Epidemiol. Infect. (2008), **136**, 1628–1637. © 2008 Cambridge University Press
doi:10.1017/S0950268808000435 Printed in the United Kingdom

Predicting mortality with pneumonia severity scores: importance of model recalibration to local settings

P. SCHUETZ[1,2*†], M. KOLLER[2†], M. CHRIST-CRAIN[1], E. STEYERBERG[3], D. STOLZ[4], C. MÜLLER[1], H. C. BUCHER[2], R. BINGISSER[1], M. TAMM[4] AND B. MÜLLER[1]

[1] *University Hospital Basel, Department of Internal Medicine, Basel, Switzerland*
[2] *Basel Institute for Clinical Epidemiology, Basel, Switzerland*
[3] *Department of Public Health, Erasmus MC, Rotterdam, The Netherlands*
[4] *Clinic of Pneumology and Pulmonary Cell Research, University Hospital Basel, Switzerland*

(*Accepted 31 January 2008; first published online 27 February 2008*)

SUMMARY

In patients with community-acquired pneumonia (CAP) prediction rules based on individual predicted mortalities are frequently used to support decision-making for in-patient *vs.* outpatient management. We studied the accuracy and the need for recalibration of three risk prediction scores in a tertiary-care University hospital emergency-department setting in Switzerland. We pooled data from patients with CAP enrolled in two randomized controlled trials. We compared expected mortality from the original pneumonia severity index (PSI), CURB65 and CRB65 scores against observed mortality (calibration) and recalibrated the scores by fitting the intercept α and the calibration slope β from our calibration model. Each of the original models underestimated the observed 30-day mortality of 11%, in 371 patients admitted to the emergency department with CAP (8·4%, 5·5% and 5·0% for the PSI, CURB65 and CRB65 scores, respectively). In particular, we observed a relevant mortality within the low risk classes of the original models (2·6%, 5·3%, and 3·7% for PSI classes I–III, CURB65 classes 0–1, and CRB65 class 0, respectively). Recalibration of the original risk models corrected the miscalibration. After recalibration, however, only PSI class I was sensitive enough to identify patients with a low risk (i.e. <1%) for mortality suitable for outpatient management. In our tertiary-care setting with mostly referred in-patients, CAP risk scores substantially underestimated observed mortalities misclassifying patients with relevant risks of death suitable for outpatient management. Prior to the implementation of CAP risk scores in the clinical setting, the need for recalibration and the accuracy of low-risk re-classification should be studied in order to adhere with discharge guidelines and guarantee patients' safety.

INTRODUCTION

Community-acquired pneumonia (CAP) is the leading cause of death from infectious diseases in western countries and health expenditures in particular for in-patient management of patients with CAP are substantial [1, 2]. Accurate assessment of disease severity, risk stratification and prediction of outcome are, therefore, prerequisites for the safe identification of patients with CAP at low risk of complications and thus suitable for outpatient management. Several international organizations have developed prediction rules and adopted guidelines to stratify patients

* Author for correspondence: Dr.med. P. Schuetz, Department of Internal Medicine and Basel Institute for Clinical Epidemiology (BICE), University Hospital Basel, Petersgraben 4, CH-4031 Basel, Switzerland.
(Email: schuetzp@uhbs.ch)
† These authors contributed equally to this work.

with CAP based on predicted mortalities for the identification of patients with CAP that may be managed in an outpatient setting in order to optimize hospital referral and lower hospital admission rates [3, 4]. The pneumonia severity index (PSI) is a widely propagated scoring system in North America that assesses the risk of death in a two-step algorithm [5]. The CURB65† score is the modified version of the British Thoracic Society (BTS) assessment tool which is based on only five predictors and used in Europe [6, 7]. The CRB65 score has been put forward as a useful substitute for the CURB65 as it does not rely on laboratory measurements and still shows acceptable discriminatory ability [7, 8].

Prior to the implementation of a statistically derived prediction score, an external validation within locally generated data should be conducted [9]. With only few exceptions [10], external validation studies of pneumonia severity scores have focused on discriminative properties, i.e. the ability of the score to distinguish patients with CAP and fatal outcome from those surviving [10–17]. Despite good discriminatory abilities, most validation studies found higher mortality rates of patients with PSI class III and CURB65 class 1 than was reported in the original studies. Because management strategies of patients with CAP depend on cut-off values of absolute predicted mortalities, it is essential that predicted risks agree with observed risks in the population in question. This is referred to as calibration. Miscalibration may lead to inadequate discharge of patients with high mortality (risk underestimation) or inadequate hospitalization of low-risk patients (risk overestimation).

The aim of our study was to validate the calibration and assess the need for recalibration of three well established pneumonia severity prediction scores in a tertiary-care setting in Switzerland.

METHODS

Study sample

For this analysis we pooled data from two randomized controlled studies enrolling patients with lower respiratory tract infections presenting to the emergency department (ED) of the University Hospital of Basel, Switzerland. The design of the two trials was similar and a complete description has been reported in detail elsewhere [18, 19]. In brief, the first trial included 243 consecutive patients with clinically suspected lower respiratory tract infections including acute and exacerbation of chronic bronchitis, and CAP, admitted from December 2002 until April 2003. The second trial included 302 patients with radiologically confirmed CAP admitted between November 2003 and February 2005. In both trials, patients were randomly assigned to procalcitonin-guided antibiotic therapy ($n=124$ or $n=151$, respectively) or to standard treatment according to guidelines ($n=119$ or $n=151$, respectively) [3, 4]. The aim of both trials was to study whether procalcitonin-guided antibiotic treatment can reduce the amount of antibiotic consumption and 30-day mortality was monitored as a secondary endpoint. The first trial measured procalcitonin only on admission, whilst in the second trial follow-up procalcitonin measurements were performed. For the purpose of this analysis, only patients with a definite diagnosis of CAP were considered. CAP was defined as the presence of a new infiltrate on chest radiograph accompanied by one, or several, acquired acute respiratory symptoms and signs such as cough, sputum production, dyspnoea, fever $>38{\cdot}0\ ^{\circ}C$, auscultatory findings of abnormal breath sounds and rales, leucocytosis $>10^{10}$ cells/l, or leucopenia $<4\times10^{9}$ cells/l [3]. In-patient or outpatient management of patients was not an exclusion criteria for either trial. Patients with other lower respiratory tract infections than CAP, including bronchitis or exacerbation of chronic obstructive pulmonary disease and asthma were not considered. Furthermore, patients with cystic fibrosis, active pulmonary tuberculosis, hospital-acquired pneumonia and severe immunosuppression (patients infected with human immunodeficiency virus infection and a CD4 count $<350\times10^{9}$/l, patients on immunosuppressive therapy after solid organ transplantation and neutropenic patients with a present neutrophil count $<500\times10^{9}$/l and patients under chemotherapy with neutrophils $500–1000\times10^{9}$/l with an expected decrease to values $<500\times10^{9}$/l) were not eligible for trial inclusion.

Patients were examined on admission to the ED by a medical resident supervised by a board-certified specialist in internal medicine. Baseline assessment included collection of clinical data and vital signs, comorbid conditions, and routine blood tests. All study forms were completed contemporaneously.

† CURB65, Confusion, Urea >7 mmol/l, Respiratory rate >30/min, low Blood pressure (systolic <90 mmHg or diastolic <60 mmHg). CRB65, the same as CURB65, but does not include urea.

https://doi.org/10.1017/S0950268808000435 Published online by Cambridge University Press

Since neither of the trials showed a significant difference between the intervention arm and the control arm regarding all-cause mortality (pooled OR 0·78, 95% CI 0·41–1·50, $P=0·46$), treatment assignment was not considered any further in this analysis. In addition, the coefficients of the calibration models did not differ between the intervention group and the control group of the study population for any of the three models assessed.

Both trials had been approved by the local Ethical Committees and registered in the Current Controlled Trials Database (ISRCTN04176397); all patients gave written informed consent.

Severity assessment and outcome

The PSI, CURB65 and CRB65 scores were calculated in all patients on the basis of the patients' unique set of prognostic indicators. Identical to the outcome definition of the original models [5, 7, 8], we used 30-day mortality as outcome for our validation study, which was collected in both trials as part of the trial safety monitoring.

Statistical analysis

We performed external validation of the three original models by assessing calibration and discrimination. We first studied calibration in a descriptive way by tabulating and plotting observed mortality across classes of predicted mortality as given with the original models [5, 7]. We then studied calibration in the context of a simple calibration model fitting the logit of the predicted mortality from the original models against the binary outcome (death or alive at 30 days) from our study population using logistic regression [20, 21]. This calibration model has the advantage of efficiency since it uses only two free parameters: an intercept α and a calibration slope β. In the ideal case of perfect validity, $\alpha=0$ and $\beta=1$. The parameters can be tested with ANOVA or Wald statistics. If α or β significantly deviate from the ideal case, then there is evidence of miscalibration and model recalibration should be performed [20, 21]. The recalibrated risk can be calculated as $\hat{\alpha}+\hat{\beta}*\log(\text{prob}/(1-\text{prob}))$ with probabilities (prob) originating from the original model and the coefficients originating from the calibration model (see Appendix). We did not compare observed mortality with predicted mortality within risk classes per model for all three models due to the low efficiency of this approach. For each model with five classes this approach would result in five one-sample proportion tests (the random observed mortality against the fixed predicted mortality). In fact we would need to spend 5 degrees of freedom (D.F.) per model instead of 2 D.F. with the calibration model approach and additionally we had the problem of multiple comparisons [20].

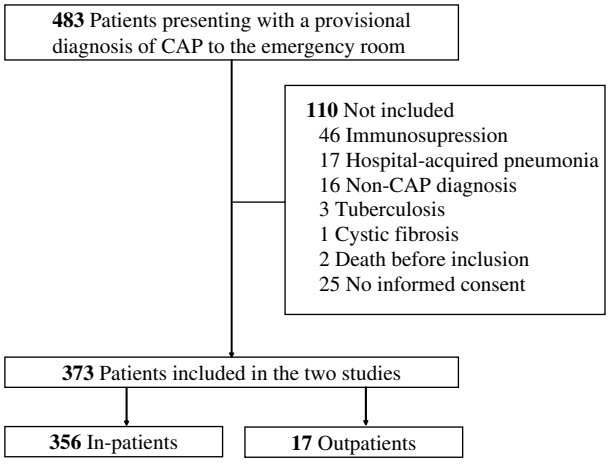

Fig. 1. Flow chart of all 373 patients included in this study. CAP, Community-acquired pneumonia.

Discrimination refers to the ability of the model to assign a higher predicted mortality to all patients with an outcome (30-day mortality) compared to patients without an outcome. We assessed discrimination using the c statistic which is equal to the area under the receiver-operating characteristics (ROC) curve. Moreover, we used the Brier score as an overall measure of model performance [22].

We used R version 2.3.1 [23] and the Design library [24] for statistical analyses.

RESULTS

Baseline characteristics

Between December 2002 and February 2005, a total of 483 (96 in the first and 387 in the second trial) consecutive patients with an initial diagnosis of CAP were screened for eligibility (Fig. 1). CAP was radiologically confirmed in 373 (87 and 286) patients who are included in this analysis. In total, 110 patients were excluded because of the use of immunosuppressive drugs ($n=46$), hospital-acquired pneumonia ($n=17$), non-CAP diagnosis ($n=16$), tuberculosis ($n=3$), cystic fibrosis ($n=1$), death before inclusion ($n=2$) or due to refusal of informed consent ($n=25$). The median age of the patients was

https://doi.org/10.1017/S0950268808000435 Published online by Cambridge University Press

73 years [inter-quartile range (IQR) 59–82 years], 84 patients (23%) were smokers with a median of 40 pack years (IQR 20–50) and 90 (24%) of the patients had an underlying chronic obstructive lung disease. Forty-nine percent of the patients ($n=184$) were randomized to receive antibiotic treatment according to procalcitonin guidance and 51% patients ($n=189$) were allocated to the control group. The majority of the patients (95·4%) were treated as in-patients with a median length of hospital stay of 11 days (IQR 6–17 days). Outpatients were predominantly in low-risk classes of PSI score (53% in class I, 29% in class II) and of CURB65 and CRB65 scores (76% each in class 0). Baseline characteristics of our study population are summarized in Table 1. For comparison, the characteristics of the study populations where the original models were developed are also given in Table 1 [5, 7].

Calibration of the three different rules

Overall, 41/373 patients died (4/96 and 37/373) and the overall mortality was 11%. The proportion of intensive care unit (ICU) admission was 8% (8/96) in the first and 10% (37/373) in the second study.

Table 2 and the calibration plots in Figure 2 illustrate predicted mortality from the original models against observed mortality within classes of predicted mortality. Compared to the observed mortality of 11%, the predicted average 30-day mortality was underestimated with each of the original models (8·4% for PSI, 5·5% for CURB65, 5·0% for CRB65, respectively). Importantly, within the low-risk classes of the original models we observed relevant mortalities of 2·6%, 5·3%, and 3·7% (PSI classes I–III, CURB65 classes 0–1, and CRB class 0, respectively) (Table 2). Low-risk predictions from each of the original models were therefore on average four times underrated compared to the observed mortalities within low-risk classes. The risk estimates in the high-risk classes were more accurate for all scores. The PSI score, but not the CURB65 and CRB65 scores slightly overestimated the risk of death in the highest risk class (Fig. 2). The four patients misclassified in the PSI and the three patients misclassified in the CURB65 score were younger [median age 67 years (IQR 62–75) and 59 years (IQR 57–61)] compared to correctly classified non-survivors [median age 79 years (IQR 70–86)].

Calibration models showed significant miscalibration of the β slope for all scores ($P<0·001$ each) underlining the necessity of recalibration (see Appendix). Calibration plots in Figure 2 show the impact of recalibration of the original models: the recalibrated mortalities are in good agreement with the observed 30-day mortality and therefore show good calibration for each model. Details on the calculation of recalibrated mortalities are given in a worked example in the Appendix. In the low-risk classes, recalibration corrected risk estimation for the PSI score (classes I–III) from 0·5% to 2·7%, for the CURB65 score (classes 0 and 1) from 1·2% to 5% and for the CRB65 score (class 0) from 0·9% to 3·5%, compared to the observed mortality of 2·5%, 5·3% and 3·7% in the corresponding risk category of each model. Within our tertiary-care setting of high-risk patients, only the recalibrated PSI was an adequate tool for the identification of low-risk patients with a predicted mortality in the low range of 1%.

Patients in the lowest risk classes of recalibrated CURB65 and CRB65 scores still had relevant mortality rates of 3·6% and 3·5% (Table 3) respectively. Only the recalibrated PSI class I score was therefore adequate for classifying patients as low risk in the range of 1%. Table 3 shows the identification of low-risk patients according to the original and the recalibrated PSI score. The original model classified 162 patients (43%) as low risk (classes I–III) with an observed mortality of 2·5%. The recalibrated PSI class I identified 41 patients (11%) as low risk with an observed mortality of 0%. Accordingly, the sensitivities and specificities of the original PSI risk model were 90·2% and 47·3%, and 100% and 12·3% for the recalibrated PSI score, respectively.

Discriminatory ability

We performed ROC analysis to assess the discriminatory ability of the three prognostic scores (Fig. 3). The PSI score had an area under the curve (AUC) of 0·72 (95% CI 0·65–0·78). The respective values for CURB65 and CRB65 scores were 0·69 (95% CI 0·61–0·77) and 0·66 (95% CI 0·58–0·73). The corresponding Brier score was 0·094, 0·096 and 0·098 for PSI, CURB65 and CRB65 indicating the lowest prediction error for the PSI score. Recalibration did not numerically affect the discriminatory ability of the models.

DISCUSSION

We performed an external validation study of three well established mortality prediction rules in 373

https://doi.org/10.1017/S0950268808000435 Published online by Cambridge University Press

Table 1. *Baseline characteristics of the study population*

Baseline characteristics	Study cohort ($n=373$)	PSI cohort ($n=2287$)	CURB65 cohort ($n=1068$)
Demographics			
Age, years*,	73 (59–82)		69 (17–100)
<0 yr (%)†	16·7	42·7	18·8
Male sex (%)	60·1	50·0	51·5
Nursing home resident (%)		8·5	0
Coexisting illnesses (%)			
Heart disease	42·4	11·1	18
Cerebrovascular disease	5.4	9·2	9
Renal dysfunction	27·1	6·7	
Liver disease	10·5	1·4	1
Chronic lung disease	24·1		35
Neoplastic disease	14·2	5·8	
Clinical signs			
Altered mental status	8·3	10·4	11·7
Rales (%)	72·3		
Systolic blood pressure (mmHg)*	130 (112–142)		
<90 mmHg (%)†	3·2	2·1	18·6
Pulse rate (bpm)*	95 (81–110)		
>124 /min (%)†	6·4	8·7	8·1
Respiratory rate	22 (20–28)		
>29/min (%)†	16·9	13·3	25·9
Temperature (°C)*	38·4 (37·6–39·2)		
<35 °C, >39·9 °C (%)†	7·8	1·6	
Pleural effusion (%)	17·2	8·9	10·9
Laboratory parameters			
Arterial pH	7·43 (7·42–7·44)		
<7·35 (%)†	4·0	3·7	
Urea (mmol/l) (%)	6·8 (4·9–11·5)		
>7 mmol/l†	47·7		33·5
>10·9 mmol/l†	27·3	14·3	
Sodium (mmol/l)	135·6 (133·0–138·0)		
<130 mmol/l (%)†	11·0	3·9	
Glucose (mmol/l)	6·6 (5·6–8·2)		
>13·9 mmol/l (%)†	4·6	4·2	
Haematocrit (%)	38 (35–40)		
<30 % (%)†	4·8	6·3	
PaO_2 (mmHg)	60·7 (58·5–63)		
<60 mmHg (%)†	27·3	20·6	25·0
Severity assessment (%)†			
PSI class I	11·0	33·8	12
PSI class II	13·4	20·9	20
PSI class III	19·0	14·3	21
PSI class IV	40·2	21·3	33
PSI class V	16·4	9·9	13
Outcome parameters (%)†			
ICU admission	9·0	9·2	4·0
Outpatient management	4·6	41·3	
Mortality	11·0	5·2	9·0

PSI, Pneumonia severity index; ICU, intensive care unit.

* Values are expressed as median and interquartile range (IQR).

† Because of rounding, percentages may not sum to 100.

https://doi.org/10.1017/S0950268808000435 Published online by Cambridge University Press

Table 2. *Observed and predicted mortality in patients with community-acquired pneumonia* ($n=373$)

	Observed		Original model		Recalibrated model	
Risk class	No. of patients per class (%)	Death (%) ($n=41$)	Expected no. of deaths	Expected mortality (%)	Expected no. of deaths	Expected mortality (%)
PSI						
I	41 (11·0)	0 (0)	0·04	0·1	0·5	1·1
II	50 (13·4)	1 (2·0)	0·3	0·6	1·6	3·1
III	71 (19·0)	3 (4·2)	0·6	0·9	2·8	3·9
IV	150 (40·2)	24 (16·0)	14·0	9·3	20·9	13·9
V	61 (16·4)	13 (21·3)	16·5	27·0	15·4	25·2
CURB65						
0	87 (23·3)	3 (3·4)	0·5	0·6	3·1	3·6
1	119 (31·9)	8 (6·7)	2·0	1·7	7·7	6·5
2	131 (35·1)	21 (16·0)	11·8	9·0	21·6	16·5
3	33 (8·9)	8 (24·2)	5·3	16·1	7·5	22·7
4/5	3 (0·8)	1 (33·3)	1·1	35·1	1·1	35·6
CRB65						
0	109 (29·2)	4 (3·7)	1·0	0·9	3·8	3·5
1	219 (58·7)	26 (11·9)	11·4	5·2	27·2	12·4
2	40 (10·7)	9 (22·5)	4·8	12·0	8·5	21·3
3	5 (1·3)	2 (40·0)	1·6	32·4	2·0	40·8
4	0 (0)		0·0	21·0		

PSI, Pneumonia severity index; CURB, Confusion, Urea >7 mmol/l, Respiratory rate >30/min, low Blood pressure (systolic <90 mmHg or diastolic <60 mmHg); CRB65, the same as CURB65, but does not include urea.

patients with CAP admitted to a tertiary-care centre in Switzerland. There was acceptable discriminatory performance but all scores markedly underestimated the mortality, particularly in the low-risk classes. This leads to misclassification of patients with a substantial mortality in the low-risk classes. As guidelines [3, 4] recommend outpatient management for low-risk patients, misclassification may result in inadequate discharge of patients with a considerable risk of death and potential legal consequences. Recalibration of the risk models corrected the miscalibration of predicted mortalities of all models under investigation. Of the recalibrated models, only the PSI was sensitive enough to accurately identify low-risk patients suitable for outpatient management.

Three different prediction rules, namely the PSI, CURB65 and CRB65 scores, have been proposed and extensively validated for risk stratification in CAP [3–6]. All three rules are originally designed to identify patients who are at low risk of death and who may hence qualify for outpatient management. Algorithms from statistical prediction models reflect the risk profile of patients embedded in a certain health-care setting where the original model was derived. Consequently, when transporting these rules to different settings at different times, validation and adaptation, if needed, is recommended [20, 21].

The original PSI, CURB65 and CRB65 classified 43%, 55% and 30% of the patients as low risk with a presumed mortality in the range of 1%. If the original models were well calibrated in our data, the observed mortalities in the low-risk classes would not have exceeded 1%. However, we observed mortalities of 2·5%, 5·3% and 3·6% indicating the need to recalibrate the models in our tertiary-care setting. Using a basic recalibration approach, the miscalibration of each model (Fig. 2) and the resulting misclassification of patients was corrected. Nevertheless, with mortality rates of 3·5% each in the lowest risk classes, the CURB65 and CRB65 scores were too insensitive to identify subjects with low mortality rates in the range of 1%. Consequently, the CURB65 and CRB65 may help to identify patients at high risk, but their ability to recognize low-risk patients is limited. Unlike the CURB65 and the CRB65, the recalibrated PSI score showed an adequate performance in the low-risk range and was able to correctly identify 11% of the patients with a mortality of 1%. Only class I of the recalibrated PSI score can therefore be used to identify patients who qualify for outpatient management.

https://doi.org/10.1017/S0950268808000435 Published online by Cambridge University Press

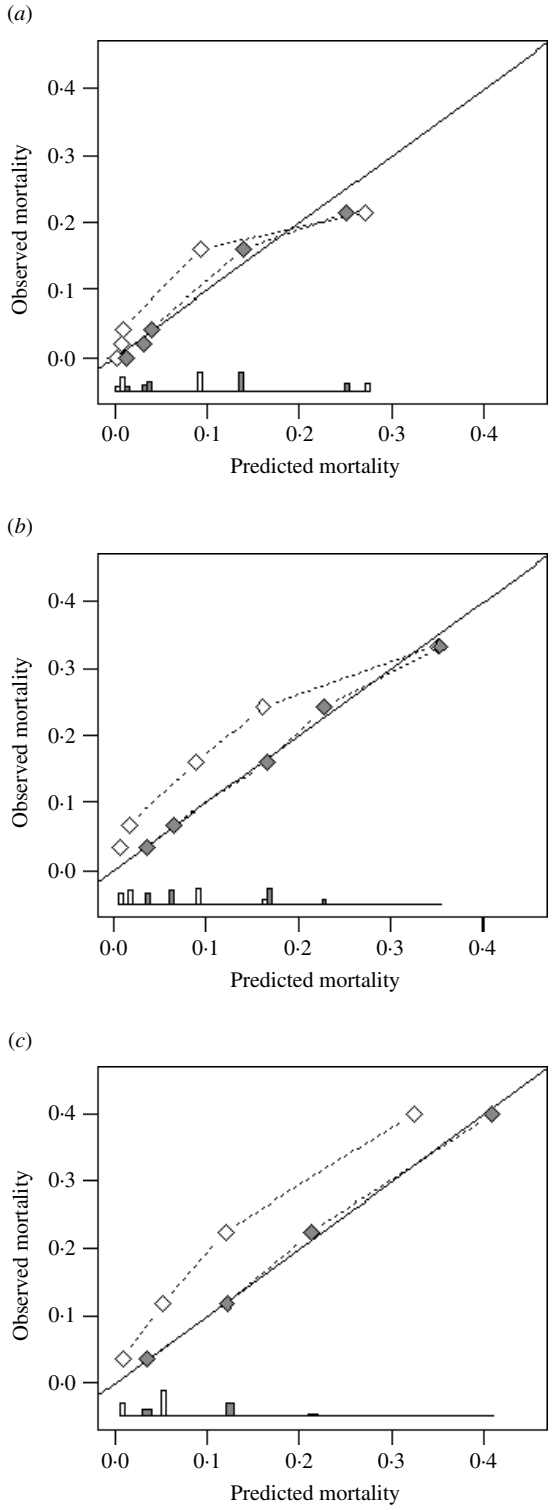

Fig. 2. Agreement between predicted and observed 30-day mortality (calibration) for three pneumonia severity prediction rules (*a*) PSI, (*b*) CURB65 and (*c*) CRB65. Observed mortality is plotted according to classes of predicted risk for each prediction rule separately. The solid line of identity represents perfect calibration of predicted risk within new patients. Correction of miscalibration (◆) after recalibration (◇).

Prior validation studies have prospectively evaluated severity scores in different clinical settings and reported high mortality rates particularly in patients of PSI class III or above and CURB65 class 1 or above [10–17]. These studies, however, focused mainly on the overall discriminatory ability of the prediction rules with varying results as expressed by differences in the area under the ROC curves. The present study extends these findings showing that an apparently adequate discriminatory model may mislead clinical decisions because of model miscalibration in a particular clinical setting. Importantly, the ROC of a prediction model is numerically not affected by miscalibration because miscalibration affects the magnitude of the predicted overall risk but not the ranking among the individual patients according to their predicted risks. Validation of both, calibration and discrimination, is thus crucial before a model which is derived at a different time and place is implemented in a clinical setting.

The present study does not *per se* question the utility of these tools, but underlines the importance of adapting these tools to local settings. The study emphasizes the importance of validation by calibration and of recalibration in the case of significant miscalibration. Although our study population differs from the original derivation population, the assessment of model calibration in the high-risk setting is from a pragmatic point of view of interest. CAP guidelines recommend the use of CAP risk scores, also in EDs, but do not specify the setting where the scores are indicated or not. In a high-risk population, as found in our study with a high proportion of referred polymorbid patients, risk scores may underestimate mortality risks, while in a low-risk setting (e.g. primary care) risk predictions may be inadequately high. Importantly, misclassified patients were found to be younger compared to correctly classified patients. As age is the strongest predictor in the risk scores, the risk of younger people may particularly be underestimated in the high-risk setting.

When evaluating a model for risk stratification, one should start using pre-existing knowledge and, if available, validate and update an existing model within the setting in question instead of building a new model from scratch with all the drawbacks of overfitting and lack of reproducibility [25]. Recalibration of existing models is attractive because of the stability which is related to the fact that only two parameters (intercept and calibration slope) are estimated [20] (see Methods section). Directly using the observed risk pertaining to

Table 3. *Accuracy of the original and the recalibrated PSI score to identify patients at low risk of death*

	PSI original (classes I–III*)	PSI recalibrated (class I*)
Number of patients assigned to low-risk class†	43% (162/373)	11% (41/373)
Mortality within low-risk class	2·5% (4/162)	0% (0/41)
Sensitivity	90·2%	100·0%
Specificity	47·3%	12·3%

PSI, Pneumonia severity index.
* Classes I–III of the original PSI score correspond to a low mortality of ⩽1%. After recalibration of the original model, only class I corresponds to a mortality of ⩽1% and therefore classifies patients as suitable for outpatient management.
† Mortality ⩽1%.

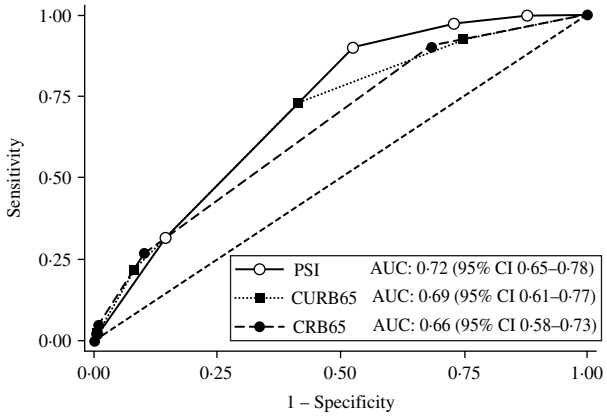

Fig. 3. Receiver-operating characteristics analysis for 30-day mortality prediction with three original pneumonia severity prediction rules (PSI, CURB65 and CRB65) in 373 patients with CAP.

a certain risk class is hampered through the potential imprecision due to the small number of observations within a certain class of predicted risk (Table 2). The recalibration approach is preferable since it is efficient and uses only two parameters which is of particular relevance in small samples as exemplified in our previous study [20].

This study included a typical spectrum of CAP patients from a university hospital in Switzerland. Our study population was different to the original study population of the PSI score in terms of age, comorbidities (e.g. renal failure, heart disease and neoplastic disease) and laboratory findings. As a study from a European tertiary-care centre, most patients were referred and selected from family physicians requesting in-patient management. Accordingly, patients had more severe pneumonia as assessed by the PSI score, the rate of outpatient management was low and mortality and rate of ICU admission was higher than in the original studies. However, we consider the population from this study as representative for the European tertiary-care setting, especially for Western Europe. In this study, the 5% of patients treated as outpatients were predominantly in the lowest risk classes of PSI, CURB65 and CRB65 scores. In comparison with 11% of low-risk patients according to the recalibrated PSI score, the management of CAP patients was reasonable after all. As outlined by guidelines, mortality prediction rules should be used to support but not replace physician decision-making about outpatient or in-patient management [3, 4]. Patients may have rare medical conditions, and patients designated as 'low risk' may have medical and psychosocial contraindications to outpatient care. Particularly, the ability to maintain oral intake, cognitive impairment, and ability to carry out activities of daily living need to be considered. Thus, determination of the initial site of care still remains an 'art of medicine' decision that, yet, may not be replaced by prediction rules [3].

Some limitations should be considered in the discussion of our results. First, the number of outcome events to perform an external validation study was rather low [26]. Second, we validated the severity scores in two trials with prospective follow-up where the issues of patient selection and representativeness of the population need to be addressed. However, the two trials consecutively included all patients with CAP, irrespective of in-patient or outpatient management. The study inclusion criteria corresponded to the criteria used in the original studies and to the criteria of CAP guidelines. In the original studies, the main reason for exclusion was non-CAP

https://doi.org/10.1017/S0950268808000435 Published online by Cambridge University Press

diagnosis and only a minority of patients was excluded because of severe contraindications such as immunosuppression or tuberculosis. It is reasonable to believe that the existing severity scores would rather underestimate the risk for rare conditions, and thus, an error because of unrepresentativeness would at least be conservative. Third, our analysis is based on predicted risks categorized in five risk classes, as issued by the original models. Preferably, we would have performed validation of the models based on a patient's individual risk using the coefficients of the original risk functions and each patient's risk profile. To the best of our knowledge the original risk functions are not published. With the full model to hand, a more differentiated picture of the performance in new data might have been possible.

In conclusion, without recalibration the original PSI, CURB65 and CRB65 scores misclassified patients with a relevant mortality as low risk in our Western European tertiary-care setting. Recalibration corrected miscalibration in each model, but only the PSI score was sensitive enough to truly identify patients at low risk. Based on this study, we advocate using the PSI prediction rule for severity assessment and consideration of outpatient management for patients with a PSI risk class I. Nevertheless, even recalibrated estimates need ongoing prospective validation and updating.

APPENDIX

Coefficients of the calibration models for the three validated risk scores

	PSI	CURB65	CRB65
Intercept $\hat{\alpha}$	−0·5167 ($P=0·13$)	−0·2321 ($P=0·61$)	0·1766 ($P=0·77$)
Calibration slope $\hat{\beta}$	0·5734 ($P<0·001$)	0·6004 ($P<0·001$)	0·7426 ($P<0·001$)

Example

Recalibration of the morality estimate from the original model in a patient with PSI class III:

$$\text{Recalibrated mortality} = 1/(1+\exp(-(\hat{\alpha}+\hat{\beta}*(\log(\text{risk}_{\text{original}}/(1-\text{risk}_{\text{original}})))))$$

$$0·039 = 1/(1+\exp(-(-0·5167+0·5734*(\log(0·9/(1-0·9)))))).$$

DECLARATION OF INTEREST

None.

ACKNOWLEDGEMENTS

We thank the staff of the clinics of Emergency Medicine, Internal Medicine and Endocrinology and the Department of Clinical Chemistry, notably Fausta Chiaverio, Martina-Barbara Bingisser, Maya Kunz, Vreni Wyss and Ursula Schild, for most helpful support during the study.

REFERENCES

1. **Macfarlane JT, *et al.*** Prospective study of aetiology and outcome of adult lower respiratory tract infections in the community. *Lancet* 1993; **341**: 511–514.
2. **Dixon RE.** Economic costs of respiratory tract infections in the United States. *American Journal of Medicine* 1985; **78**: 45–51.
3. **Niederman MS, *et al.*** Guidelines for the management of adults with community-acquired pneumonia. Diagnosis, assessment of severity, antimicrobial therapy, and prevention. *American Journal of Respiratory and Critical Care Medicine* 2001; **163**: 1730–1754.
4. **Woodhead M, *et al.*** Guidelines for the management of adult lower respiratory tract infections. *European Respiratory Journal* 2005; **26**: 1138–1180.
5. **Fine MJ, *et al.*** A prediction rule to identify low-risk patients with community-acquired pneumonia. *New England Journal of Medicine* 1997; **336**: 243–250.
6. **Lim WS, *et al.*** Study of community acquired pneumonia aetiology (SCAPA) in adults admitted to hospital: implications for management guidelines. *Thorax* 2001; **56**: 296–301.
7. **Lim WS, *et al.*** Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax* 2003; **58**: 377–382.
8. **Neill AM, *et al.*** Community acquired pneumonia: aetiology and usefulness of severity criteria on admission. *Thorax* 1996; **51**: 1010–1016.
9. **Justice AC, Covinsky KE, Berlin JA.** Assessing the generalizability of prognostic information. *Annals of Internal Medicine* 1999; **130**: 515–524.
10. **Flanders WD, *et al.*** Validation of the pneumonia severity index. Importance of study-specific recalibration. *Journal of General Internal Medicine* 1999; **14**: 333–340.
11. **Yan Man S, *et al.*** Prospective comparison of three predictive rules for assessing severity of community-acquired pneumonia in Hong Kong. *Thorax* 2007; **62**: 348–353.
12. **Capelastegui A, *et al.*** Validation of a predictive rule for the management of community-acquired pneumonia. *European Respiratory Journal* 2006; **27**: 151–157.

13. **Marras TK, Gutierrez C, Chan CK.** Applying a prediction rule to identify low-risk patients with community-acquired pneumonia. *Chest* 2000; **118**: 1339–1343.
14. **Buising KL, *et al.*** A prospective comparison of severity scores for identifying patients with severe community acquired pneumonia: reconsidering what is meant by severe pneumonia. *Thorax* 2006; **61**: 419–424.
15. **Ewig S, *et al.*** Validation of predictive rules and indices of severity for community acquired pneumonia. *Thorax* 2004; **59**: 421–427.
16. **Spindler C, Ortqvist A.** Prognostic score systems and community-acquired bacteraemic pneumococcal pneumonia. *European Respiratory Journal* 2006; **28**: 816–823.
17. **Aujesky D, *et al.*** Prospective comparison of three validated prediction rules for prognosis in community-acquired pneumonia. *American Journal of Medicine* 2005; **118**: 384–392.
18. **Christ-Crain M, *et al.*** Effect of procalcitonin-guided treatment on antibiotic use and outcome in lower respiratory tract infections: cluster-randomised, single-blinded intervention trial. *Lancet* 2004; **363**: 600–607.
19. **Christ-Crain M, *et al.*** Procalcitonin guidance of antibiotic therapy in community-acquired pneumonia: a randomized trial. *American Journal of Respiratory and Critical Care Medicine* 2006; **174**: 84–93.
20. **Steyerberg EW, *et al.*** Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine* 2004; **23**: 2567–2586.
21. **Steyerberg EW, *et al.*** Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology* 2001; **54**: 774–781.
22. **Poses RM, Cebul RD, Centor RM.** Evaluating physicians' probabilistic judgments. *Medical Decision Making* 1988; **8**: 233–240.
23. **R Development Core Team.** R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2006.
24. **Harrell FE.** Design: Design Package. R package version 2.0-12, 2005.
25. **van Houwelingen HC.** Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine* 2000; **19**: 3401–3415.
26. **Vergouwe Y, *et al.*** Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology* 2005; **58**: 475–483.