BMJ Open

Predicting mortality in patients treated differently: updating and external validation of a prediction model for nursing home residents with dementia and lower respiratory infections

Simone P Rauh,[1,2] Martijn W Heymans,[1,2] David R Mehr,[3] Robin L Kruse,[3] Patricia Lane,[4] Neil W Kowall,[5] Ladislav Volicer,[6] Jenny T van der Steen[7,8]

To cite: Rauh SP, Heymans MW, Mehr DR, *et al*. Predicting mortality in patients treated differently: updating and external validation of a prediction model for nursing home residents with dementia and lower respiratory infections. *BMJ Open* 2016;**6**:e011380. doi:10.1136/bmjopen-2016-011380

▸ Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2016-011380).

Received 3 February 2016
Revised 12 July 2016
Accepted 3 August 2016

For numbered affiliations see end of article.

Correspondence to
Simone P Rauh;
s.rauh@vumc.nl

ABSTRACT

Objective: To evaluate whether a model that was previously developed to predict 14-day mortality for nursing home residents with dementia and lower respiratory tract infection who received antibiotics could be applied to residents who were not treated with antibiotics. Specifically, in this same data set, to update the model using recalibration methods; and subsequently examine the historical, geographical, methodological and spectrum transportability through external validation of the updated model.

Design: 1 cohort study was used to develop the prediction model, and 4 cohort studies from 2 countries were used for the external validation of the model.

Setting: Nursing homes in the Netherlands and the USA.

Participants: 157 untreated residents were included in the development of the model; 239 untreated residents were included in the external validation cohorts.

Outcome: Model performance was evaluated by assessing discrimination: area under the receiver operating characteristic curves; and calibration: Hosmer and Lemeshow goodness-of-fit statistics and calibration graphs. Further, reclassification tables allowed for a comparison of patient classifications between models.

Results: The original prediction model applied to the untreated residents, who were sicker, showed excellent discrimination but poor calibration, underestimating mortality. Adjusting the intercept improved calibration. Recalibrating the slope did not substantially improve the performance of the model. Applying the updated model to the other 4 data sets resulted in acceptable discrimination. Calibration was inadequate only in one data set that differed substantially from the other data sets in case-mix. Adjusting the intercept for this population again improved calibration.

Conclusions: The discriminative performance of the model seems robust for differences between settings. To improve calibration, we recommend adjusting the intercept when applying the model in settings where different mortality rates are expected. An impact study may evaluate the usefulness of the two prediction models for treated and untreated residents and whether it supports decision-making in clinical practice.

Strengths and limitations of this study

- An existing prediction model for treated residents was updated and validated for untreated residents, thus combining prior knowledge on predictors of 14-day mortality for nursing home residents with dementia and lower respiratory tract infections treated with antibiotics with new knowledge on residents not treated with antibiotics.
- The generalisability of the updated model for the untreated residents was evaluated by externally validating the model in four data sets from two countries, allowing for an evaluation of the historical, geographical, methodological and spectrum validity of the model.
- Through adjustment of the intercept only, the model was generalisable to different settings which supports the principle of a stepwise approach using cumulative data to improve prognostic modelling.
- The relatively small sample sizes of the data sets did not allow for re-estimating the predictors or for extending the original prediction model by adding new predictors; however, recalibration methods may suffice to improve model performance.

INTRODUCTION

Lower respiratory tract infections (LRI), including pneumonia, are an important cause of death in nursing home residents and may be the ultimate cause of death in one-third to two-thirds of patients with dementia.[1–4] Physicians and families involved in the care of patients with dementia often must decide whether or not to treat LRI, and patients' wishes are often unclear.[5] A variety of factors relate to the decision to withhold antibiotic treatment, and these may differ by country; for example, in the USA, sicker patients are more likely to be treated, whereas the reverse is true in the Netherlands.[6]

To inform prognosis and support physicians in decision-making, a prediction model was previously developed in a Dutch population of nursing home residents with dementia and LRI (more specifically, a physician's diagnosis of pneumonia, mostly without X-ray),[3] predicting mortality within 14 days for residents treated with antibiotics. The model showed good discrimination and adequate calibration in the Dutch development population.[3] This model was validated externally in a US study population of residents with LRI in which discrimination was good and calibration was adequate.[3] However, a clinical impact study indicated that physicians perceived usefulness in practice as suboptimal because the prediction was valid only for residents treated with antibiotics.[7] The current study therefore aims to define a model that can adequately predict mortality in nursing home residents with dementia and LRI who are not treated with antibiotics.

Predicting a particular outcome may start with existing prediction models rather than developing new prediction models for each new data set, because in this way, prior knowledge is combined with new knowledge.[8 9] Furthermore, for clinical practice, it is impractical to have a variety of competing models to choose from. For instance, more than 60 models have been developed that predict breast cancer prognosis[10] and more than 100 models that predict outcome after brain trauma.[11] We previously found that, despite profound differences between residents treated with and residents treated without antibiotics, predictors for mortality within 1 week were largely similar between those two groups.[12] Therefore, rather than developing yet another prediction model, our aim was to evaluate whether the model that was previously developed to predict 14-day mortality for nursing home residents with dementia and LRI who received antibiotics[3] could be applied to residents who were not treated with antibiotics. In this way, spectrum transportability of the original model was studied. In addition, our aims were to examine whether the model performance in the untreated residents improves by applying available updating strategies,[9 13–15] and to externally validate this updated model in four other data sets, thus studying the historical, geographical, methodological and spectrum transportability of the updated model.

METHODS

Description of the development data set

The Dutch Pneumonia Study prospectively included 706 nursing home residents in 61 nursing homes between October 1996 and July 1998.[12 16] The inclusion criteria were (1) psychogeriatric disease (almost all dementia); (2) nursing home residence for ≥4 weeks; and (3) pneumonia diagnosis as judged by a physician (mostly without X-ray).

For this study, we only included residents who were not treated with antibiotics (n=165). We excluded residents who did not have a diagnosis of dementia (n=6), with unknown mortality status (n=1), and one resident who had missing values on 4 of the 8 predictors, resulting in 157 eligible cases for analyses.

Description of the external validation data sets

To externally validate the prediction model, four data sets were used. The Missouri Lower Respiratory Infection (LRI) Study prospectively included 1406 episodes of LRI in 1044 nursing home residents in 36 nursing homes between August 1995 and September 1998.[17 18] We selected the episodes of residents who were not treated with antibiotics (n=254). We excluded episodes without a dementia diagnosis (n=78). The Missouri LRI Study defined dementia as either a diagnosis of dementia or a score of ≥3 on the MDS Cognitive Performance Scale.[3 19] Eligible cases (176) represented 162 residents.

The 'Dutch 2006–2007' study prospectively included 72 nursing home residents in 54 nursing homes between July 2006 and August 2007.[20 21] Of these 54 nursing homes, 53 previously participated in the Dutch Pneumonia Study. We selected the residents not treated with antibiotics (n=15). We excluded residents without a dementia diagnosis (n=1) or with unknown mortality status (n=1). There were 13 cases eligible for analyses.

The 'Bedford US' study prospectively included 110 episodes of LRI in 94 nursing home residents in the dementia special care unit of a US Department of Veterans Affairs nursing home between February 2004 and November 2008.[22] All residents were diagnosed with dementia. We selected the episodes untreated with antibiotics (n=25). There were 25 eligible cases in 25 residents for analyses.

Of the Dutch End of Life in Dementia (DEOLD) study, we included 155 episodes of 110 nursing home residents identified prospectively or retrospectively after death in 34 nursing homes between January 2007 and March 2010.[23] All residents were diagnosed with dementia. We selected the episodes untreated with antibiotics (n=25). There were 25 eligible cases in 24 residents for analyses.

All studies have been approved by the local medical ethics committees and, when this was deemed necessary by the medical ethics review committee, family or proxies provided informed consent.

Original prediction model

The original prediction model was developed in the Dutch Pneumonia Study in nursing home residents with dementia and LRI who were treated with antibiotics,[3] to

Table 1 Original prediction model for treated residents[3] (logistic regression model, after internal validation)

Predictor	Regression coefficient
Intercept	−6.263
Male gender	0.447
Respiratory rate (per unit)	0.027
Respiratory difficulty (y/n)	0.667
Pulse rate (per unit)	0.019
Decreased alertness (y/n)	0.692
Insufficient fluid intake (y/n)	0.561
Eating dependency*	0.771
Pressure sore (y/n)	0.557

*Per point more dependent on a 3-point scale (0, independent; 1, need for assistance; 2, fully dependent).

predict 14-day mortality (table 1). The model included: gender (male/female), respiratory rate (breaths per minute), respiratory difficulty (yes/no), pulse rate (beats per minute), decreased alertness (yes/no), insufficient fluid intake (yes/no), eating dependency (independent/need for assistance/fully dependent) and pressure sore (any grade; yes/no). Respiratory rates and pulse rates were truncated at 12 and 60 breaths per minute and at 50 and 160 beats per minute, respectively, to avoid undue influence of outliers (the latter value was not exceeded in the data used for this study).

This prediction model showed excellent discrimination (area under the receiver operating characteristic curve (AUC)=0.80) and adequate calibration (Hosmer and Lemeshow (H&L) goodness-of-fit statistic: p=0.23) in the Dutch development population.[3] After internal validation of the model using bootstrapping techniques, external validation in a US population showed acceptable discrimination (AUC=0.74) and adequate calibration (H&L statistic: p=0.67).[3]

Developing the model for antibiotic-untreated residents

Step 1: Validation of the original prediction model in untreated residents

First, we tested whether the original prediction model for residents treated with antibiotics[3] was also valid for residents of the Dutch Pneumonia Study who were not treated with antibiotics. Prior research in this study population has shown that the residents not treated with antibiotics are more severely ill and have higher mortality than the residents who were treated with antibiotics.[12 16] Therefore, we anticipated a need to update the model that was developed and internally validated for the residents treated with antibiotics to improve model performance in the residents not treated with antibiotics.

Step 2: Updating of the original prediction model

After applying the original prediction model to the untreated residents of the Dutch Pneumonia Study, the same data set was used to update the model according to the new case-mix of untreated residents, thus improving the performance of the model for the untreated residents. Steyerberg[14] describes eight different methods to update a prediction model, including methods for recalibration, model revision and model extension.

In our study, we employed methods for recalibration. First (update 1), the intercept of the original prediction model was recalibrated: the regression coefficients of the original prediction model were applied to the untreated residents of the Dutch Pneumonia Study, fixed at their original values, while the intercept was the only free parameter. Thus, we corrected for differences in mortality rate between treated and untreated residents.

Second (update 2), the intercept and the overall calibration slope were recalibrated: the linear predictor (LP) was calculated for each patient by multiplying the values of the regression coefficients of the original prediction model by the values of the corresponding predictor variables for each individual patient. Next, a prediction model was constructed with only the LP as a predictor, estimating two parameters: an intercept and a regression coefficient for the LP (ie, the calibration slope or the shrinkage factor). This recalibrates the original regression coefficients.

We refrained from additional steps, such as re-estimating the predictors and extending the model with new predictors, to avoid problematic overfitting of coefficients due to our relatively small sample (n=157 untreated cases in the development data set).

Step 3: External validation of the updated model

Finally, the updated model was externally validated in four data sets: the Missouri LRI Study, the Dutch 2006–2007 study, the Bedford US study and the DEOLD study. Owing to considerable differences in resident characteristics and in the LRI definition between the Missouri LRI Study (LRI assessed by project nurses using clinical criteria) and the other three studies (physician's diagnosis of pneumonia), the model was validated separately in the Missouri LRI Study and in the other three data sets combined.

Validation studies can address several types of transportability: historical/temporal, geographical, methodological and spectrum/domain transportability.[9 14 24] Residents in the Missouri LRI Study were historically comparable to residents in the development data set (1995–1998 vs 1996–1998), but these data sets differed in geographical (USA vs the Netherlands) and methodological aspects (different diagnosis of LRI) and in case-mix (table 2). This enabled a study of the geographical, methodological and spectrum transportability of the updated model with data from the Missouri LRI Study.

Residents in the Dutch 2006–2007 study, the Bedford US study and the DEOLD study (hereafter referred to as the three combined external validation data sets) differed from residents in the development data set in historical

Table 2 Description of the development and external validation patient data sets

Resident characteristic	Data set Dutch Pneumonia Study (development data set)	Missouri LRI Study	Three combined external validation data sets*
Number of untreated patients	157	176	63
14-day mortality, number (%)	138 (87.9)	24 (13.6)†‡	51 (81.0)
Age, mean (SD); range	82.6 (7.8); 59–98	85.2 (8.1); 60–104†‡	82.9 (6.0); 67–99
Dementia severity/cognitive performance, mean score (SD)	BANS-S 20.5 (3.9)	CPS 4.5 (1.3)	–
Gender, % female	61.8%	81.3%†‡	33.3%†
Respiratory rate, mean (SD); range§	29.8 (9.3); 12–60	26.5 (6.7); 12–44 †‡	29.3 (8.4); 12–56
Respiratory difficulty, %	66.2%	19.3%†‡	77.8%
Pulse rate, mean (SD); range¶	97.6 (17.8); 50–144	85.3 (14.9); 52–140†‡	92.6 (17.9); 50–148
Decreased alertness, %	75.2%	27.3%†‡	79.4%
Insufficient fluid intake, %	78.3%	8.0%†‡	74.6%
Eating dependency, %			
Independent	0.6%	10.8%†‡	4.8%†
Need for assistance	5.7%	50.0%	22.2%
Fully dependent	93.6%	39.2%	73.0%
Pressure sore, %	28.0%	11.4%†	20.6%

*Three combined external validation data sets: the Dutch 2006–2007 study, the Bedford US study and the DEOLD study combined.
†Significantly different (p<0.05) compared with the Dutch Pneumonia Study; tested with t-tests for continuous, χ^2 for dichotomous and χ^2 including test for trend for categorical variables.
‡Significantly different (p<0.05) compared with the three combined external validation data sets.
§Truncated at 12 and 60 breaths per minute.
¶Truncated at 50 beats per minute.
BANS-S, Bedford Alzheimer Nursing Severity-Scale;[31] CPS, Cognitive Performance Scale.[19]

aspects (2006–2007, 2004–2008, 2007–2010 vs 1996–1998) but were comparable to residents in the development data set in methodological and spectrum aspects. Geographically, residents of the Dutch 2006–2007 study were very similar to residents of the development data set because 53 of the 54 participating nursing homes in the Dutch 2006–2007 study participated in the Dutch Pneumonia Study (the development data set). The DEOLD study was mostly conducted in different Dutch nursing homes than in the other two Dutch studies, whereas the Bedford US study was conducted in the USA. Therefore, the three combined external validation data sets enable a study of historical and, to some extent, geographical transportability of the updated model.

Statistical analyses

Comparing samples

Differences in resident characteristics between the development data set, the Missouri LRI Study and the three combined external validation data sets were tested using χ^2 tests for dichotomous variables, χ^2 tests for trend for ordinal categorical variables and independent-samples t-tests for continuous variables.

Model performance

To evaluate the performance of the original prediction model applied to the untreated residents of the Dutch Pneumonia Study (step 1), the performance of the two updated models (ie, recalibration of intercept only, and slope combined with the intercept) in the Dutch Pneumonia Study (step 2) and the performance of the updated model in the external validation data sets (step 3), discrimination and calibration were assessed. For discrimination, AUCs were assessed, considering AUCs of 0.70–0.79 to indicate acceptable, 0.80–0.89 to indicate excellent and >0.90 to indicate outstanding discrimination.[25] Calibration was assessed by H&L statistics (non-significant values indicate adequate calibration) and by calibration graphs, comparing observed and predicted mortality rates in deciles of the predicted mortality risk. Further, the added value of the new intercept (update 1) or calibration slope (update 2) was tested for significance using the Wald-statistic, and difference in overall performance between two models was assessed using a likelihood-ratio test (LR-test).

In addition, reclassification tables were produced to compare classifications of patients as low or high risk between the old and new models, that is, whether patients were better classified using the new models.[26] Next, the net reclassification index (NRI) was calculated,[26] using a cut-off point of 80% to define low versus high mortality risk. This cut-off point was chosen because a survey study showed that 73% of the clinicians considered mortality risks of 75–90% as sufficiently high to justify withholding antibiotics.[20] NRI for events was calculated as the proportion of events classified up (ie, from <80% to ≥80%) minus the proportion of events classified down, and can be interpreted as the improvement in sensitivity between the two different models.[26 27] NRI for non-events was calculated as the proportion of non-events classified down minus the proportion of non-events classified up, and can be interpreted as the improvement in specificity between the two different models.[26 27] Total NRI was the sum of NRI for events and NRI for non-events and could theoretically range from −2 (if sensitivity and specificity would deteriorate from 100% to 0%) to +2 (if sensitivity and specificity would improve from 0% to 100%). Finally, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated using the same cut-off of 80% to compare classifications of patients as low or high risk between the original and the updated models.

Missing predictor values

Consistent with the imputation strategy that was used previously in the Dutch Pneumonia Study and the Missouri LRI Study,[3] we used the same imputation strategy in this study: for the continuous variables, respiratory rate (17.8% missing) and pulse rate (19.1% missing), hot-deck imputation was applied, using the variables tachypnoea and tachycardia, respectively. For dichotomous and categorical variables (<2.5% missing per variable), modes were imputed.

Sensitivity analysis

In the Missouri LRI Study and in the DEOLD study, multiple episodes could be included per resident. To evaluate whether the inclusion of multiple episodes per resident could have affected our results, we assessed the performance of the updated model in subsets of the external validation data sets including only the first episode for each resident (14 and 1 episode were removed in the Missouri LRI Study and in the DEOLD study, respectively).

Significance and software

A two-sided p value of <0.05 was considered statistically significant. Statistical analyses were performed using SPSS V.19 and R software V.2.15.0, using the packages ‘rms’, ‘pROC’ and ‘ResourceSelection’.

RESULTS

Resident characteristics

Several resident characteristics differed between the data sets (table 2 and online supplementary table SA). In the development data set, 14-day mortality was 87.9%. Among the external validation data sets, 14-day mortality was lowest in the Missouri LRI Study (13.6%) and highest in the Dutch 2006–2007 study (92.3%). Residents in the Missouri LRI Study were significantly older, more often female and less severely ill compared to residents of either the development data set or the three combined external validation data sets. The percentage of pressure sores was lower in the Missouri LRI Study. Residents in the three combined external validation data sets were significantly less often fully dependent in eating and more often male (with the Bedford US study including mostly male residents) than residents in the development data set.

Step 1: validation of the original prediction model

The original prediction model applied to the untreated residents of the Dutch Pneumonia Study showed excellent predictive performance (AUC=0.80, 95% CI 0.79 to 0.82; table 3). However, calibration was poor (H&L statistic: p<0.001) and the calibration plot showed that predicted values were systematically too low (figure 1A). Based on a cut-off point of 80%, specificity was 100%, but sensitivity was only 1%.

Step 2: updating of the original prediction model

Consistent with the underestimation of mortality in the previous step, we found that updating the intercept (update 1) resulted in an increase in the intercept of 2.66 (table 3), and adding this new intercept to the model led to a significant Wald-statistic (p<0.001, not shown). The calibration plot improved considerably (figure 1B) and the H&L statistic showed adequate fit (p=0.38). Since the ranking of the predicted probabilities is not affected by the updated intercept, discrimination did not change. Reclassification of the first updated model compared with the original prediction model (cut-off 80%) showed that 84% of the events were reclassified from low risk (<80%) to high risk (≥80%), and thus were reclassified correctly, but also 53% of the non-events were reclassified from low risk to high risk, and thus were reclassified incorrectly (tables 4 and 5), resulting in an NRI of 0.31 (p=0.01). Sensitivity increased to 85%, although specificity decreased to 47%. This led to a decrease in PPV from 100% to 92% and an increase in NPV from 12% to 30%.

The model with the updated intercept and calibration slope (update 2) resulted in a non-significant Wald-statistic for the new calibration slope (p=0.30, not shown) and no significant improvement in overall model performance (LR-test p=0.28). Further, the H&L statistic showed adequate fit (p=0.35), but the calibration plot deteriorated compared with the model including only an updated intercept (figure 1C). Again, the ranking of the predicted probabilities, and thus discrimination, did not change. Reclassification of the model including a recalibrated intercept and slope compared with the model including a recalibrated intercept showed only minor improvement: 5% of the non-events were reclassified from high risk to low risk; specifically, one non-event was reclassified correctly (table 6). Finally, there was no change in sensitivity and only minor improvement in specificity from 47% to 53%. Therefore, we selected the model with the updated intercept only (update 1) for external validation (see online supplementary table SB).

Step 3: external validation of the updated model

Applying the updated model to the Missouri LRI Study showed that discrimination was acceptable (AUC=0.76,

Table 3 Model performance in the Dutch Pneumonia Study (development data set)

		Step 2	
	Step 1	**Update 1**	**Update 2**
Discrimination			
AUC (95% CI)	0.80 (0.79 to 0.82)	0.80 (0.79 to 0.82)	0.80 (0.79 to 0.82)
Sensitivity	1%	85%	85%
Specificity	100%	47%	53%
PPV	100%	92%	93%
NPV	12%	30%	32%
Calibration (H&L statistic*)			
χ^2	178.6	8.6	8.9
p Value	<0.001	0.38	0.35
Overall performance (LR-test)			
χ^2	–	–	1.18†
p Value	–	–	0.28
Reclassification			
NRI‡	–	0.31§	0.05†
p Value	–	0.09	0.32
Parameters			
Model intercept¶	−6.263	−3.607	−5.359
Calibration intercept**	–	2.656	0.904
Calibration slope††	–	–	1.33

Step 1: Original prediction model applied to untreated residents.
Step 2: Updating the original model.
Update 1: recalibrating intercept.
Update 2: recalibrating intercept and slope.
*H&L statistic: non-significant p values indicate adequate fit.
†Compared with update 1.
‡Cut-off point 80%.
§Compared with step 1.
¶Intercept of the new model, after updating.
**Deviation from original intercept (−6.263).
††To recalibrate the original regression coefficients, all original regression coefficients are multiplied by the calibration slope or shrinkage factor.
AUC, area under the curve; H&L statistic, Hosmer and Lemeshow goodness-of-fit statistic; LR-test, likelihood ratio test; NPV, negative predictive value; NRI, net reclassification index; PPV, positive predictive value; χ^2: Chi-square.

95% CI 0.74 to 0.78), but the predicted probabilities were systematically too high (figure 1D), and the fit was inadequate (H&L statistic: p<0.001; table 7). We therefore examined whether updating the intercept for this data set would improve model performance. Consistent with the overestimation of mortality in this data set, we found that updating the intercept resulted in a significant decrease of 2.407 (p<0.001; data not shown), leading to a final intercept of −6.014 (online supplementary table SB). Applying this new model to the Missouri LRI Study resulted in an improvement in calibration, statistically (H&L statistic: p=0.16, not shown) and visually (figure 1E).

In the three combined external validation data sets (table 5), discrimination was excellent (AUC=0.83, 95% CI 0.79 to 0.86). The H&L statistic indicated that calibration was adequate (p=0.09), although the calibration plot (figure 1F) showed that predicted values were overestimated for residents with lower observed probabilities. Again, we examined whether updating the intercept for this data set would improve model performance. Updating the intercept resulted in a small non-significant decrease in the intercept of 0.38 (p=0.28; data not shown). Although this update improved calibration statistically (H&L statistic: p=0.25; data not shown), the calibration plot showed only a small improvement (figure 1G), and reclassification of this model compared with the previous model showed that 18% of the events were reclassified from high risk to low risk (data not shown), and thus were reclassified incorrectly, while 8% of the non-events were reclassified correctly, resulting in an NRI of −0.09 (p=0.36). Since the calibration plot showed overestimation of the predicted values for residents in the three combined external validation data sets with lower observed probabilities (figure 1F, G), we additionally examined whether updating the calibration slope would improve the model for these data sets. However, this resulted in a non-significant Wald-statistic for the new calibration slope (p=0.26; data not shown) and no significant improvement in overall model performance (LR-test p=0.22).

Sensitivity analysis

Applying the updated model to a subset of the Missouri LRI Study including only the 162 first episodes did not lead to substantial differences in the performance of the model: AUC=0.75 (95% CI 0.73 to 0.77), H&L statistic:

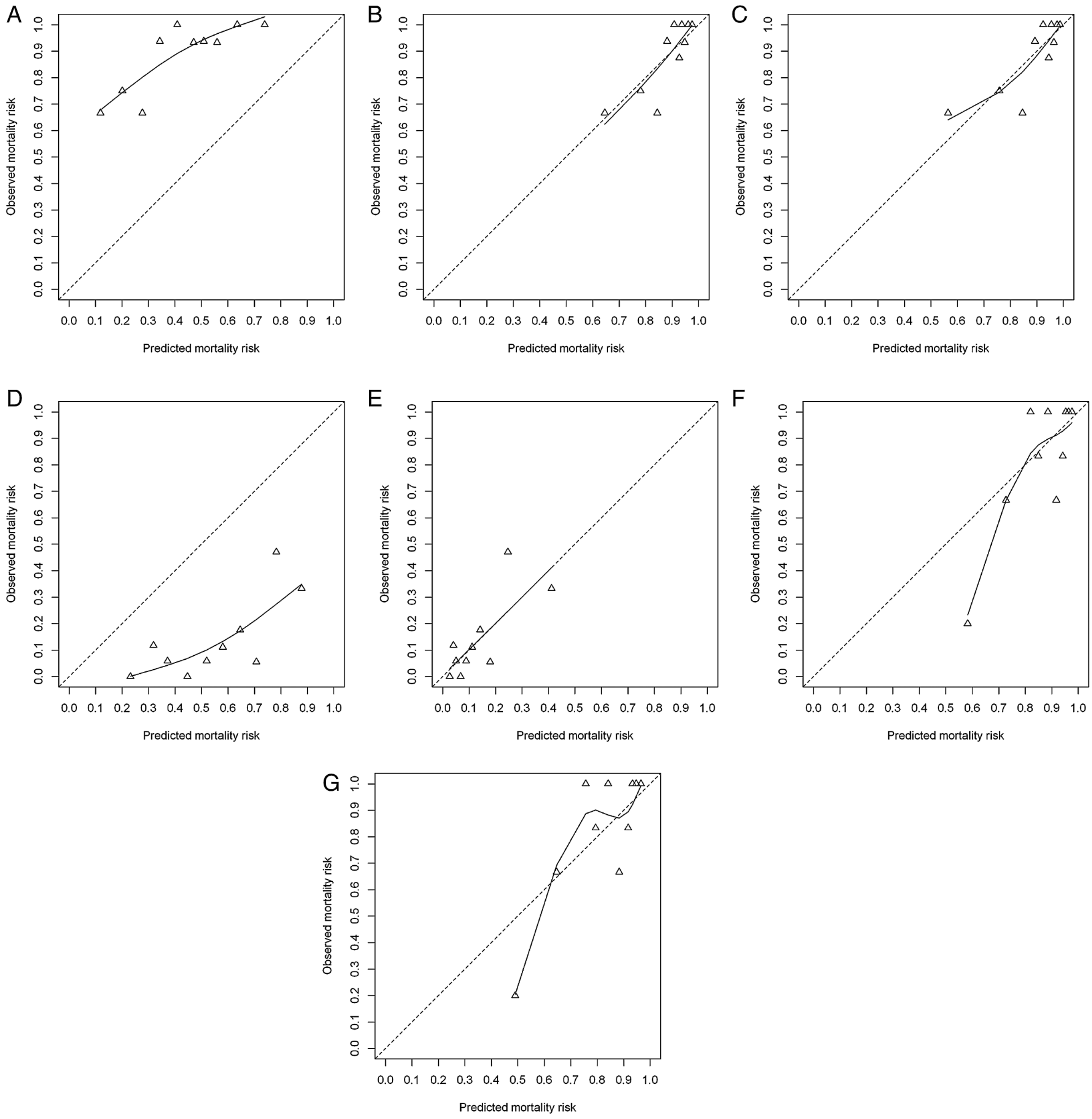

Figure 1 Calibration plots. The dotted line indicates perfect calibration. The triangles represent the observed and predicted mortality rates in deciles of the predicted mortality risk. The solid line is a smoothed spline curve. (A) Calibration plot of the original prediction model applied to untreated residents of the Dutch Pneumonia Study (step 1). (B) Calibration plot of update 1: model with recalibrated intercept applied to untreated residents of the Dutch Pneumonia Study (step 2). (C) Calibration plot of update 2: model with recalibrated intercept and calibration slope applied to untreated residents of the Dutch Pneumonia Study (step 2). (D) Calibration plot of external validation: model with recalibrated intercept applied to untreated residents of the Missouri LRI Study (step 3). (E) Calibration plot of additional update: model with additional recalibration of the intercept for the untreated residents of the Missouri LRI Study. (F) Calibration plot of external validation: model with recalibrated intercept applied to untreated residents of the three combined external validation data sets (step 3). (G) Calibration plot of additional update: model with additional recalibration of the intercept for the untreated residents of the three combined external validation data sets.

p<0.001. Similarly, for the three combined external validation data sets, applying the updated model to a subset including only the 62 first episodes did not lead to substantial differences in the performance of the model: AUC=0.83 (95% CI 0.79 to 0.86), H&L statistic: p=0.052.

DISCUSSION

Our aim was to evaluate whether a previously developed model that predicts short-term mortality in nursing home residents with dementia and LRI treated with antibiotics[3] could be applied to residents who were not treated with antibiotics. We found that in the

Table 4 Reclassification index (number of participants)

		Update 1 compared with original model	Update 2 compared with update 1
NRI (p value)		0.31 (p=0.01)	0.05 (p=0.32)
Events (n=138)	Classified up	0.84 (116)	0.00 (0)
	Classified down	0.00 (0)	0.00 (0)
Non-events (n=19)	Classified up	0.53 (10)	0.00 (0)
	Classified down	0.00 (0)	0.05 (1)

NRI, net reclassification index.
Reclassification based on a cut-off point of 80%.
Original model: original prediction model applied to untreated residents of the Dutch Pneumonia Study.
Update 1: recalibration intercept.
Update 2: recalibration intercept and slope.

Table 5 Reclassification table: update 1 compared with original model

	Update 1		
Original model	<80%	≥80%	Total
<80% risk			
n with event	21	116	137
n without event	9	10	19
≥80% risk			
n with event	0	1	1
n without event	0	0	0
Total	30	127	157

Original model: original prediction model applied to untreated residents.
Update 1: recalibration intercept.

Table 6 Reclassification table: update 2 compared with update 1

	Update 2		
Update 1	<80%	≥80%	Total
<80% risk			
n with event	21	0	21
n without event	9	0	9
≥80% risk			
n with event	0	117	117
n without event	1	9	10
Total	31	126	157

Update 1: recalibration intercept.
Update 2: recalibration intercept and slope.

development cohort, the Dutch Pneumonia Study, discrimination was excellent in the untreated residents, but calibration was poor and mortality was underestimated. This may relate to the fact that the untreated residents were, compared with the treated residents, more severely ill on average and also had a higher mortality rate. Updating the model using recalibration methods improved calibration in the untreated residents. This indicates that the discriminative performance of the original prediction model may be robust for differences in spectrum aspects, while calibration may be inadequate in new settings and require correction for differences in mortality rates.

In line with this, external validation of the updated model in the Dutch 2006–2007 study, the Bedford US study and the DEOLD study combined—populations with a case-mix and a mortality rate similar to the Dutch Pneumonia Study—showed an excellent discrimination and adequate calibration. However, in the Missouri LRI Study, where the untreated residents were less severely ill and the mortality rate was much lower compared with the other data sets, the updated model showed acceptable discrimination but inadequate calibration, systematically overestimating mortality. Nevertheless, an additional update of the intercept in this population improved calibration considerably. We conclude that while the discriminative performance of the updated model is robust for differences in historical, geographical, methodological and spectrum aspects between settings, calibration can be inadequate in new settings with different illness severity and mortality rates. We therefore recommend using the updated model for untreated residents in populations with an illness severity and a mortality rate comparable to the untreated residents of the Dutch Pneumonia Study, but to update the intercept of the model when applying the model in a setting with a different illness severity or where a different mortality rate is expected. For example, in populations with relatively low illness severity and low mortality such as the Missouri LRI Study, an intercept of −6.0 is required (instead of −3.6). Online supplementary table SB provides examples of how to apply the prediction model in clinical practice.

In the three combined external validation data sets, an additional update of the intercept was not deemed necessary, although the calibration plot did show overestimated mortality risks for residents with lower observed risks. Since the data in this area were sparse (11 residents), we cannot determine whether this reflects random error or actual overestimated predictions for residents with lower observed risks. Although updating the intercept resulted in a small improvement in calibration both statistically (assessed with the H&L statistic) and visually (assessed with calibration graphs), it worsened classification as low versus high risk. Physicians may find the classification of high-risk or low-risk patients important for clinical practice,[7] which implies that correct classification might be more important for clinical practice than correct calibration, and therefore this additional update was not an improvement in this population.

A limitation of this study is that the sample sizes of the data sets were relatively small, which did not allow for re-estimating predictors or extension of the original prediction model with new predictors. However, recalibration methods may suffice to improve model performance, especially when discrimination is already adequate in a new setting.[8 14 28] In our study, we indeed

Table 7 Performance of updated model (update 1; update of intercept) in all data sets

	Dutch Pneumonia Study (n=157)	Missouri LRI Study (n=176)	Three combined external validation data sets (n=63)
Discrimination			
AUC (95% CI)	0.80 (0.79 to 0.82)	0.76 (0.74 to 0.78)	0.83 (0.79 to 0.86)
Calibration			
χ^2	8.6	174.0	13.6
p Value	0.38	<0.001	0.09
Type of transportability*			
Historical/temporal	(Reference)	–	+
Geographical	(Reference)	+	+/–
Methodological	(Reference)	+	–
Spectrum/domain	(Reference)	+	–
Mortality rate	88%	14%	81%

Three combined external validation data sets: Dutch 2006–2007 study, the Bedford US study and the DEOLD study combined.
*Compared with data set 1: Dutch Pneumonia Study.
+, aspect differs from data set 1: Dutch Pneumonia Study.
–, aspect is similar to data set 1: Dutch Pneumonia Study.
AUC, area under the curve; χ^2: Chi-square.

found that an update of the intercept led to a considerable improvement in calibration. Further, two of the data sets included multiple episodes per resident. However, sensitivity analysis showed that excluding those multiple episodes did not considerably change the performance of the model. Finally, we validated our model in US and Dutch populations. Another validation study might be needed before using the model in other countries.

A limitation of using reclassification tables and measures of sensitivity, specificity, NPV and PPV was that a cut-off point is needed to define high versus low risk. We chose a cut-off point of 80% based on research among Dutch elderly care physicians,[20] but this cut-off may be highly culturally sensitive. Another cut-off point would lead to different values of these measures.

A strength of this study is that an existing prediction model for treated residents was successfully validated and updated for untreated residents instead of creating another new prediction model. In this way, prior knowledge on predictors of 14-day mortality for nursing home residents with dementia and LRI treated with antibiotics was combined with new knowledge on residents not treated with antibiotics. Several studies have emphasised the importance of validating existing models and have expressed concerns about the fact that validation studies are still rare.[8 9 29 30] Moreover, after updating the original model for the untreated residents, we externally validated it, which has also been recommended in the literature.[8 9 30]

The original prediction model was also converted to a scoring system that can be used in clinical practice. Our research indicates that the same scoring system could be used for the untreated residents by adapting the intercept only. In the data sets included in our study, the decision to either treat a resident or to withhold treatment was not based on randomised study designs. In contrast, it has been shown that physicians might base this decision on illness severity, and therefore, the lower mortality rates for treated compared with untreated residents are not necessarily the result of their treatment.[16] When the two scoring systems are used simultaneously in clinical practice, it is important to caution against incorrect interpretation as to the effect of antibiotics in the same (individual) residents.

In conclusion, we demonstrated that the prediction model that was developed to predict 14-day mortality for nursing home residents with dementia and antibiotic-treated LRI[3] showed excellent discrimination in nursing home residents with dementia and LRI who were not treated with antibiotics, and only required adjustment of the intercept to improve calibration. Further, the intercept had to be adjusted again when the model for untreated residents was applied to an untreated population that differed substantially in overall mortality. We therefore conclude that the discriminative performance of the prediction model seems robust for differences between settings, and, to improve calibration, recommend adjusting the intercept when applying the model in a setting where a different mortality rate is expected. A clinical impact study may evaluate the usefulness of the two prediction models—the original developed model for the treated residents and the updated model for the untreated residents—in clinical practice.

Author affiliations
[1]Department of Epidemiology and Biostatistics, VU University Medical Centre, Amsterdam, The Netherlands
[2]EMGO Institute for Health and Care Research, VU University Medical Centre, Amsterdam, The Netherlands
[3]Department of Family and Community Medicine, School of Medicine, University of Missouri, Columbia, Missouri, USA
[4]E.N. Rogers Memorial Veterans Hospital, Geriatric Research Education Clinical Center, Bedford, Massachusetts, USA
[5]VA Boston Healthcare System, Department of Veterans Affairs and Boston University Alzheimer Disease Center at BU School of Medicine, Boston, Massachusetts, USA
[6]School of Aging Studies, University of South Florida, Tampa, Florida, USA
[7]Leiden University Medical Center, Department of Public Health and Primary Care, Leiden, The Netherlands

[8]Radboud University Medical Center, Department of Primary and Community Care, Nijmegen, The Netherlands

Contributors SPR contributed to conception and design, data analysis and interpretation, drafting the article. MWH contributed to conception and design, data analysis and interpretation, revising the article critically for important content. DRM, RLK, PL, NWK and LV contributed to data collection, revising the article critically for important content. JTvdS contributed to conception and design, data collection, data analysis and interpretation, revising the article critically for important content. All authors approved the final version of the manuscript.

Funding This work was supported by the Netherlands Organisation for Scientific Research (NWO), career award for JTvdS under the Innovational Research Incentives Scheme, Vidi grant number 917.11.339. The Dutch Pneumonia Study was originally funded by the Dutch Ministry of Healthcare, Welfare and Sport (grant numbers. 37851, 37852, 37853, 37832) and the Society 'Het Zonnehuis'. The Missouri LRI Study was originally funded by the Agency for Healthcare Quality and Research (grant HS08551) and a Generalist Physician Faculty Scholars Award to DRM from the Robert Wood Johnson Foundation. The Dutch 2006–2007 study was originally funded by 'Stichting Wetenschaps Bevordering Verpleeghuiszorg' (Society Promoting Scientific Research in Nursing Home Care) (SWBV). The Bedford US Study was originally funded by the Department of Veterans Affairs and NIH Center (grant P30 AG13846). The DEOLD study was originally supported by a career award for JTvdS of the Netherlands Organisation for Scientific Research (NWO; Veni 916.66.073), by ZonMw the Netherlands Organisation for Health Research and Development (grant number 1151.0001) and by the VU University Medical Center, EMGO Institute for Health and Care Research, Department of General Practice & Elderly Care Medicine and Department of Public and Occupational Health, Amsterdam, The Netherlands. The funders had no role in the study design, data collection, data synthesis, data interpretation or writing of the report.

Competing interests None declared.

Ethics approval This paper includes data from five previously conducted studies (no data were collected for the purpose of this study). The Dutch Pneumonia Study, the Dutch 2006–2007 study and the DEOLD study were approved by the Medical Ethics Review Committee of the VU University Medical Center in Amsterdam. The Missouri LRI Study was approved by institutional Review Boards at the University of Missouri-Columbia Health Sciences Center and Washington University School of Medicine. Additionally, several hospital Institutional Review Boards conducted their own evaluations before granting permission to review medical records, and two nursing homes had ethical review panels that approved the project. The Bedford US study was approved by the local Institutional Review Board of the E.N. Rogers Memorial Veterans Hospital in Bedford, Massachusetts.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Public posting of individual level participant data is not covered by an informed patient consent procedure. A data set containing pseudonyms can be obtained by collaborating scientists on approval of a scientific project proposal. Any questions about the data can be directed to the corresponding author.

REFERENCES

1. Attems J, König C, Huber M, *et al.* Cause of death in demented and non-demented elderly inpatients; an autopsy study of 308 cases. *J Alzheimers Dis* 2005;8:57–62.
2. Brunnström HR, Englund EM. Cause of death in patients with dementia disorders. *Eur J Neurol* 2009;16:488–92.
3. van der Steen JT, Mehr DR, Kruse RL, *et al.* Predictors of mortality for lower respiratory infections in nursing home residents with dementia were validated transnationally. *J Clin Epidemiol* 2006;59:970–9.
4. Keene J, Hope T, Fairburn CG, *et al.* Death and dementia. *Int J Geriat Psychiatry* 2001;16:969–74.
5. van der Steen JT, Muller MT, Ooms ME, *et al.* Decisions to treat or not to treat pneumonia in demented psychogeriatric nursing home patients: development of a guideline. *J Med Ethics* 2000;26:114–20.
6. van der Maaden T, Hendriks SA, de Vet HCW, *et al.* Antibiotic use and associated factors in patients with dementia: a systematic review. *Drugs Aging* 2015;32:43–56.
7. van der Steen JT, Albers G, Licht-Strunk E, *et al.* A validated risk score to estimate mortality risk in patients with dementia and pneumonia: barriers to clinical impact. *Int Psychogeriatr* 2011;23:31–43.
8. Janssen KJ, Vergouwe Y, Kalkman CJ, *et al.* A simple method to adjust clinical prediction models to local circumstances. *Can J Anaesth* 2009;56:194–201.
9. Toll DB, Janssen KJ, Vergouwe Y, *et al.* Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol* 2008;61:1085–94.
10. Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. In: Lyman GH, Burstein H, eds. *Breast cancer translational therapeutic strategies*. New York: Informa Healtcare, 2007:11–25.
11. Perel P, Edwards P, Wentz R, *et al.* Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak* 2006;6:38.
12. van der Steen JT, Ooms ME, van der Wal G, *et al.* Withholding or starting antibiotic treatment in patients with dementia and pneumonia: prediction of mortality with physicians' judgment of illness severity and with specific prognostic models. *Med Decis Making* 2005;25:210–21.
13. Steyerberg EW, Borsboom GJ, van Houwelingen HC, *et al.* Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;23:2567–86.
14. Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*. Springer, 2009.
15. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, 2001.
16. van der Steen JT, Ooms ME, Adèr HJ, *et al.* Withholding antibiotic treatment in pneumonia patients with dementia: a quantitative observational study. *Arch Intern Med* 2002;162:1753–60.
17. Mehr DR, Binder EF, Kruse RL, *et al.* Clinical findings associated with radiographic pneumonia in nursing home residents. *J Fam Pract* 2001;50:931–7.
18. Mehr DR, Binder EF, Kruse RL, *et al.* Predicting mortality in nursing home residents with lower respiratory tract infection: the Missouri LRI study. *JAMA* 2001;286:2427–36.
19. Morris JN, Fries BE, Mehr DR, *et al.* MDS cognitive performance scale(C). *J Gerontol* 1994;49:M174–82.
20. van der Steen JT, Helton MR, Ribbe MW. Prognosis is important in decisionmaking in Dutch nursing home patients with dementia and pneumonia. *Int J Geriatr Psychiatry* 2009;24:933–6.
21. van der Steen JT, Meuleman-Peperkamp I, Ribbe MW. Trends in treatment of pneumonia among Dutch nursing home patients with dementia. *J Palliat Med* 2009;12:789–95.
22. van der Steen JT, Lane P, Kowall NW, *et al.* Antibiotics and mortality in patients with lower respiratory infection and advanced dementia. *J Am Med Dir Assoc* 2012;13:156–61.
23. van der Steen JT, Ribbe MW, Deliens L, *et al.* Retrospective and prospective data collection compared in the Dutch End Of Life in Dementia (DEOLD) study. *Alzheimer Dis Assoc Disord* 2014;28:88–94.
24. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130:515–24.
25. Hosmer J, Lemeshow S, Sturdivant RX. *Assessing the fit of the model. Applied logistic regression*. John Wiley & Sons, 2013:177.
26. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, *et al.* Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27:157–72.
27. Leening MJG, Vedder MM, Witteman JCM, *et al.* Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Ann Intern Med* 2014;160:122–31.
28. Janssen KJ, Moons KG, Kalkman CJ, *et al.* Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008;61:76–86.
29. Bouwmeester W, Zuithoff NP, Mallett S, *et al.* Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9:1–12.
30. Moons KG, Kengne AP, Grobbee DE, *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;98:691–8.
31. Volicer L, Hurley AC, Lathi DC, *et al.* Measurement of severity in advanced Alzheimer's disease. *J Gerontol* 1994;49:M223–6.