Intensive Care Med (2011) 37:1258–1268
DOI 10.1007/s00134-011-2265-6

REVIEW

Lilian Minne
Jeroen Ludikhuize
Evert de Jonge
Sophia de Rooij
Ameen Abu-Hanna

Prognostic models for predicting mortality in elderly ICU patients: a systematic review

Received: 12 November 2010
Accepted: 19 April 2011
Published online: 7 June 2011

Electronic supplementary material
The online version of this article (doi:10.1007/s00134-011-2265-6) contains supplementary material, which is available to authorized users.

L. Minne (✉) · A. Abu-Hanna
Department of Medical Informatics,
Academic Medical Center, Room J1b-124,
P.O. Box 22660, 1100 DD Amsterdam,
The Netherlands
e-mail: L.Minne@amc.uva.nl
Tel.: +31-20-5666893
Fax: +31-20-6919840

J. Ludikhuize
Department of Quality, Process
and Innovation, Academic Medical Center,
P.O. Box 22660, 1100 DD Amsterdam,
The Netherlands

S. de Rooij
Department of Geriatrics,
Academic Medical Center, P.O. Box 22660,
1100 DD Amsterdam, The Netherlands

E. de Jonge
Department of Intensive Care,
Leiden University Medical Center,
P.O. Box 9600, 2300 RC Leiden,
The Netherlands

Abstract *Purpose:* To systematically review prognostic research literature on development and/or validation of mortality predictive models in elderly patients. *Methods:* We searched the Scopus database until June 2010 for articles aimed at validating prognostic models for survival or mortality in elderly intensive care unit (ICU) patients. We assessed the models' fitness for their intended purpose on the basis of barriers for use reported in the literature, using the following categories: (1) clinical credibility, (2) methodological quality (based on an existing quality assessment framework), (3) external validity, (4) model performance, and (5) clinical effectiveness. *Results:* Seven studies were identified which met our inclusion criteria, one of which was an external validation study. In total, 17 models were found of which six were developed for the general adult ICU population and eleven specifically for elderly patients. Cohorts ranged from 148 to 12,993 patients and only smaller ones were obtained prospectively. The area under the receiver operating characteristic curve (AUC) was most commonly used to measure performance (range 0.71–0.88). The median number of criteria met for clinical credibility was 4.5 out of 7 (range 2.5–5.5) and 17 out of 20 for methodological quality (range 15–20). *Conclusions:* Although the models scored relatively well on methodological quality, none of them can be currently considered sufficiently credible or valid to be applicable in clinical practice for elderly patients. Future research should focus on external validation, addressing performance measures relevant for their intended use, and on clinical credibility including the incorporation of factors specific for the elderly population.

Keywords Prognostic models · Elderly · Intensive care · Validity · Predictive performance · Clinical credibility

Introduction

Prognostic models may serve different purposes in the intensive care unit (ICU) [1, 2]. First, they may be used for risk adjustment in benchmarking when comparing outcomes of patients admitted to different ICUs. Second, identification of high-risk or low-risk subgroups may be used for triage or for risk stratification of patients in clinical trials. Finally, prognostic models could be used to support individual decision-making (e.g. end-of-life decisions or informing patients and their families). The simplified acute physiology score (SAPS) [3], acute physiology and chronic health evaluation (APACHE) [4], and mortality prediction model (MPM) [5] families are commonly used models originally designed to predict mortality in a general adult ICU population. These "general" models, however, have also been applied in elderly and very elderly populations (defined by various thresholds on age starting from 65 years) [6].

Elderly people represent a rapidly growing distinctive subgroup of patients admitted to ICUs with higher prevalence of co-morbidity, cognitive and functional impairment, and mortality [7]. Several studies found that in elderly people not age per se, but other factors related to old age are predictive of mortality, including diagnosis, co-morbidity, and pre-morbid cognitive and functional status [8–13]. Although general models use age as a proxy of these factors, they may not sufficiently correct for them as calendar age and biological age diverge at older age. In addition, elderly patients, who are usually excluded from clinical trials, react differently to diagnostic procedures and medication than the younger population [14]. Moreover, there is some evidence suggesting that older patients in the ICU are treated differently from younger ones even when they have the same severity of illness [15]. For these reasons, new, more "specific" models incorporating these factors have been developed specifically to predict outcome in elderly and very elderly ICU patients.

There is no literature review, however, narrative or systematic, that describes these models. Of special interest is what is known about the validity of these models and whether they can be used in clinical practice. In this systematic review we aimed to answer the following research questions: (1) which prognostic models (general or specific) are validated in an elderly or very elderly ICU population, and (2) to what extent can these models reliably be used for the purpose they were developed for? Aside from describing the general characteristics of such studies we resort to the literature on barriers to the implementation of prognostic models in practice in order to extract and assess relevant descriptors.

Materials and methods

Search strategy and data sources

The Scopus database (from January 1966 to June 2010) was searched on the basis of title, abstract, and keywords for research articles and reviews using the query shown in Fig. 1. Scopus comprises, among others, large bibliographic databases such as Medline and Embase. In addition to the articles retrieved by the electronic search, references of all included articles as well as the references of articles citing them were screened.

Inclusion criteria

Two reviewers (LM and JL) independently screened all titles and abstracts of research articles and reviews written in English, and applied the following inclusion criteria: (1) the study aimed to validate the performance of a prognostic model in predicting mortality and/or survival; (2) mortality and/or survival was defined for a specific endpoint (e.g. 30 days); (3) patients were admitted to the ICU; and (4) patients were at least 65 years old, termed 'elderly patients' in this review. Possible differences were resolved by consulting a third independent reviewer (AAH). Review articles were only used for reference in our discussion.

Extracted information

Each study was described by general descriptors (study design, number of included patients, age subgroups, mortality rate, model type, and stated purpose of the model) and descriptors relevant for assessing the fitness of the model for its intended use. The stated purpose of a model was described by using the following categories: (1) benchmarking, (2) identification of high- or low-risk subgroups, (3) individual decision-making.

For assessing the models' fitness for use, we took into account the main barriers reported in the literature for using prognostic models [1, 16]; these are model complexity and lack of trust in the models due to clinical (in)credibility and lack of evidence (e.g. lack of external validation). These imply the following prerequisites for useful models as reported elsewhere [17–19]: clinical credibility and clinical effectiveness, good methodological quality, external validation, and good performance. We summarized these aspects in the following categories: (1) clinical credibility (including model complexity), (2) methodological quality, (3) external validation, (4) model performance, and (5) clinical effectiveness (evidence that the model has a positive impact in practice). The same two reviewers who conducted the screening process (LM and JL) together scored the included studies on these categories. These reviewers were not involved in any of these reviewed studies.

Clinical credibility

To measure clinical credibility, we used the following criteria based on those reported elsewhere [1, 16]: (1) elderly-specific factors (e.g. co-morbidity, pre-morbid cognitive and functional status) are included, (2) there are no abrupt risk changes due to the use of thresholds to categorize continuous data, (3) data are obtained in a timely fashion (i.e. available before a decision is to be made), (4) data are obtained reliably (i.e. variables are not

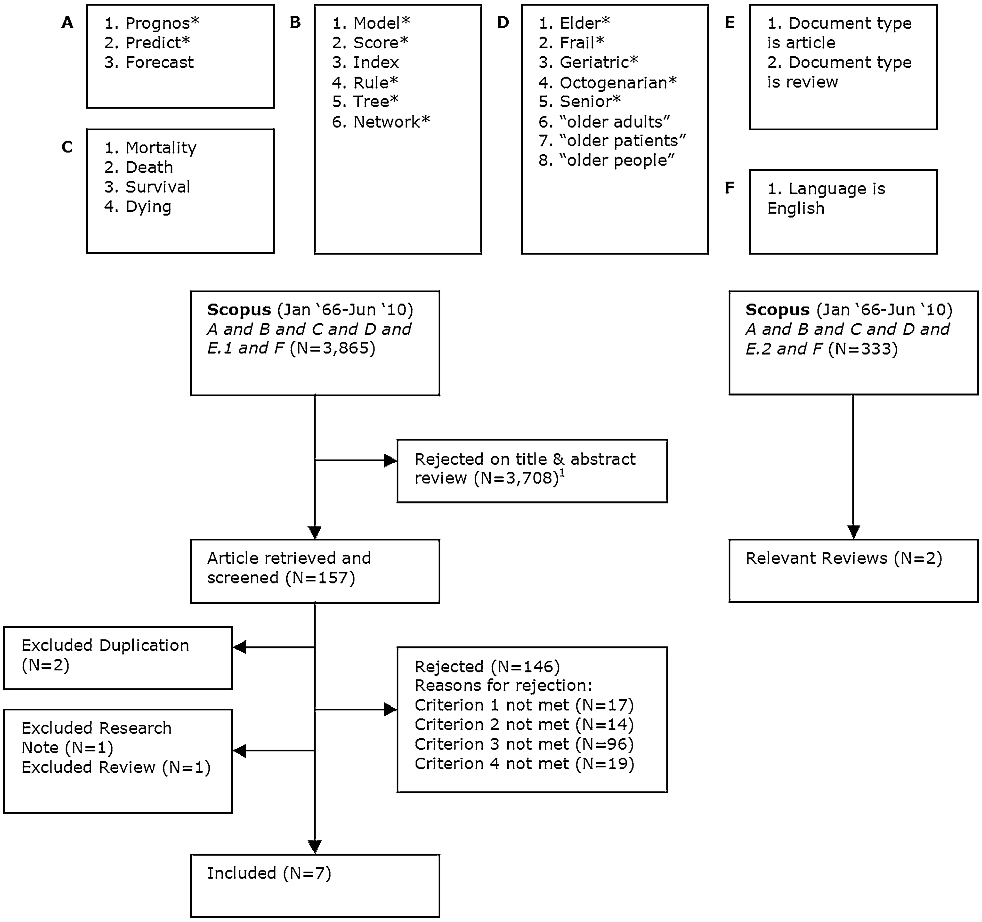
A
1. Prognos*
2. Predict*
3. Forecast
B
1. Model*
2. Score*
3. Index
4. Rule*
5. Tree*
6. Network*
C
1. Mortality
2. Death
3. Survival
4. Dying
D
1. Elder*
2. Frail*
3. Geriatric*
4. Octogenarian*
5. Senior*
6. "older adults"
7. "older patients"
8. "older people"
E
1. Document type is article
2. Document type is review
F
1. Language is English
Scopus (Jan '66-Jun '10) A and B and C and D and E.1 and F (N=3,865)
Scopus (Jan '66-Jun '10) A and B and C and D and E.2 and F (N=333)
Rejected on title & abstract review (N=3,708)1
Article retrieved and screened (N=157)
Relevant Reviews (N=2)
Excluded Duplication (N=2)
Excluded Research Note (N=1)
Excluded Review (N=1)
Rejected (N=146)
Reasons for rejection:
Criterion 1 not met (N=17)
Criterion 2 not met (N=14)
Criterion 3 not met (N=96)
Criterion 4 not met (N=19)
Included (N=7)

Fig. 1 Search strategy. The items within the *boxes A–E* are connected by the OR operator. [1]Main reasons for rejection: description of disease course or specific population rather than predicting patients' outcomes; not specific for a population of elderly patients; and predicted outcome is not the desired one (e.g. falls, functional decline, time to death)

subjective), (5) predictions are easy to generate, (6) predictions are obtained in an understandable fashion (i.e. the described model is not a "black box"). We added to this list the following criterion from 2000 [19]: (7) the range of predictions (minimum and maximum) in the sample is described (if the maximal prediction is low or the range of predictions is small, the prognostic information provided by the model may be too weak to influence individual treatment decisions). Although clinical credibility is influenced by methodological quality, we describe methodological quality separately because it is also meaningful to consider it on its own.

Methodological quality

A checklist with quality criteria for prognostic studies, described by us [20] and based on work by Hayden et al. [21], was used to appraise the methodological quality of the included studies (Online Resource 1). The checklist addresses the following components: study participation (e.g. study population described and representing source population), prognostic factor measurement (e.g. prognostic factors defined and measured appropriately), outcome measurement (e.g. outcome defined), and analysis (e.g. appropriate description and implementation of analysis).

External validation

To assess the extent to which newly developed models are being used or validated by others, we reported (1) whether the study developed a new model or externally validated an existing model, (2) whether newly developed models were validated by others, and (3) the number of times studies were cited by others (as a proxy, albeit weak, for the relevance of the models).

Model performance

We reported the performance of each model as described in the original study. An explanation of the meaning and interpretation of the performance measures used to describe model performance can be found in Online Resource 2.

Clinical effectiveness

This relates to whether there is evidence, obtained by impact studies, showing that the models brought positive change in either the process (e.g. how decisions are made) or patient outcome. We checked whether there were (positive) impact studies reported for the given models. To this end we searched all articles referring to the respective models for impact studies.

Results

Search results

Of 3,865 articles initially identified, seven met the inclusion criteria and were included in this review (Fig. 1). No additional articles were found after cross-referencing. Two relevant reviews were found that were used for reference in our discussion. No differences in inclusion decisions between the two reviewers arose, and they were always able to reach consensus in scoring the included studies.

Study characteristics

The characteristics of the included studies were heterogeneous (Table 1). In three prospective [6, 22, 23] and four retrospective [24–27] studies, cohorts ranged from 148 to 12,993 patients, minimum age of participants from 65 to 85 years, and hospital mortality from 6.3 to 54.8%. Two studies did not define outcome as hospital mortality, but used 30-day mortality [24] and 1-year survival [23] instead. Finally, two studies focused on a subgroup of elderly ICU patients, *Clostridium difficile*-associated disease (CDAD) [24] and pneumonia [23].

Reported models

In the seven included studies, a total of 17 models are validated. Of these, six models were developed for a general ICU population [6, 25, 26] and 11 models specifically for an elderly ICU population [22–27]. One of the general models was adjusted for the elderly population in which it was validated [26]. In two studies general and specific models were directly compared to each other [25, 26], in one study only general models were validated [6], and in four studies only specific models [22–24, 27]. Six general and seven specific models were developed by logistic regression [6, 22–27], one specific model by recursive partitioning [26], and three by PRIM [25] (for an explanation of these techniques see Online Resource 2).

The stated purpose of the models was benchmarking in two studies [25, 26], identification of high-risk subgroups in five studies [6, 22, 23, 25, 26], and support of individual decision-making in seven studies [6, 22–27].

Clinical credibility

None of the 17 models fulfilled all 7 criteria for clinical credibility (Table 2). The median score was 4.5 (out of 7), the highest score was 5.5 in three models, and the lowest was 2.5 in one model. Fifteen models met less than five criteria, and two models less than four. Criterion 1 (inclusion of elderly-specific factors) was the least often met (only in five models), followed by criterion 2 (no use of arbitrary thresholds to categorize continuous data; met in 6 models).

Methodological quality

Table 3 shows the results of checking the included studies against a set of 20 quality criteria. The number of criteria that were met ranged from 15 to 20; the median was 17. Although all studies fulfilled most criteria in the other three components, two studies [22, 27] did not meet three or more criteria in the analysis component. Low scores in methodological quality mainly concerned validation performed on the developmental set itself [22], lack of comparison to a reference model [22, 24–27], and not measuring performance in terms of both discrimination and calibration [22, 24, 25, 27].

External validation

One study externally validated existing models, e.g. SAPS-II, APACHE-II, and MPM-II [6]. The other six studies either developed one or more new specific models [22, 24–27] or internally validated a model they developed earlier [23]. Newly developed models were never validated by others. The number of times models were cited by others ranged from 0.3 to 4 per year.

Model performance

Model performance is summarized in Table 4. Most studies [6, 22–24, 26] measured performance in terms of discrimination by calculating the area under the receiver operating characteristic curve (AUC; five studies), whereas one study [27] used the Goodman–Kruskal γ statistic to measure discrimination. Moreover, three studies [6, 23, 26] measured calibration using the Hosmer–Lemeshow statistics, two studies [25, 26] calculated the positive predictive value (PPV), and one study [26] measured accuracy by calculating Brier scores. Note that the PPV is dependent on the cut-off point used.

Table 1 Summary of included studies

Study (country)	Study design (period)	*N*/age (subgroup)	Mortality (%)	Outcome	Model type	Model description	Model validation	Own model	Validated by others	Number of times cited by others[a] (converted per year)	Stated purpose
Zilberberg et al. [24] (USA)	Retrospective (Jan 04 to Dec 05)	148/65+ (CDAD)	45.3	30-day mortality	LR	Age 75+^{A}, lack of history of CRD, lack of leukocytosis, presence septic shock, APACHE-II 20+^{B}	Internal validation using bootstrapping	Yes	No	3 (3)	Individual decision-making
Nannings et al. [25] (NL)	Retrospective (Jan 97 to Oct 05)	12,993/80+	34.5	Hospital mortality	PRIM	Subgroup A: 24 h urine production < 0.83 l, mechanical ventilation at 24 h after admission, lowest systolic blood pressure during first 24 h < 75 mmHg, lowest pH during first 24 h < 7.3, and medical or unscheduled surgical reason for admission^{C}	Internal validation using an independent validation set	Yes	No	3 (1.5)	Benchmarking Identification of high-risk subgroups Individual decision-making
						Subgroup B: lowest systolic blood pressure during first 24 h < 70 mmHg, 24 h urine production < 0.9 l, and lowest pH value during first 24 h < 7.3 or >7.6^{C}		Yes	No		
						Subgroup C: GCS < 5^{C}		Yes	No		
					LR	SAPS-II^{D} on all subgroups		No	–		
de Rooij et al. [26] (NL)	Retrospective (Jan 97 to Dec 03)	6,867/80+	31.3	Hospital mortality	Classification tree	GCS, planned/unplanned surgery, 24 h urine, bicarbonate, urea, syst ABP, mechanical ventilation 24 h after adm, pH^{C}	Internal validation using an independent validation set	Yes	No	3 (1)	Benchmarking Identification of high-risk subgroups Individual decision-making
					LR	SAPS-II^{D}		No	–		
						Recalibrated SAPS-II^{D}		No	–		
Torres et al. [22] (ES)	Prospective (Mar 00 to Oct 00)	412/65−, 65–80,80+	7.8	IMCU mortality	LR	APACHE-II^{B}, TISS-28, Barthel index, diagnosis, stroke	Apparent validation on training set	Yes	No	16 (4)	Individual decision-making
			6.3	Hospital mortality		APACHE-II^{B}, TISS-28, diagnosis, stroke		Yes	No		
			25	2-year mortality after discharge		Charlson index, age^{E}		Yes	No		
			33.5	2-year + hospital mortality		Age^{E}, APACHE-II, Barthel index, Charlson index, TISS-28		Yes	No		
Nierman et al. [27] (USA)	Retrospective (Jan 96 to Dec 97)	455/85+	24.6	Hospital mortality	LR	Age^{E}, gender, independent ADL, assistance with ADL, type of ICU, heart rate at ICU admission, count of organ system failures, count of ICU procedures	Internal validation using an independent validation set	Yes	No	35 (3.89)	Individual decision-making
Sikka et al. [6] (USA)	Prospective (Jun 96 to Sep 99)	253/75+ (pneumonia)	54.8	Hospital mortality	LR	SAPS-II^{D}	External validation	No	–	16 (1.6)	Individual decision-making
						APACHE-II^{B}		No	–		
						MPM-II^{E}		No	–		

Table 1 continued

Study (country)	Study design (period)	*N*/age (subgroup)	Mortality (%)	Outcome	Model type	Model description	Model validation	Own model	Validated by others	Number of times cited by others[a] (converted per year)	Stated purpose
Jandziol and Ridley [23] (UK)	Prospective (Apr 95 to Sep 96)	555/70+	43	1-year survival	LR	Extreme age (>85 years)[A], diagnostic category, acute physiological derangement	External validation	Yes	–	3 (0.3)	Individual decision-making

USA United States, *NL* the Netherlands, *ES* Spain, *UK* United Kingdom, *N* number of participants, *CDAD Clostridium difficile*-associated disease, *IMCU* intermediate care unit, *LR* logistic regression, *PRIM* patient rule induction method, *CRD* chronic respiratory disease, *APACHE* acute physiology and chronic health evaluation, *GCS* Glasgow coma score, *SAPS* simplified acute physiology score, *syst ABP* systolic arterial blood pressure, *adm* admission, *TISS* therapeutic intervention scoring system, *ADL* activities of daily living, *ICU* intensive care unit, *MPM* mortality prediction model

[a] Scopus, date 20 July 2010

Superscript capitals in the column model description denote the way in which age is represented in the model

[A] Age is represented as a dichotomous variable by using a specific cut-off point

[B] Age is represented as a discrete variable by the following 5 categories: ≤44, 45–54, 55–64, 65–74, ≥75

[C] Age is not included in the model itself, but the model is restricted to patients aged 80 years and older

[D] Age is represented as a discrete variable using the following 6 categories: <40, 40–59, 60–69, 70–74, 75–79, ≥80

[E] Age is represented as a continuous variable

Clinical effectiveness

There were no reported studies on the impact of any of the models on either clinical process or patient outcomes.

Discussion

In the seven included studies, 17 models were found that were subject to validation for risk estimation in an elderly or very elderly ICU population (defined by different age thresholds starting from 65 years). Of these, six are general (i.e. developed for a general adult ICU population) and 11 are specific models (i.e. developed specifically for an elderly ICU population). The added value of specific models with respect to general ones is addressed in only two studies that did not find differences [25, 26]. The specific models addressed in these two studies did not include elderly-specific factors, such as pre-morbid cognitive and functional status or co-morbidity, however.

To assess whether models could be used for the purpose they were developed for, we focused on clinical credibility of the model, methodological quality of the study, external validity, the model's reported performance, and clinical effectiveness. Generally, although the studies score relatively well on methodological quality (most quality criteria are met by most studies with a median of 17 out of 20), there are opportunities for improvement in clinical credibility (median number of criteria met is 4.5 out of 7). In addition, external validation is scarce and done only for the general models in small patient samples. Moreover, most model development studies were rarely cited by others and there are no studies about the impact of models on either clinical practice or patient outcomes.

The performance of models can be considered reasonable, but most studies did not apply a set of performance measures covering more than one aspect of model performance. All studies reported individual decision-making as a possible use of prognostic models, of which one study only measured discrimination, whereas another only measured positive predictive values. Although the importance of calibration and accuracy may be debatable in individual decision-making, discrimination between survivors and non-survivors is obviously essential. Two studies intended to use prognostic models for benchmarking, but only one of them measured performance in terms of discrimination, calibration, and accuracy. Moreover, three studies did not describe the range of predicted probabilities, and none of them used the R^2 measure. R^2, which is equal to $1 - \text{Brier score}/[M \times (1 - M)]$, where M is the overall mortality risk, adjusts the Brier score to the prevalence of mortality. For these reasons, there is still no sufficient evidence that these models can be used for these purposes. Exceptions are the models evaluated by de Rooij et al. [26] and Sikka et al. [6]

Table 2 Criteria for clinical credibility

Study	Model	Elderly-specific factors are included	No abrupt risk changes due to the use of thresholds	Data obtained in timely fashion	Data obtained reliably, i.e. not subjective	Easy to generate predictions	Understandable model[a]	Prediction range described
Zilberberg et al. [24]	LR	–	–	±	±	+	±	–
Nannings et al. [25]	Subgroup A	–	–	+	+	+	+	NA
	Subgroup B	–	–	+	+	+	+	NA
	Subgroup C	–	–	+	±	+	+	NA
	SAPS-II	–	–	+	+	+	±	+
de Rooij et al. [26]	Classification tree	–	–	+	±	+	+	+
	SAPS-II	–	–	+	+	+	±	+
	Recalibrated SAPS-II	–	–	+	+	+	±	+
Torres et al. [22]	LR 1	+	+	+	–	+	±	–
	LR 2	+	+	+	±	+	±	–
	LR 3	+	+	+	–	+	±	–
	LR 4	+	+	+	–	+	±	–
Nierman et al. [27]	LR	+	+	+	–	+	±	–
Sikka et al. [6]	SAPS-II	–	–	+	+	+	±	+
	APACHE-II	–	–	+	+	+	±	+
	MPM-II	–	+	+	+	+	±	+
Jandziol and Ridley [23]	LR	–	–	+	+	+	±	+

APACHE acute physiology and chronic health evaluation, *LR* logistic regression, *MPM* mortality prediction model, *NA* not applicable, *SAPS* simplified acute physiology score

[a] + for symbolic models, ± for (generalized) linear models, – for non-linear "black box" models

Table 3 Quality criteria

	Zilberberg et al. [24]	Nannings et al. [25]	de Rooij et al. [26]	Torres et al. [22]	Nierman et al. [27]	Sikka et al. [6]	Jandziol and Ridley [23]
1. Study participation							
Description of setting and study period	+	+	+	+	+	+	+
Description of inclusion and exclusion criteria	+	+	+	+	+	+	+
Description of patient mix	+	+	+	±	+	±	+
Number of patients reported	+	+	+	+	+	+	+
Number of patients >100	+	+	+	+	+	+	+
Mortality rate reported	+	+	+	+	+	+	+
Description of patient characteristics	+	+	+	+	+	+	+
Study population represents source population	+	+	+	+	+	±	+
2. Prognostic factor measurement							
Definition of all prognostic factor(s) evaluated	+	+	+	+	+	+	±
Description of type of model(s)	+	+	+	+	+	+	±
Description of % of participants with complete data and handling of missing values	–	–	+	–	–	+	–
3. Outcome measurement[a]							
Definition of outcome of interest	+	+	+	+	+	+	+
4. Analysis							
Description of all evaluation measures	+	+	+	+	+	+	+
Description of model building strategy	+	+	+	+	+	+	±
Description of test method	+	+	+	±	+	+	+
Both aspects of discrimination and calibration evaluated	–	–	+	–	–	+	+
Separate test set used for testing	+	+	+	–	–	+	+
Sufficient presentation of data to assess adequacy of the analysis	+	+	+	+	+	+	±
No selective reporting of results	+	+	+	+	+	+	+
Comparison to reference model	–	+	+	–	–	+	–

[a] As a result of our inclusion criteria in this review all studies score one point in this component

Table 4 Model performance

Study	Model	AUC/AUC ± SD/ AUC (95% CI)	Brier	PPV/PPV (95% CI)	Hosmer–Lemeshow statistics	Goodman–Kruskal γ statistic
Zilberberg et al. [24]	LR	0.740 (0.663–0.817)				
Nannings et al. [25]	Subgroups			Subgroup A: 0.918 Subgroup B: 0.895 Subgroup C: 0.873		
	SAPS-II			Subgroup A: 0.918 Subgroup B: 0.895 Subgroup C: 0.873		
de Rooij et al. [26]	Classification tree	0.77 ± 0.01	0.16	T0.5: 0.69 (0.64–0.73) T0.7: 0.85 (0.8–0.89) T0.8: 0.88 (0.83–0.91)		
	SAPS-II	0.77 ± 0.01	0.16	T0.5: 0.68 (0.64–0.72) T0.7: 0.78 (0.73–0.82) T0.8: 0.83 (0.77–0.87)	*H* statistic = 64.3 ($p < 0.00001$) *C* statistic = 89 ($p < 0.00001$)	
	Recalibrated SAPS-II	0.77 ± 0.01	0.16	T0.5: 0.71 (0.67–0.76) T0.7: 0.81 (0.76–0.86) T0.8: 0.88 (0.81–0.92)	*H* statistic = 9.5 ($p = 0.49$) *C* statistic = 21.6 ($p = 0.02$)	
Torres et al. [22]	LR 1	0.88 (0.82–0.93)				
	LR 2	0.81 (0.75–0.87)				
	LR 3	0.77 (0.71–0.82)				
	LR 4	0.79 (0.74–0.84)				
Nierman et al. [27]	LR					Min −2.557, mean −0.487, max 2.502
Sikka et al. [6]	SAPS-II	0.752 ± 0.053			*H* statistic = 16.38 ($p < 0.05$)	
	APACHE-II	0.711 ± 0.049			*H* statistic = 19.03 ($p < 0.01$)	
	MPM-II	0.747 ± 0.054			*H* statistic = 9.87 ($p > 0.1$)	
Jandziol and Ridley [23]	LR	0.75			*H* statistic $p < 0.05$	

APACHE acute physiology and chronic health evaluation, *LR* logistic regression, *MPM* mortality prediction model, *SAPS* simplified acute physiology score, *AUC* area under the receiver operating characteristic curve, *CI* confidence interval, PPV positive predictive value, *T* threshold, e.g. T0.5 means the threshold used for calculating the PPV is 0.5, *min* minimum, *max* maximum, *SD* standard deviation

(APACHE-II, SAPS-II, MPM-II and a classification-tree model), but extensive external validation of these models is still needed in the elderly patient population.

Our study has the following strengths. First, to our knowledge, this is the first systematic review providing an overview of prognostic models used to predict mortality or survival in an elderly ICU population. Second, we focused on what we believe are prerequisites for appropriate choice and use of prognostic models in clinical practice: methodological quality of the study, clinical credibility and external validity of the model, whether performance measures relevant to the intended use were reported, and model performance. Third, we used extensive search criteria to identify relevant studies. Finally, a comprehensive set of quality criteria was used for methodological appraisal of studies. Limitations of our study include the fact that, although based on literature, our checklist for clinical credibility and methodological quality inevitably suffers a certain degree of subjectivity. We provided, however, a comprehensive overview of how prognostic research can be improved in order to support the wider acceptance of prognostic models. Another limitation is that model use in practice may be influenced not only by methodological quality and credibility but also by various implicit social, organizational, or other aspects which cannot be assessed from the published studies themselves.

Two other studies reviewed the literature about prognostic tools or variables for outcome prediction in critically ill elderly patients. However, these studies either focused on risk factors alone [9] or focused on scoring systems predicting outcome in patients that are not yet admitted to the ICU [28].

Two recent reviews by Mallett et al. [17] assessed the reporting of methods in studies developing prognostic models in cancer (no age restrictions) and the reporting of their performance [18]. Although they studied a different patient population, they also found that external validation of models is scarce. They also reported on findings that do not concord with ours. Using quality criteria comparable to those used in our checklist, they found that methodology and reporting of methods and model performance were generally poor. Apparently, reporting of prognostic models is of higher quality in the field of critical care than in the field of cancer.

Accepting any of the included models in clinical practice for either benchmarking, high-risk subgroup selection, or individual decision-making in the elderly ICU population is still premature as both external validation and evidence for the validity of the models for their intended use are very scarce. To help alleviate these problems, assuming the goal is to employ the models in practice, researchers should better consider the model's intended use and report on the appropriate performance measures, but preferably also using a set of performance measures covering aspects of accuracy (e.g. Brier score, R^2), discrimination (e.g. AUC), as well as calibration (e.g. Hosmer–Lemeshow statistics). Moreover, there is a need for external validation studies (preferably conducted independently by different researchers in a different research setting as authors tend to confirm the validity of their own models [19]) and studies on model acceptance and their impact on clinical practice (i.e. are they clinically effective?). These kinds of studies could increase physicians' trust and knowledge on how to use prognostic models, and thereby reduce barriers for using models in clinical practice [1]. It should be noted, however, that the older the individual patient becomes, the smaller their proportion is in the developmental and/or validation samples. Thus the predictions for these patients will be less reliable.

Besides lack of validation studies, lack of physicians' trust in the models is also caused by lack of clinical credibility [16]. Indeed, we found caveats in elements that determine clinical credibility. First, most models did not include elderly-specific factors (e.g. premorbid cognitive and functional status or co-morbidity), although there is increasing evidence that in elderly patients not age per se, but elderly-specific factors are predictive of mortality [8–13]. Useful instruments to measure these include the Katz [29] or Barthel [30] activities of daily living (ADL) index, the informant questionnaire on cognitive decline in the elderly, short form (IQCODE-sf) [31], and the Charlson co-morbidity index [32], although these may be hard to collect accurately for all patients, especially in an ICU environment. Moreover, predictions may improve when including information arising during ICU stay, such as the sequential organ failure assessment (SOFA) scores [20]. Second, more than half of the models used arbitrary thresholds for categorizing continuous variables, such as age, even though, for example, it would be unlikely that the prognosis for someone aged 75 years and 0 days would be very different from someone aged 74 years and 364 days, but they would fall in different categories if the threshold is 75 years [16]. Third, some data are obtained subjectively, but admittedly these mostly include elderly-specific factors that are sometimes hard to measure objectively. Fourth, most studies did not report the range of probabilities provided by their models, although, for example, there is no use for a model providing a range of probabilities between 0.1 and 0.3 if the intention is to inform individual treatment decisions pertaining to survival [19]. In addition, predictions that markedly deviate from the overall risk are more useful than probabilities that are close to it.

Focus of future research should include external validation and improving existing models, and not only the development of new models. Models for the elderly patients may, however, be improved by adding pre-morbid cognitive and physical status as additional predictors, and/or short-term post-discharge cognitive and physical status as additional outcome measures [9, 33]. In addition, for purposes of supporting individual decision-making at the ICU, it could be beneficial to use information (e.g. SOFA scores) about patients during their entire stay (as physicians do) rather than using data collected only in the first 24 h of admission, which are collected mainly for risk adjustment when comparing patient outcomes in different ICUs.

Previous research has shown that patients' preferences towards life-sustaining treatments are highly dependent on their probabilities on a good outcome [34, 35]. In fact, patients often prefer palliative care aiming at comfort and relief of pain if chances for survival are very low or if survival is associated with high burden. Future research should focus on the acceptability by doctors, patients, and their families of different forms of prognostic information and whether this might influence the process and outcome of decision-making.

Conclusion

Current prognostic models predicting mortality in the elderly and very elderly ICU populations are not mature enough for use in clinical practice either because they are not credible enough and/or not yet extensively validated. This applies for any of their reported uses: benchmarking, high-risk subgroup selection, or individual decision-making. There is no evidence that elderly-specific models (e.g. models developed specifically for an elderly or very elderly ICU population) perform better than general models (e.g. models developed for a general adult ICU population), but direct comparison of specific and general models is scarce. Moreover, elderly-specific factors (e.g. co-morbidity, and cognitive and functional status) are not included in these elderly-specific models, although there is increasing evidence that these, and not age per se, are predictive of mortality in elderly people.

To obtain wider acceptance of prognostic models the focus of future research should be on external validation, addressing the performance measures relevant for their intended use, and on models' clinical credibility, including the incorporation of factors specific for the elderly population. In addition, it might be useful to use information about patients during their entire stay (e.g. SOFA scores) and to predict cognitive and physical status next to survival. Finally, future research could focus on the question whether doctors, patients, and their families value receiving more tailored prognostic information, in which form, and whether provision of this information influences the process and outcome of their decision-making.

Conflict of interest The author(s) declare that they have no competing interests, but they note that some co-authors of this review were co-authors of two of the reviewed studies. These co-authors, however, were not involved in selecting and scoring the quality of reviewed studies.

References

1. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG (2009) Prognosis and prognostic research: what, why, and how? BMJ 338:b375
2. Steyerberg EW (2009) Clinical prediction models: a practical approach to development, validation, and updating. Springer, Berlin
3. Le Gall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA 270:2957–2963
4. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) APACHE II: a severity of disease classification system. Crit Care Med 13:818–829
5. Teres D, Lemeshow S, Avrunin JS, Pastides H (1987) Validation of the mortality prediction model for ICU patients. Crit Care Med 15:208–213
6. Sikka P, Jaafar WM, Bozkanat E, El-Solh AA (2000) A comparison of severity of illness scoring systems for elderly patients with severe pneumonia. Intensive Care Med 26:1803–1810
7. US Bureau of the Census, Cheeseman Day J (1993) Population projections of the United States, by age, sex, race, and hispanic origin: 1993 to 2050. Current population reports, US Government Printing Office, Washington, pp 25–1104
8. Ricou B, Merlani P (2008) What limits for acute care in the elderly? Curr Opin Anaesthesiol 21:380–385
9. de Rooij SE, Abu-Hanna A, Levi M, de Jonge E (2005) Factors that predict outcome of intensive care treatment in very elderly patients: a review. Crit Care 9:R307–R314
10. Boumendil A, Somme D, Garrouste-Org Guidet B (2007) Should elderly patients be admitted to the intensive care unit? Intensive Care Med 33:1252–1262
11. Ryan D, Conlon N, Phelan D, Marsh B (2008) The very elderly in intensive care: admission characteristics and mortality. Crit Care Resusc 10:106–110
12. Chelluri L, Grenvik A, Silverman M (1995) Intensive care for critically ill elderly: mortality, costs, and quality of life. Review of the literature. Arch Intern Med 155:1013–1022
13. Rodriguez-Reganon I, Colomer I, Frutos-Vivar F, Manzarbeitia J, Rodriguez-Manas L, Esteban A (2006) Outcome of older critically ill patients: a matched cohort study. Gerontology 52:169–173
14. Spinewine A (2008) Adverse drug reactions in elderly people: the challenge of safer prescribing. BMJ 336:956–957
15. O'Brien JM Jr, Aberegg SK, Ali NA, Diette GB, Lemeshow S (2009) Results from the national sepsis practice survey: predictions about mortality and morbidity and recommendations for limitation of care orders. Crit Care 13:R96
16. Wyatt JC (1995) Prognostic models: clinically useful or quickly forgotten? BMJ 311:1539–1541
17. Mallett S, Royston P, Dutton S, Waters R, Altman DG (2010) Reporting methods in studies developing prognostic models in cancer: a review. BMC Med 8:20
18. Mallett S, Royston P, Waters R, Dutton S, Altman DG (2010) Reporting performance of prognostic models in cancer: a review. BMC Med 8:21
19. Altman DG, Royston P (2000) What do we mean by validating a prognostic model? Stat Med 19:453–473
20. Minne L, Abu-Hanna A, de Jonge E (2008) Evaluation of SOFA-based models for predicting mortality in the ICU: a systematic review. Crit Care 12:R161
21. Hayden JA, Cote P, Bombardier C (2006) Evaluation of the quality of prognosis studies in systematic reviews. Ann Intern Med 144:427–437
22. Torres OH, Francia E, Longobardi V, Gich I, Benito S, Ruiz D (2006) Short-and long-term outcomes of older patients in intermediate care units. Intensive Care Med 32:1052–1059
23. Jandziol AK, Ridley SA (2000) Validation of outcome prediction in elderly patients. Anaesthesia 55:107–112
24. Zilberberg MD, Shorr AF, Micek ST, Doherty JA, Kollef MH (2009) Clostridium difficile-associated disease and mortality among the elderly critically ill. Crit Care Med 37:2583–2589
25. Nannings B, Abu-Hanna A, de Jonge E (2008) Applying PRIM (patient rule induction method) and logistic regression for selecting high-risk subgroups in very elderly ICU patients. Int J Med Inform 77:272–279
26. de Rooij SE, Abu-Hanna A, Levi M, de Jonge E (2007) Identification of high-risk subgroups in very elderly intensive care unit patients. Crit Care 11:R33
27. Nierman DM, Schechter CB, Cannon LM, Meier DE (2001) Outcome prediction model for very elderly critically ill patients. Crit Care Med 29:1853–1859

28. Hudson S, Boyd O (2007) Criteria for admission to the ICU and scoring systems for severity of illness. Surgery 25:117–121
29. Katz S, Ford AB, Moskowitz RW, Jackson BA, Jaffe MW (1963) Studies of illness in the aged: the new index of ADL: a standardized measure of biological and psychological function. JAMA 185:914–919
30. Mahoney FI, Barthel DW (1965) Functional evaluation: the Barthel index. Md State Med J 14:61–65
31. Jorm AF (1994) A short form of the informant questionnaire on cognitive decline in the elderly (IQCODE): development and cross-validation. Psychol Med 24:145–153
32. Charlson ME, Pompei P, Ales KL, MacKenzie CR (1987) A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 40:373–383
33. Malani PN (2009) Functional status assessment in the preoperative evaluation of older adults. JAMA 302:1582–1583
34. Fried TR, Bradley EH, Towle VR, Allore H (2002) Understanding the treatment preferences of seriously ill patients. N Engl J Med 346:1061–1066
35. Murphy DJ, Burrows D, Santilli S, Kemp AW, Tenner S, Kreling B, Teno J (1994) The influence of the probability of survival on patients' preferences regarding cardiopulmonary resuscitation. N Engl J Med 330:545–549