or use of formal calibration measures. Applicability to the target population of older people was generally good, but 18 (24%) of 74 validations had unclear concerns over applicability because they included some participants younger than 65 years,[25–30] only military veterans,[31,32] or only male participants.[33]

Discussion

This systematic review examined 74 validations of 64 unique prediction models for all-cause mortality of older adults over a variety of time periods up to 3 years. The methodological quality of validations was generally poor, with one in ten not reporting discrimination and two-thirds not reporting calibration. The examined prediction models used a wide variety of predictors, with 70% using morbidities, and approximately half using age and sex. The most common group of morbidities used as predictors was cardiovascular diseases, followed by cancer and neurological or psychiatric diseases, but there was considerable heterogeneity between models. For 1-year mortality prediction, discrimination was poor for a

	Number of models with acceptable discrimination* (%)	RR (95% CI)	p value
Predicting 1-month mortality (N=27)			
Model uses morbidities as a predictor			
No (n=11)	6 (55%)	Ref	··
Yes (n=16)	9 (56%)	1·03 (0·52–2·06)	0·931
Model uses age as a predictor			
No (n=10)	7 (70%)	Ref	··
Yes (n=17)	8 (47%)	0·67 (0·35–1·28)	0·229
Model uses sex as a predictor			
No (n=15)	7 (47%)	Ref	··
Yes (n=12)	8 (67%)	1·43 (0·73–2·80)	0·299
Model adopts clinical assessment data			
No (n=12)	8 (67%)	Ref	··
Yes (n=15)	7 (47%)	0·70 (0·36–1·37)	0·299
Model adopts questionnaire or self-report data			
No (n=12)	7 (58%)	Ref	··
Yes (n=15)	8 (53%)	0·91 (0·47–1·79)	0·794
Number of predictors			
1–10 (n=12)	5 (42%)	Ref	··
11–20 (n=6)	3 (50%)	1·20 (0·42–3·41)	0·738
>20 (n=9)	7 (78%)	1·87 (0·88–3·97)	0·105
Timepoint for prediction			
During inpatient admission (n=13)	6 (46%)	Ref	··
During emergency department attendance (n=7)	4 (57%)	1·24 (0·52–2·95)	0·630
Any time (n=7)	5 (71%)	1·55 (0·73–3·28)	0·255
Predicting 1-year mortality (N=46)			
Model uses morbidities as a predictor			
No (n=13)	8 (62%)	Ref	··
Yes (n=33)	24 (73%)	1·18 (0·73–1·91)	0·493
Model uses age as a predictor			
No (n=12)	5 (42%)	Ref	··
Yes (n=34)	27 (79%)	1·91 (0·96–3·80)	0·067
Model uses sex as a predictor			
No (n=22)	11 (50%)	Ref	··
Yes (n=24)	21 (88%)	1·75 (1·12–2·73)	0·014
Model adopts clinical assessment data			
No (n=35)	27 (77%)	Ref	··
Yes (n=11)	5 (46%)	0·59 (0·30–1·15)	0·123
Model adopts questionnaire or self-report data			
No (n=39)	28 (72%)	Ref	··
Yes (n=7)	4 (57%)	0·80 (0·41–1·56)	0·505
Number of predictors			
1–10 (n=19)	10 (53%)	Ref	··
11–20 (n=10)	7 (70%)	1·33 (0·74–2·40)	0·342
>20 (n=17)	15 (88%)	1·68 (1·06–2·66)	0·028
Timepoint for prediction			
During inpatient admission (n=13)	7 (54%)	Ref	··
During emergency department attendance (n=6)	2 (33%)	0·62 (0·18–2·14)	0·448
Any time (n=25)	21 (84%)	1·56 (0·92–2·65)	0·101
During CGA (n=2)	2 (100%)	1·86 (1·12–3·07)	0·016

CGA=comprehensive geriatric assessment. RR=relative risk. *Acceptable discrimination indicates an area under the receiver operating characteristic curve or Harrell's C statistic ≥0·70.

Table 4: **Model characteristics associated with prediction models having acceptable discrimination**

third of the models (AUC or C statistic <0·70) and only excellent (AUC 0·81) in one study. For the minority explicitly reporting their judgement of calibration, it was generally reported to be adequate or good. Models with more than 20 predictors that included sex as a predictor and were used during CGA were more likely to have acceptable discrimination, but there was no evidence that studies using data from clinical assessment were superior to studies that only used routine data.

Strengths of this systematic review include the performance of a comprehensive literature search in major databases and reporting consistent with both CHARMS and PROBAST; however, the systematic review also has several limitations. First, the heterogeneity of the studies precluded meta-analysis, in part because studies that framed themselves as evaluating the performance of commonly used indices such as Deyo CCI or Romano CCI actually evaluated a variety of different models (from models only using the morbidity index in prediction to those with varying numbers of other predictors). For instance, Gagne and colleagues[45] evaluated the performance of Romano CCI, but also included age, race, and gender as predictors, which we considered to be an evaluation of Romano CCI plus covariates rather than Romano CCI alone. The main implication is that there is no tool with evidence of good performance from multiple external validations. Second, our understanding of calibration is limited because only a third of studies reported any calibration data, and many of those with at least some calibration data (most commonly a calibration plot) did not provide an interpretation of their meaning. Third, although some prediction models were shown to have adequate performance in studies with low risk of bias, many did not report their study population in enough detail to be sure that the findings apply to diverse populations, particularly regarding race or ethnicity (or both). Fourth, although all studies examined populations with mean or median age 65 years or older, some studies included some younger people because they allowed adults younger than 65 years.[25,27,29] In principle, including people in these age groups might mean predictive performance would be evaluated as more favourable than if the model was examined only in older people (as age is such a strong predictor of mortality), but these models did not show better than acceptable performance. Fifth, studies are all from middle-income or high-income countries, and the results might not apply to low-income or middle-income countries where life expectancy is shorter. Finally, restricting the analyses to external validation studies means that some potentially superior prediction tools were not considered (eg, QMortality, where discrimination was excellent in internal validation);[49] however, prediction tool performance is typically worse in external validation (although it can be improved by recalibrating to local data),[50] and external validation is therefore recommended before clinical use.

Previous reviews examining the predictive value of models using administrative or routine electronic data (such as CCI and its derivatives) in the whole population have concluded that models using various morbidity indices somewhat outperform models only using age and sex,[50] and that models that included ECI as a predictor most consistently have the highest discrimination for predicting mortality.[51] The C statistics observed in the general adult population or in subpopulations defined by disease are variable, but often higher than those observed in this systematic review (which is probably at least partly explained by examining performance in populations with a wider age range, since age is a dominant predictor).[51] Our findings in studies of older people, however, are consistent with other reviews of prediction performance in older adults with dementia[13] and residents of nursing homes,[14] where discrimination in external validation was commonly poor and never better than acceptable. Our findings are also consistent with review findings that models predicting emergency hospital admission in older people perform poorly.[52] Our finding that calibration was uncommonly examined and rarely explicitly interpreted is consistent with other reviews of prediction models for older adults.[14,52–54]

Although clinical guidelines recommend predicting life expectancy to identify people for whom care might need personalisation,[1] existing mortality prediction tools for older adults usually only have acceptable discrimination (the ability to distinguish between those who die and those who do not) and most have uncertain calibration (the extent to which prediction is accurate for groups of patients). At a minimum, users of these tools should, therefore, be very cautious in interpreting the meaning of a prediction, both for population risk stratification and in particular for predicting mortality risk of an individual.

In terms of research, better mortality prediction tools for older adults are needed; however, how best to develop such tools, in terms of optimal data sources and predictors, is unclear. This systematic review found that prediction was better if sex and possibly age were included in the model, and that prediction was better for models with more predictors, but there was no evidence that including data from clinical assessment (such as CGA) improved model performance. Future research would ideally extend existing approaches to modelling, including by comparing prediction tools using different combinations of routine data and bespoke data (eg, clinical assessment or self-report of function) to clarify if models using bespoke data have a meaningful advantage that outweighs the fact that they are harder to deploy at scale (which would have many advantages in terms of transferability of models). In addition, all models currently use data measured at a single timepoint, typically without any interaction terms, and the potential values of models that better account for interactions and time-varying predictors need evaluating. For existing and new models, high-quality external validation that robustly examines calibration and discrimination[10] and that examines performance in important subgroups (eg, by age group, gender, race or ethnicity, or presence of dementia) is needed, as good performance overall can conceal poor performance in crucial subgroups.[3] Ideally, external validations should include head-to-head comparisons of different prediction models in the same population to further inform model choice.[55]

Conclusion

This systematic review synthesised 74 external validations of 64 unique prediction models for predicting short to medium term, all-cause mortality of older adults. Methodological quality was variable, reporting was often poor, performance in terms of discrimination was rarely better than adequate, and calibration was usually uncertain. Development and robust validation of better mortality prediction tools for older adults is needed to support the personalisation of care in the face of short life expectancy.

Contributors

BG, LH, CP, SS, NIL, MJL, AA, JDF, and PG conceptualised the systematic review. BG and LH secured the funding. LH, SA, and BG designed the literature search strategies. LH and BG performed title screening, study selection, and data extraction. LH conducted the risk of bias and applicability assessment. LH, SS, AA, NIL, and BG completed the data synthesis and analysis. LH and BG developed the first draft of the manuscript. All authors interpreted the results and read and approved the final manuscript.

Declaration of interests

We declare no competing interests.

Acknowledgments

The systematic review was funded by the National Institute for Health Research (NIHR) Artificial Intelligence and Multimorbidity: Clustering in Individuals, Space and Clinical Context (NIHR202639) and the Advanced Care Research Centre at the University of Edinburgh (established by a research grant from Legal and General Group). The funders had no role in conduct of the study, interpretation, or the decision to submit for publication. The views expressed are those of the authors and not necessarily those of the NIHR, the Department of Health and Social Care, or Legal and General.

References

1 National Institute for Health and Care Excellence. Multimorbidity: clinical assessment and management. London: National Institute for Health and Care Excellence, 2016.

2 Lee SJ, Leipzig RM, Walter LC. Incorporating lag time to benefit into prevention decisions for older adults. *JAMA* 2013; **310:** 2609–10.

3 Livingstone SJ, Morales DR, McMinn M, Eke C, Donnan P, Guthrie B. Effect of competing mortality risks on predictive performance of the QFracture risk prediction tool for major osteoporotic fracture and hip fracture: external validation cohort study in a UK primary care population. *BMJ Med* 2022; **1:** e000316.

4 Livingstone S, Morales DR, Donnan PT, et al. Effect of competing mortality risks on predictive performance of the QRISK3 cardiovascular risk prediction tool in older people and those with comorbidity: external validation population cohort study. *Lancet Healthy Longev* 2021; **2:** e352–61.

5 WHO. Ageing and health. Geneva: World Health Organization. Oct 1, 2022. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health (accessed March 13, 2023).

6 Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* 2006; **144:** 201–09.

7 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.

8 Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol* 1992; **45:** 613–19.

9 Romano PS, Roos LL, Jollis JG. Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives. *J Clin Epidemiol* 1993; **46:** 1075–79, discussion 1081–90.

10 Snell KIE, Levis B, Damen JAA, et al. Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA). *BMJ* 2023; **381:** e073538.

11 Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015; **68:** 279–89.

12 Souwer ETD, Bastiaannet E, Steyerberg EW, Dekker JT, van den Bos F, Portielje JEA. Risk prediction models for postoperative outcomes of colorectal cancer surgery in the older population - a systematic review. *J Geriatr Oncol* 2020; **11:** 1217–28.

13 Smith EE, Ismail Z. Mortality risk models for persons with dementia: a systematic review. *J Alzheimers Dis* 2021; **80:** 103–11.

14 Zhang S, Zhang K, Chen Y, Wu C. Prediction models of all-cause mortality among older adults in nursing home setting: a systematic review and meta-analysis. *Health Sci Rep* 2023; **6:** e1309.

15 Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021; **372:** n71.

16 Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Med Care* 1998; **36:** 8–27.

17 Fishman PA, Goodman MJ, Hornbrook MC, Meenan RT, Bachman DJ, O'Keeffe Rosetti MC. Risk adjustment using automated ambulatory pharmacy data: the RxRisk model. *Med Care* 2003; **41:** 84–99.

18 Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014; **11:** e1001744.

19 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.

20 Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019; **170:** 51–58.

21 Hosmer DW, Lemeshow S. Applied logistic regression. New York, NY: John Wiley & Sons, 2000.

22 Hartman N, Kim S, He K, Kalbfleisch JD. Pitfalls of the concordance index for survival outcomes. *Stat Med* 2023; **42:** 2179–90.

23 Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019; **17:** 230.

24 Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 1997; **16:** 965–80.

25 Armiñanzas C, Velasco L, Calvo N, Portilla R, Riancho JA, Valero C. CURB-65 as an initial prognostic score in internal medicine patients. *Eur J Intern Med* 2013; **24:** 416–19.

26 Greenwald S, Chamoun GF, Chamoun NG, et al. Risk Stratification Index 3.0, a broad set of models for predicting adverse events during and after hospital admission. *Anesthesiology* 2022; **137:** 673–86.

27 Ha DT, Dang TQ, Tran NV, Pham TN, Nguyen ND, Nguyen TV. Development and validation of a prognostic model for predicting 30-day mortality risk in medical patients in emergency department (ED). *Sci Rep* 2017; **7:** 46474.

28 Jung HW, Kim JW, Han JW, et al. Multidimensional geriatric prognostic index, based on a geriatric assessment, for long-term survival in older adults in Korea. *PLoS One* 2016; **11:** e0147032.

29 Moman RN, Loprinzi Brauer CE, Kelsey KM, Havyer RD, Lohse CM, Bellolio MF. PREDICTing mortality in the emergency department: external validation and derivation of a clinical prediction tool. *Acad Emerg Med* 2017; **24:** 822–31.

30 Perkins AJ, Kroenke K, Unützer J, et al. Common comorbidity scales were similar in their ability to predict health care costs and mortality. *J Clin Epidemiol* 2004; **57:** 1040–48.

31 Lu CY, Barratt J, Vitry A, Roughead E. Charlson and Rx-Risk comorbidity indices were predictive of mortality in the Australian health care setting. *J Clin Epidemiol* 2011; **64:** 223–28.

32 Radomski TR, Zhao X, Hanlon JT, et al. Use of a medication-based risk adjustment index to predict mortality among veterans dually-enrolled in VA and medicare. *Healthc* 2019; **7:** S2213-0764(18)30230-6.

33 Mnatzaganian G, Ryan P, Norman PE, Hiller JE. Accuracy of hospital morbidity data and the performance of comorbidity scores as predictors of mortality. *J Clin Epidemiol* 2012; **65:** 107–15.

34 Desai MM, Bogardus ST Jr, Williams CS, Vitagliano G, Inouye SK. Development and validation of a risk-adjustment index for older patients: the high-risk diagnoses for the elderly scale. *J Am Geriatr Soc* 2002; **50:** 474–81.

35 Bahadirli S, Kurt E, Rohat AK, Kurt SZE, Sanri E, Bulut M. Evaluation and comparison of screening tools used to predict the adverse outcomes of elderly patients in the emergency department. *Acta Med Mediter* 2021; **37:** 1133–39.

36 O'Caoimh R. Validation of the Risk Instrument for Screening in the Community (*RISC*) among older adults in the emergency department. *Int J Environ Res Public Health* 2023; **20:** 3734.

37 Salvi F, Morichi V, Lorenzetti B, et al. Risk stratification of older patients in the emergency department: comparison between the Identification of Seniors at Risk and Triage Risk Screening Tool. *Rejuvenation Res* 2012; **15:** 288–94.

38 Zelis N, Buijs J, de Leeuw PW, van Kuijk SMJ, Stassen PM. A new simplified model for predicting 30-day mortality in older medical emergency department patients: the rise up score. *Eur J Intern Med* 2020; **77:** 36–43.

39 Bryant K, Sorich MJ, Woodman RJ, Mangoni AA. Validation and adaptation of the multidimensional prognostic index in an older Australian cohort. *J Clin Med* 2019; **8:** 1820.

40 Pilotto A, Ferrucci L, Franceschi M, et al. Development and validation of a multidimensional prognostic index for one-year mortality from comprehensive geriatric assessment in hospitalized older patients. *Rejuvenation Res* 2008; **11:** 151–61.

41 Pilotto A, Sancarlo D, Aucella F, et al. Addition of the multidimensional prognostic index to the estimated glomerular filtration rate improves prediction of long-term all-cause mortality in older patients with chronic kidney disease. *Rejuvenation Res* 2012; **15:** 82–88.

42 Sancarlo D, D'Onofrio G, Franceschi M, et al. Validation of a Modified-Multidimensional Prognostic Index (m-MPI) including the Mini Nutritional Assessment Short-Form (MNA-SF) for the prediction of one-year mortality in hospitalized elderly patients. *J Nutr Health Aging* 2011; **15:** 169–73.

43 Mehta HB, Li S, An H, Goodwin JS, Alexander GC, Segal JB. Development and validation of the summary Elixhauser comorbidity score for use with ICD-10-CM-coded data among older adults. *Ann Intern Med* 2022; **175:** 1423–30.

44 Mayo NE, Nadeau L, Levesque L, Miller S, Poissant L, Tamblyn R. Does the addition of functional status indicators to case-mix adjustment indices improve prediction of hospitalization, institutionalization, and death in the elderly? *Med Care* 2005; **43:** 1194–202.

45 Gagne JJ, Glynn RJ, Avorn J, Levin R, Schneeweiss S. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *J Clin Epidemiol* 2011; **64:** 749–59.

46 Smolin B, Levy Y, Sabbach-Cohen E, Levi L, Mashiach T. Predicting mortality of elderly patients acutely admitted to the department of internal medicine. *Int J Clin Pract* 2015; **69:** 501–08.

47 Clegg A, Bates C, Young J, et al. Development and validation of an electronic frailty index using routine primary care electronic health record data. *Age Ageing* 2016; **45:** 353–60.

48 Duarte CW, Black AW, Murray K, et al. Validation of the Patient-Reported Outcome Mortality Prediction Tool (PROMPT). *J Pain Symptom Manage* 2015; **50:** 241–7.e6.

49 Hippisley-Cox J, Coupland C. Development and validation of QMortality risk prediction algorithm to estimate short term risk of death and assess frailty: cohort study. *BMJ* 2017; **358:** j4208.

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

51 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.

52 Klunder JH, Panneman SL, Wallace E, et al. Prediction models for the prediction of unplanned hospital admissions in community-dwelling older adults: a systematic review. *PLoS One* 2022; **17:** e0275116.

53 Gao Y, Chen Y, Hu M, et al. Characteristics and quality of diagnostic and risk prediction models for frailty in older adults: a systematic review. *J Appl Gerontol* 2022; **41:** 2113–26.

54 Van Grootven B, van Achterberg T. Prediction models for functional status in community dwelling older adults: a systematic review. *BMC Geriatr* 2022; **22:** 465.

55 Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol* 2008; **61:** 1085–94.