

RESEARCH ARTICLE

Open Access

Assessing discriminative ability of risk models in clustered data

David van Klaveren^{1*}, Ewout W Steyerberg¹, Pablo Perel² and Yvonne Vergouwe¹

Abstract

Background: The discriminative ability of a risk model is often measured by Harrell's concordance-index (c-index). The c-index estimates for two randomly chosen subjects the probability that the model predicts a higher risk for the subject with poorer outcome (concordance probability). When data are clustered, as in multicenter data, two types of concordance are distinguished: concordance in subjects from the same cluster (within-cluster concordance probability) and concordance in subjects from different clusters (between-cluster concordance probability). We argue that the within-cluster concordance probability is most relevant when a risk model supports decisions within clusters (e.g. who should be treated in a particular center). We aimed to explore different approaches to estimate the within-cluster concordance probability in clustered data.

Methods: We used data of the CRASH trial (2,081 patients clustered in 35 centers) to develop a risk model for mortality after traumatic brain injury. To assess the discriminative ability of the risk model within centers we first calculated cluster-specific c-indexes. We then pooled the cluster-specific c-indexes into a summary estimate with different meta-analytical techniques. We considered fixed effect meta-analysis with different weights (equal; inverse variance; number of subjects, events or pairs) and random effects meta-analysis. We reflected on pooling the estimates on the log-odds scale rather than the probability scale.

Results: The cluster-specific c-index varied substantially across centers (*IQR* = 0.70-0.81; $I^2 = 0.76$ with 95% confidence interval 0.66 to 0.82). Summary estimates resulting from fixed effect meta-analysis ranged from 0.75 (equal weights) to 0.84 (inverse variance weights). With random effects meta-analysis – accounting for the observed heterogeneity in c-indexes across clusters – we estimated a mean of 0.77, a between-cluster variance of 0.0072 and a 95% prediction interval of 0.60 to 0.95. The normality assumptions for derivation of a prediction interval were better met on the probability than on the log-odds scale.

Conclusion: When assessing the discriminative ability of risk models used to support decisions at cluster level we recommend meta-analysis of cluster-specific c-indexes. Particularly, random effects meta-analysis should be considered.

Keywords: Clustered data, Concordance, Discrimination, Meta-analysis, Prediction, Risk model

Background

Assessing the performance of a risk model is of great practical importance. An essential aspect of model performance is separating subjects with good outcome from subjects with poor outcome (discrimination) [1]. The concordance probability is a commonly used measure of discrimination reflecting the association between model predictions and true outcomes [2,3]. For binary outcome

data it is the probability that a randomly chosen subject from the event group has a higher predicted probability of having an event than a randomly chosen subject from the non-event group. For time-to-event outcome data it is the probability that, for a randomly chosen pair of subjects, the subject who experiences the event of interest earlier in time has a lower predicted value of the time to the occurrence of the event. For both kinds of outcome data the concordance probability is often estimated with Harrell's concordance (c)-index [2].

* Correspondence: d.vanklaveren.1@erasmusmc.nl

¹Department of Public Health, Erasmus MC, Dr. Molewaterplein 50, Rotterdam 3015 GE, The Netherlands

Full list of author information is available at the end of the article

In risk modelling, clustered data are frequently used. A typical example is multicenter patient data, i.e. data of patients who are treated in different centers with similar inclusion criteria across the centers. Patients treated in the same center are nevertheless more alike than patients from different centers. A comparable type of clustering may occur in patients treated in different countries or in patients treated by different caregivers in the same center. Similarly, in public health research the study population is often clustered in geographical regions like countries, municipalities or neighbourhoods. It has been suggested that clustering should be taken into account in the development of risk models to obtain unbiased estimates of predictor effects [4]. This can be done by using a multi-level logistic regression model for binary outcomes or a frailty model for time-to-event outcomes [5,6].

It would be natural to take clustering also into account when measuring the performance of a risk model. For multilevel models, it has been proposed to consider the concordance probability of subjects within the same cluster (within-cluster concordance probability) separately from the concordance probability of subjects in different clusters (between-cluster concordance probability) [7,8]. We propose using the within-cluster concordance probability when risk models are used to support decisions within clusters, e.g. in clinical practice where decisions on interventions are commonly taken within centers. A valuable risk model should then be able to separate subjects within the same cluster into those with good outcome and poor outcome. We consider the within-cluster concordance probability more relevant in this context than the between-cluster or overall concordance probability.

Here, we aimed to estimate the within-cluster concordance probability from clustered data. We explored different meta-analytic methods for pooling cluster-specific concordance probability estimates with an illustration in predicting mortality among patients suffering from traumatic brain injury.

Methods

Mortality in traumatic brain injury patients

We present a case study of predicting mortality after Traumatic Brain Injury (TBI). Risk models using baseline characteristics provide adequate discrimination between patients with good and poor 6-month outcomes after TBI [9,10]. We used patients enrolled in the Medical Research Council Corticosteroid Randomisation after Significant Head Injury [11] trial (registration ISRCTN74459797, <http://www.controlled-trials.com/>), who were recruited between 1999 and 2004. This was a large international double-blind, randomized placebo-controlled trial of the effect of early administration of a 48-h infusion of methylprednisolone on outcome after head injury. The trial included 10,008 adults clustered in 239 centers

with Glasgow Coma Scale (GCS) [12] Total Score ≤ 14 , who were enrolled within 8 hours after injury. By design the patient inclusion criteria were equal in all 239 centers.

We considered patients with moderate or severe brain injury (GCS Total Score ≤ 12) and observed 6-month Glasgow Outcome Scale (GOS) [13]. Patients who were treated in one of 35 European centers with more than 5 patients experiencing the event ($n = 2,081$), were used to assess the discriminative ability of a prediction model developed with data from 35 centers. Patients who were treated in one of 21 Asian centers with more than 5 patients experiencing the event ($n = 1,421$) were used to assess the discriminative ability at external validation.

We used a Cox proportional hazards model with age, GCS Motor Score and pupil reactivity as covariates similar to previously developed risk models [9,10]. We modelled center with a Gamma frailty (random effect) to account for heterogeneity in mortality among centers. We estimated parameters on the European selection of patients with the R package survival [14,15]. As center effect estimates are unavailable when using a risk model in new centers, we calculated individual risk predictions applying the Gamma frailty mean of 1 for each patient.

Cluster-specific concordance probabilities

We estimated the concordance probability within each cluster by Harrell's c-index [2], i.e. the proportion of all usable pairs of subjects in which the predictions are concordant with the outcomes. A pair of subjects is usable if we can determine the ordering of their outcomes. For binary outcomes, pairs of subjects are usable if one of the subjects had an event and the other did not. For time-to-event outcomes, pairs of subjects are usable if their failure times are not equal and at least the smallest failure time is uncensored. For a usable subject pair the predictions are concordant with the outcomes if the ordering of the predictions is equal to the ordering of the outcomes. Values of the c-index close to 0.5 indicate that the model does not perform much better than a coin-flip in predicting which subject of a randomly chosen pair will have a better outcome. Values of the c-index near 1 indicate that the model is almost perfectly able to predict which subject of a randomly chosen pair will have a favourable outcome. We estimated the variances of the cluster-specific c-indexes with a method proposed by Quade [16]. Formulas are provided in Appendix 1.

Pooling cluster-specific concordance probability estimates

The within-cluster concordance probability C_w can be estimated by pooling the cluster-specific concordance probability estimates into a weighted average. Previously, the cluster-specific concordance probability estimates were pooled with the number of usable subject pairs as

weights [7,8]. Here, we define eight different ways for pooling of cluster-specific estimates – both on the probability scale and on the log-odds scale – based on fixed effect meta-analysis and random effects meta-analysis.

We consider a dataset with subjects in K clusters. Let m_k be the number of subjects and e_k be the number of events in cluster k . We denote the number of usable subject pairs – pairs of subjects for whom we can determine the ordering of their outcomes – in cluster k by n_k . The cluster-specific concordance probability estimate for cluster k is denoted by \hat{C}_k with sampling variance estimate $\hat{\sigma}_k^2$.

Fixed effect meta-analysis

Fixed effect meta-analysis assumes that one common within-cluster concordance probability C_W exists that applies to all clusters. The observed cluster-specific estimates vary only because of chance created from sampling subjects. Fixed effect meta-analysis with cluster weights w_k results in:

$$\hat{C}_W = \frac{\sum_k w_k \hat{C}_k}{\sum_k w_k} \text{ with } \hat{\sigma}_{\hat{C}_W}^2 = \frac{\sum_k w_k^2 \hat{\sigma}_k^2}{\left(\sum_k w_k\right)^2} \quad (1)$$

The simplest approach would be to apply equal weights, $w_k = 1/K$ for each cluster (method 1). This estimator is quite naive when the cluster size varies, because small clusters are given the same weight as large clusters and information about the precision of the cluster-specific estimates is ignored. Heuristic choices of weights taking the cluster size into account are the number of subjects, $w_k = m_k$ (method 2), or the number of events, $w_k = e_k$ (method 3). Analogous to the definition of the c-index a fourth option is the number of usable subject pairs as weights, $w_k = n_k$ (method 4). The pooled estimate is then equal to the proportion of all usable within-cluster subject pairs in which the predictions and outcomes are concordant. Another choice of meta-analysis weights are the inverse variances, $w_k = 1/\hat{\sigma}_k^2$ (method 5). These weights express the precision of the cluster-specific estimates and are commonly used in meta-analysis of study-specific treatment effects.

Random effects meta-analysis

In our context a random effects meta-analysis considers that the cluster-specific estimates vary not only because of sampling variability but also because of differences in true concordance probabilities. This is appropriate for high values of I^2 [17]. I^2 measures the proportion of variability in cluster-specific estimates that is due to between-cluster heterogeneity rather than chance. Random effects meta-analysis assumes that cluster-specific concordance probabilities C_k are distributed about mean μ with between-cluster variance τ^2 , with the observed \hat{C}_k

normally distributed about C_k with sampling variance σ_k^2 . The mean within-cluster concordance probability estimate $\hat{\mu}$ is the average of the cluster-specific estimates with the inverse variances as weights (method 6):

$$\hat{\mu} = \frac{\sum_k w_k \hat{C}_k}{\sum_k w_k}, \hat{\sigma}_{\hat{\mu}}^2 = \frac{\sum_k w_k^2 (\hat{\sigma}_k^2 + \hat{\tau}^2)}{\left(\sum_k w_k\right)^2} \\ = \frac{1}{\sum_k w_k} \text{ with } w_k = 1/(\hat{\sigma}_k^2 + \hat{\tau}^2) \quad (2)$$

For estimation of the between-cluster variance τ^2 we used the DerSimonian and Laird [18] method. Alternative estimators for τ^2 can be found in DerSimonian and Kacker [19].

With the additional assumption of normally distributed C_k we can derive a prediction interval for the within-cluster concordance probability C_W in a new or unspecified cluster [20]. If τ^2 were known, then $\hat{\mu} \sim N(\mu, \hat{\sigma}_{\hat{\mu}}^2)$ and $C_W \sim N(\mu, \tau^2)$ imply (assuming independence of C_W and $\hat{\mu}$ given μ) that $C_W - \hat{\mu} \sim N(0, \tau^2 + \hat{\sigma}_{\hat{\mu}}^2)$. Hence the within-cluster concordance probability C_W in a new cluster is normally distributed, with mean $\hat{\mu}$ and variance $\tau^2 + \hat{\sigma}_{\hat{\mu}}^2$ (Figure 1). Since τ^2 is estimated, we assume $\frac{C_W - \hat{\mu}}{\sqrt{\hat{\tau}^2 + \hat{\sigma}_{\hat{\mu}}^2}}$ to take a more conservative t-distribution with $K - 2$ degrees of freedom instead of the standard normal distribution

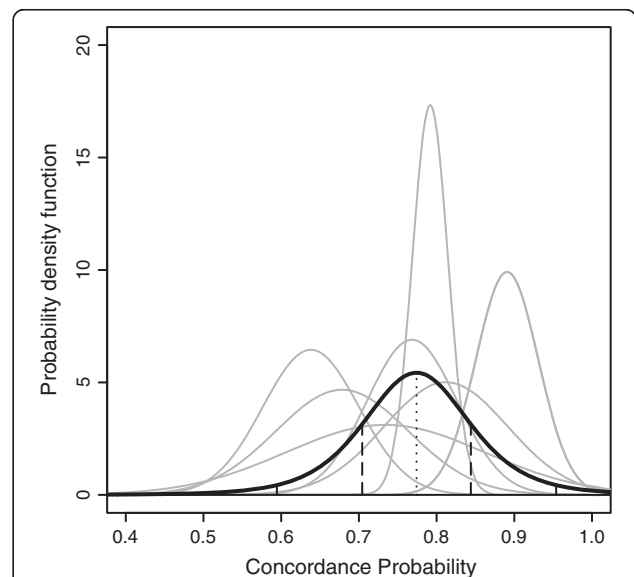


Figure 1 Example of random effects meta-analysis of concordance probability estimates in 7 clusters. Cluster-specific estimates are in grey. Under the assumption of normally distributed cluster-specific concordance probabilities, the predictive distribution resulting from a random effects meta-analysis is in black. The mean, the mean \pm one standard deviation and the 2.5 and 97.5 percentiles of the predictive distribution are plotted with vertical lines.

[20]. Thus, a 95% prediction interval of the within-cluster concordance probability C_W in an unspecified cluster can be approximated by: $\hat{\mu} \pm t_{K-2}^{0.975} \sqrt{\hat{\tau}^2 + \hat{\sigma}_\mu^2}$ with $t_{K-2}^{0.975}$ denoting the 97.5% percentile of the t-distribution with $K - 2$ degrees of freedom.

Meta-analysis scale

When calculating a prediction interval of the within-cluster concordance probability C_W , Riley et al [21] advised to perform a random effects meta-analysis on a scale that helps meet the normality assumption for the random effects. When the normality assumption of the random effects model holds, the C_k are normally distributed with mean μ and variance $\tau^2 + \sigma_k^2$. As a consequence, the standardized residuals z_k defined below should approximately have a standard normal distribution:

$$z_k = (\hat{C}_k - \hat{\mu}) / \sqrt{\hat{\tau}^2 + \hat{\sigma}_k^2} \tag{3}$$

To consider if the normality assumption is valid we used a normal probability plot of z_k and applied the Shapiro-Wilk test to z_k [22]. In a normal probability plot z_k is plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality. The Shapiro-Wilk test returns the probability of obtaining the test-statistic as least as extreme as the observed one, under the null-hypothesis that z_k are normally distributed (p-value). When the p-value is above significance level α , say 5%, the null hypothesis that z_k is normally distributed is not rejected.

Since the concordance probability is restricted to [0, 1] the normality assumption of random effects meta-analysis may be violated. We considered inverse variance weighted meta-analysis on the log-odds scale as an alternative

approach (methods 7 and 8 for fixed effect and random effects meta-analysis respectively). The resulting estimators for the within-cluster concordance probability are defined in Appendix 2. The normality assumption on log-odds scale was again assessed by the normal probability plot and the Shapiro-Wilk test.

Table 1 contains a summary of the eight pooling methodologies described above. For all the meta-analyses we used the R package rmeta [14,23].

Results

The European patients were slightly older in comparison with the Asian patients (median age 36 vs. 31 years) and were more likely to have the worst GCS Motor Score of 1, i.e. no motor response (21% versus 4%) compared to the Asian patients (Table 2). However, 6 month mortality was lower in the European patients (27%) than in the Asian patients (35%).

We found that 6-month mortality was clearly associated with higher age, worse GCS Motor Score and less pupil reactivity (Table 3). Heterogeneity in mortality among European centers was substantial as indicated by the hazard ratio of 1.7 for the 75 percentile versus the 25 percentile of the random center effect, based on the quartiles of the Gamma frailty distribution with mean 1 and variance estimate 0.146.

Among European centers (overall c-index 0.80) the c-indexes varied substantially with an interquartile range of 0.70 to 0.81 (Figure 2). Pooled concordance probability estimates resulting from fixed effect meta-analysis ranged from 0.75 (equal weights) to 0.84 (inverse variance weights). Random effects meta-analysis (method 6) led to a mean concordance probability estimate $\hat{\mu} = 0.77$, a between-cluster variance estimate $\hat{\tau}^2 = 0.0072$ and a wide 95% prediction interval (0.60 to 0.95) reflecting the strong

Table 1 Overview of the 8 methods for pooling of cluster-specific concordance probability estimates

	Fixed effect meta-analysis Assuming the same true (logit) concordance probability within each cluster	Random effects meta-analysis Assuming variation in true (logit) concordance probabilities across clusters
Probability scale		
Meta-analysis of cluster-specific estimates of the concordance probability	1. Equal weight for each cluster 2. Number of subjects in the cluster 3. Number of subjects in the cluster with an event 4. Number of usable subject pairs within the cluster 5. Inverse of the cluster-specific sampling variance estimate	6. Inverse of the sum of the cluster-specific sampling variance estimate and the between-cluster variance estimate
Log-odds scale		
Meta-analysis of cluster-specific estimates of the logit concordance probability	7. Inverse of the cluster-specific sampling variance estimate on log-odds scale	8. Inverse of the sum of the cluster-specific sampling variance estimate on log-odds scale and the between-cluster variance estimate on log-odds scale

Table 2 Patient characteristics in selected European and Asian centers

Characteristic	Measure or Category	Europe		Asia	
Age (years)	Median (25–75 percentile)	36	(24–53)	31	(22–43)
GCS Motor score	No response (1)	445	(21%)	55	(4%)
	Extension (2)	134	(6%)	96	(7%)
	Abnormal flexion (3)	176	(8%)	124	(9%)
	Normal flexion (4)	321	(15%)	261	(18%)
	Localizes/obeys (5/6)	1,005	(48%)	885	(62%)
Pupil reactivity	No pupil reacted	291	(14%)	129	(9%)
	One pupil reacted	123	(6%)	117	(8%)
	Both pupils reacted	1,667	(80%)	1,175	(83%)
Six-month mortality	Dead	553	(27%)	495	(35%)
Patients	Total	2,081		1,421	
Centers	Total	35		21	
Patients per center	Median (25–75 percentile)	33	(21–64)	34	(20–66)

heterogeneity in the cluster-specific concordance probabilities ($I^2 = 0.76$ with 95% confidence interval 0.66 to 0.82). Random effects meta-analysis on log-odds scale (method 8) led to similar results, but with a somewhat smaller asymmetric prediction interval (0.58 to 0.89).

Large differences in pooling weights, together with heterogeneity in the cluster-specific concordance probabilities, led to very different pooled estimates. We analysed the pooling weights to explain the differences in pooled estimates (Figure 3). The patient-weighted estimate was dominated by center 2 with 494 of the 2,081 patients. The event-weighted estimate was dominated by center 12 with 107 out of 553 events. The patient-pair-weighted estimate was heavily determined by both center 2 and center 12 as the number of usable patient pairs is related to the number of patients times the number of events. The fixed effect inverse-variance weighted estimate was also strongly influenced by centers with high number of patients or

events, because the standard errors of the cluster-specific estimates depend heavily on the number of patients and events. Furthermore, the fixed effect inverse-variance weighted estimate was upwardly influenced by center 1 as a result of the small standard error relative to the small number of patients and events. The random effects inverse-variance weighted estimate was much less dominated by particular centers and close to the equally weighted estimate because of the large amount of heterogeneity. The standard error on the log-odds scale increased with increasing c-index according to Equation 10 in Appendix 2 and therefore put less weight on the centers with a high concordance probability estimate resulting in lower pooled estimates. The large standard errors for centers with high c-index also decreased the heterogeneity ($I^2 = 0.61$ with 95% confidence interval 0.44 to 0.73) on the log-odds scale resulting in more similar weights for fixed effect and random effects meta-analysis.

To check the validity of the normality assumption in the random effects meta-analyses, we calculated standardized residuals (Equation 3), both on the probability and the log-odds scale. The standardized residuals better fitted to the standard normal distribution on the probability scale than on the log-odds scale (Figure 4, p-values for rejection of the normality null hypothesis of 0.666 on probability scale and of 0.030 on log-odds scale).

To illustrate the comparison in an external validation setting, we repeated the analysis of the within-cluster concordance probability in Asian centers with the same risk model (Figure 5). Among Asian clusters (overall c-index 0.74) the c-indexes varied less (*IQR* 0.71–0.78), which was reflected in a lower proportion of variation among clusters that is due to heterogeneity rather than chance ($I^2 = 0.32$ with 95% confidence interval 0 to 0.60). As a result, different pooling methodologies led to more

Table 3 Associations between predictors and 6-month mortality in European centers

Characteristic	Level	HR (95 % CI)	
Age (years)	47 versus 23*	2.1	(1.9–2.4)
GCS Motor score	No response (1)	3.1	(2.4–4.0)
	Extension (2)	2.8	(2.0–3.8)
	Abnormal flexion (3)	2.4	(1.7–3.2)
	Normal flexion (4)	1.5	(1.1–2.0)
	Localizes/obeys (5/6)	1.0	(ref)
Pupil reactivity	No pupil reacted	2.8	(2.3–3.5)
	One pupil reacted	1.7	(1.2–2.3)
	Both pupils reacted	1.0	(ref)
Center random effect	75 versus 25 percentile	1.7	

*Interquartile range.

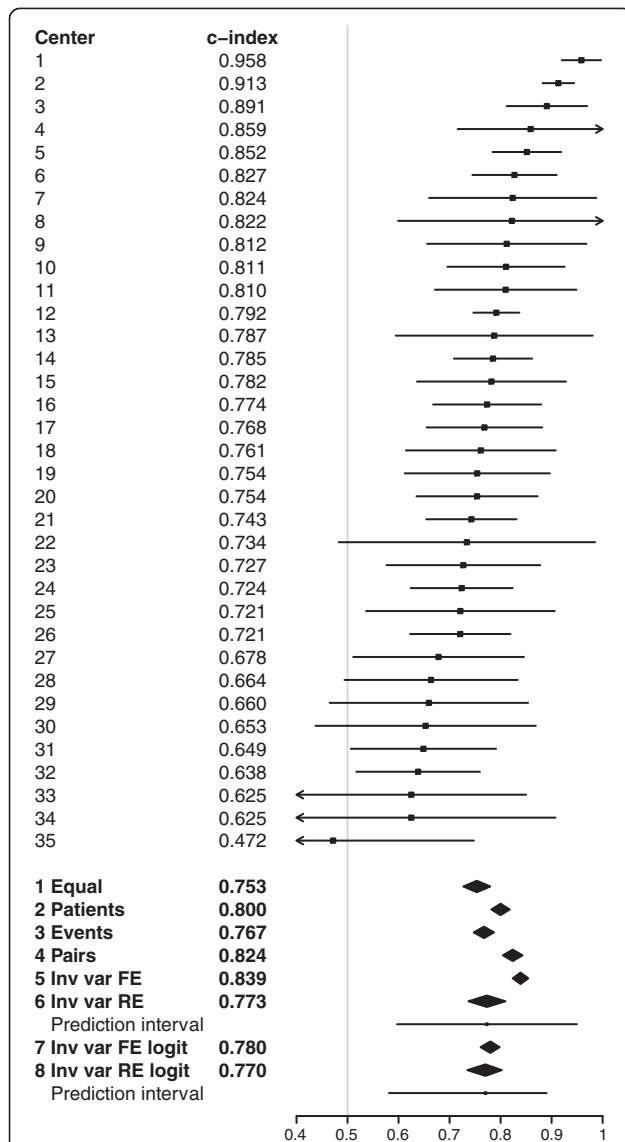


Figure 2 Center-specific and pooled concordance probability estimates with 95% confidence intervals for European centers.
 For pooled estimates based on random effects meta-analysis a 95% prediction interval for the concordance probability is presented by a horizontal line. 1 Equal = Fixed effect meta-analysis with equal weights; 2 Patients = Fixed effect meta-analysis with number of patients as weights; 3 Events = Fixed effect meta-analysis with number of events as weights; 4 Pairs = Fixed effect meta-analysis with number of usable patient pairs as weights; 5 Inv var FE = Fixed effect meta-analysis with inverse variance weights; 6 Inv var RE = Random effects meta-analysis with inverse variance weights; 7 Inv var FE logit = Fixed effect meta-analysis with inverse variance weights on log-odds scale; 8 Inv var RE logit = Random effects meta-analysis with inverse variance weights on log-odds scale.

similar pooled estimates, because differences in cluster weights have less impact when cluster-specific estimates are more alike. Based on random effects meta-analysis, estimates of the mean within-cluster concordance probability and the between-cluster variance were $\hat{\mu} = 0.75$ and

$\hat{\tau}^2 = 0.0013$ respectively. The resulting prediction interval (0.67 to 0.83) was much smaller than for the European clusters. The heterogeneity disappeared on the log-odds scale ($I^2 = 0$) leading to equal estimates by fixed effect and random effects meta-analysis.

Discussion

We studied how to assess the discriminative ability of risk models in clustered data. The within-cluster concordance probability is an important measure for risk models when these models are used to support decisions on interventions within the clusters. The within-cluster concordance probability can be estimated by pooling cluster-specific concordance probability estimates (e.g. c-indexes) with a meta-analysis, similar to pooling of study-specific treatment effect estimates. We considered different pooling strategies (Table 1) and recommend random effects meta-analysis in case of substantial variability – beyond chance – of the concordance probability across clusters [20,21]. To decide if the meta-analysis should be undertaken on the probability scale or the log-odds scale we suggest considering the normality assumptions on both scales by normal probability plots and Shapiro-Wilk tests of the standardized residuals.

The illustration of predicting 6-month mortality after TBI prompted the use of random effects meta-analysis because of the strong difference – beyond chance – in concordance probability among centers. This was clearly visualized by the forest plot in Figure 2. Random effects meta-analysis results can be summarized by the mean concordance probability and a 95% prediction interval for possible values of the concordance probability. By definition, these results give insight into the variation of the discriminative ability among centers as opposed to fixed effect meta-analysis results [20,21]. By comparing normal probability plots and Shapiro-Wilk test results based on the standardized residuals we concluded the random effects meta-analysis results on probability scale to be the most appropriate (Figure 4). Although the methodology is illustrated with time-to-event outcomes of traumatic brain injury patients, it is also applicable to binary outcomes.

Even if a risk model contains regression coefficients that are optimal for the data in each cluster, differences in case mix may lead to different concordance probabilities across clusters [24]. Furthermore, predictor effects may vary because of cluster-specific circumstances, also leading to different cluster-specific concordance probabilities. Given the variability beyond chance in our case study, we consider a random effects meta-analysis of the cluster-specific c-indexes as most appropriate.

The assumption of random effects meta-analysis is that underlying concordance probabilities among clusters are

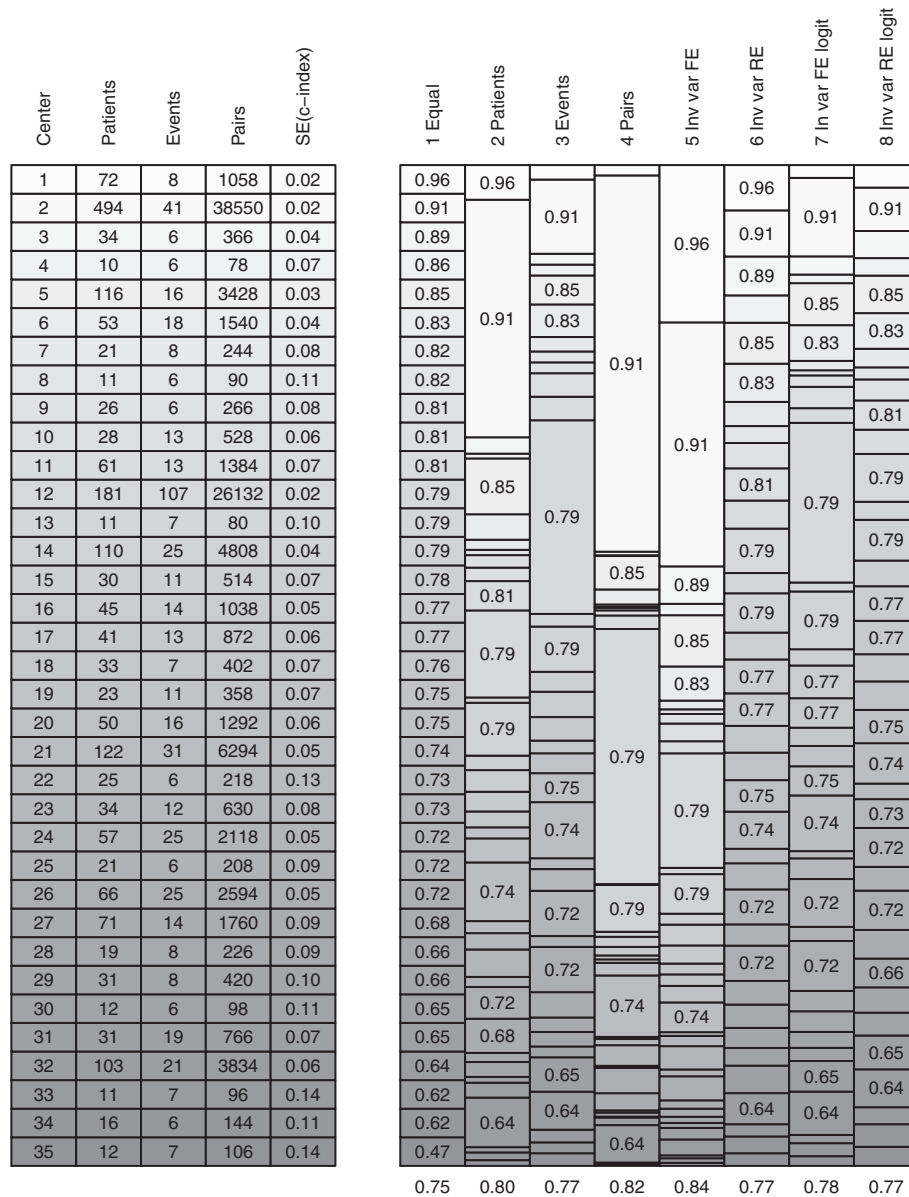
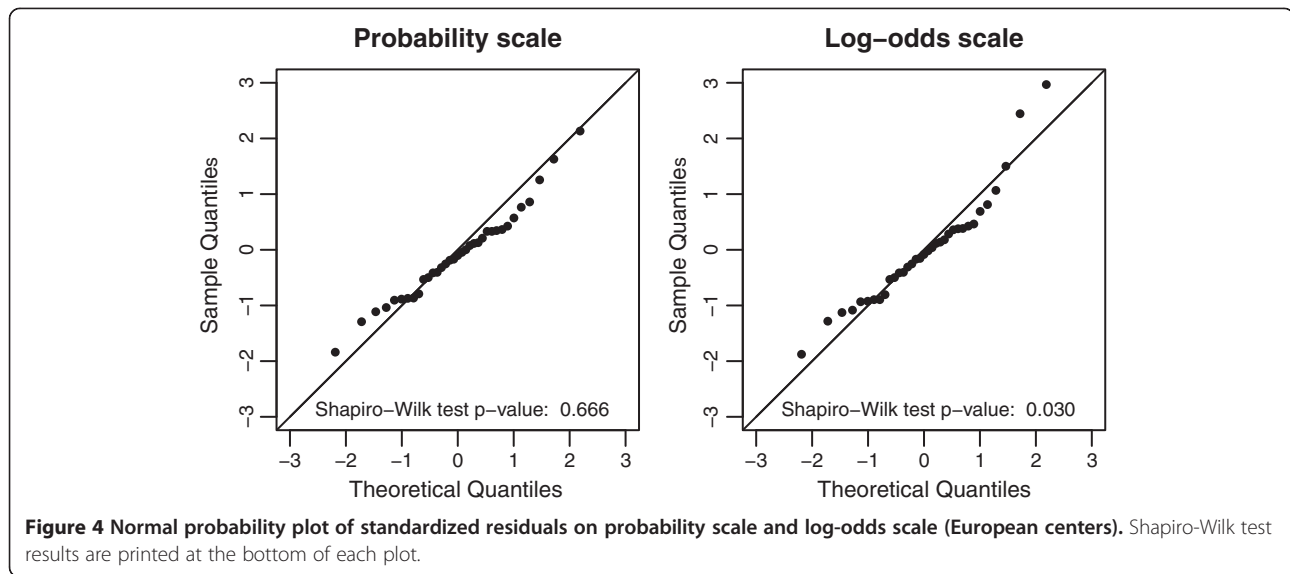


Figure 3 Meta-analysis pooling weights for European centers. For methods 1 to 8 the weights are represented by the height of the bars on the right hand side of the Figure. C-indexes are printed in the bars if the cluster weight was at least equal to the average weight. 1 Equal = Fixed effect meta-analysis with equal weights; 2 Patients = Fixed effect meta-analysis with number of patients as weights; 3 Events = Fixed effect meta-analysis with number of events as weights; 4 Pairs = Fixed effect meta-analysis with number of usable patient pairs as weights; 5 Inv var FE = Fixed effect meta-analysis with inverse variance weights; 6 Inv var RE = Random effects meta-analysis with inverse variance weights; 7 Inv var FE logit = Fixed effect meta-analysis with inverse variance weights on log-odds scale; 8 Inv var RE logit = Random effects meta-analysis with inverse variance weights on log-odds scale.

exchangeable, i.e. cluster-specific concordance probabilities are expected to be non-identical, yet identically distributed [20]. If part of the variation can be explained by cluster characteristics, a meta-regression – assuming partial exchangeability – of the concordance probability estimates with cluster characteristics as covariates is preferable.

We chose to analyse the concordance probability as it is the most commonly used measure of discriminative ability of a risk model. However, the same logic of pooling cluster-specific performance measure estimates can be applied to any other performance measure, like the discrimination slope, the explained variation (R^2) or the Brier score [25].



We used Harrell’s c-index to estimate cluster-specific concordance probabilities together with Quade’s formula for the cluster-specific variances of the c-index [2,16]. The same methodology of pooling cluster-specific performance measure estimates can be applied to other concordance probability estimators and its variances. Other estimators for the concordance probability in time-to-event data can be found in Gönen and Heller [26] and Uno et al [27]. These estimators are especially favourable when censoring varies by cluster as they are shown to be less sensitive to censoring distributions. Other variance estimators are described by Hanley and McNeil [28], and DeLong et al [29] for binary outcome data and by Nam and D’Agostino [30] and Pencina and D’Agostino [3] for time-to-event outcome data. The variance of the concordance probability estimate can also be estimated with a bootstrap procedure [31].

Conclusion

We recommend meta-analysis of cluster-specific c-indexes when assessing discriminative ability of risk models used to support decisions at cluster level. Particularly, random effects meta-analysis should be considered as it allows for and provides insight into the variability of the concordance probability among clusters.

Appendix 1

The concordance probability is defined as the probability that a randomly chosen subject pair with different outcomes is concordant. For a randomly chosen subject pair (i, j) with outcomes Y_i and Y_j and model predictions \hat{Y}_i and \hat{Y}_j the concordance probability C is:

$$C = \Pr(\hat{Y}_i < \hat{Y}_j | Y_i < Y_j) \quad (4)$$

Harrell’s c-index [2] estimates the concordance probability by the proportion of all usable pairs of subjects (n_u) in which the predictions and outcomes are concordant (n_c), with tied predictions (n_t) counted as 1/2:

$$\hat{C} = \frac{n_c + n_t/2}{n_u} \quad (5)$$

For binary outcomes y , pairs of subjects are usable if one of the subjects had an event and the other did not. The number of usable subject pairs n_u , the number of concordant subject pairs n_c and the number of tied subject pairs n_t are:

$$\begin{aligned} n_u &= \sum_i \sum_j I(y_i < y_j) \\ n_c &= \sum_i \sum_j I(y_i < y_j \text{ and } \hat{y}_i < \hat{y}_j) \\ n_t &= \sum_i \sum_j I(y_i < y_j \text{ and } \hat{y}_i = \hat{y}_j) \end{aligned} \quad (6)$$

For time-to-event outcomes y , pairs of subjects are usable if their survival times are not equal and at least the smallest survival time is uncensored. We have to add the restriction that the smallest observation y_i of each subject pair is uncensored, denoted by $\delta_i = 1$:

$$\begin{aligned} n_u &= \sum_i \sum_j I(y_i < y_j \text{ and } \delta_i = 1) \\ n_c &= \sum_i \sum_j I(y_i < y_j \text{ and } \delta_i = 1 \text{ and } \hat{y}_i < \hat{y}_j) \\ n_t &= \sum_i \sum_j I(y_i < y_j \text{ and } \delta_i = 1 \text{ and } \hat{y}_i = \hat{y}_j) \end{aligned} \quad (7)$$

The variance of the c-index can be estimated according to Quade [16]:

$$\hat{\sigma}_{\hat{C}}^2 = \frac{\sum n_{u,i}^2 (\sum n_{c-d,i})^2 - 2 \sum n_{u,i} \sum n_{c-d,i} \sum n_{u,i} n_{c-d,i} + (\sum n_{u,i})^2 \sum n_{c-d,i}^2}{(\sum n_{u,i})^4} \quad (8)$$

All summations over i with $n_{u,i}$ and $n_{c-d,i}$ the number of usable and the number of concordant minus discordant subject pairs of which subject i is one:

$$\begin{aligned} n_{u,i} &= \sum_j I(y_i < y_j \text{ and } \delta_i = 1) \\ n_{c,i} &= \sum_j I(y_i < y_j \text{ and } \delta_i = 1 \text{ and } \hat{y}_i < \hat{y}_j) \\ n_{d,i} &= \sum_j I(y_i < y_j \text{ and } \delta_i = 1 \text{ and } \hat{y}_i > \hat{y}_j) \\ n_{c-d,i} &= n_{c,i} - n_{d,i} \end{aligned} \quad (9)$$

Appendix 2

Based on the delta method, a variance estimator for the logit of the c-index is:

$$\begin{aligned} \text{var}(\text{logit}(\hat{C})) &= \text{var}\left(\log\left(\frac{\hat{C}}{1-\hat{C}}\right)\right) \\ &= \frac{\text{var}(\hat{C})}{(\hat{C}(1-\hat{C}))^2} \end{aligned} \quad (10)$$

We used this variance estimator to perform a meta-analysis on log-odds scale. The pooling weights (method 7) for a fixed effect inverse variance meta-analysis on log-odds scale are:

$$w_k = \left[\frac{\hat{\sigma}_k^2}{(\hat{C}_k(1-\hat{C}_k))^2} \right]^{-1} \quad (11)$$

The pooling weights (method 8) for a random effects inverse variance meta-analysis on log-odds scale are:

$$w_k = \left[\frac{\hat{\sigma}_k^2}{(\hat{C}_k(1-\hat{C}_k))^2} + \hat{\tau}^2 \right]^{-1} \quad (12)$$

The resulting pooled estimates together with confidence and prediction intervals are transformed back to probability scale.

Abbreviations

c-index: Concordance-index; CRASH: Corticosteroid randomisation after significant head injury; GCS: Glasgow coma scale; GOS: Glasgow outcome scale; IQR: Interquartile range.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DK, ES and YV designed the study. PP participated in the collection of data and organisation of the databases from which this manuscript was developed. DK and YV analysed the data and wrote the first draft of the manuscript. All authors contributed to writing the manuscript and read and approved the final manuscript.

Acknowledgements

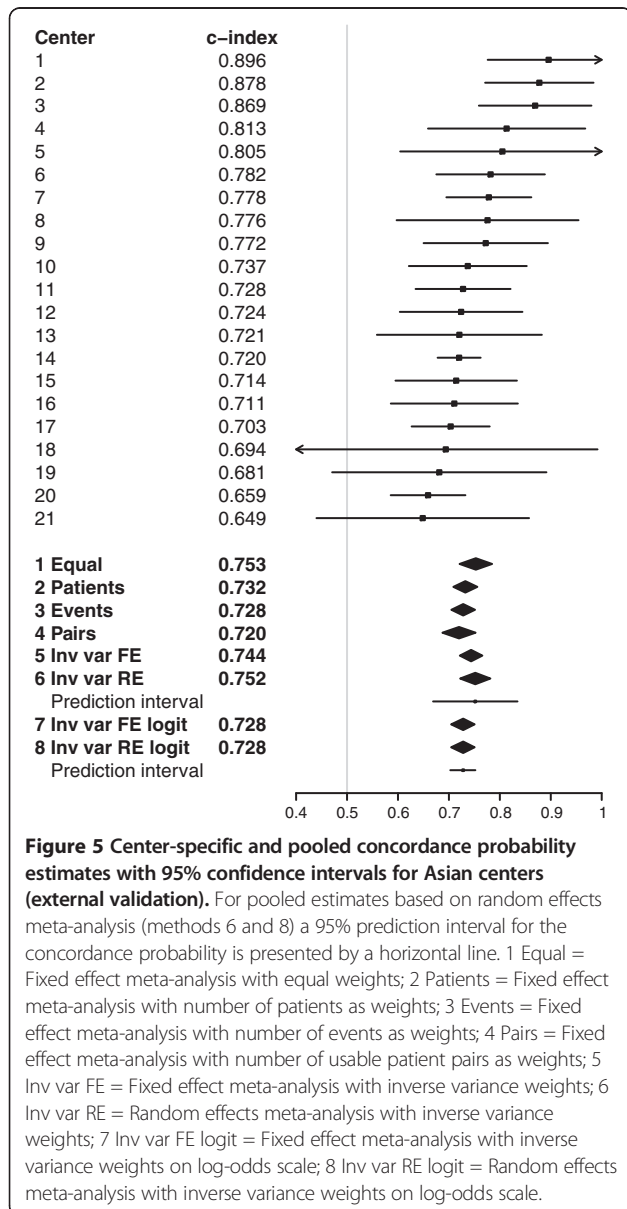
The authors express their gratitude to all of the principal investigators of the CRASH trial for providing the data. We thank Prof. Emmanuel Lesaffre (Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands) for helpful comments.

Funding

This work was supported by the Netherlands Organisation for Scientific Research (grant 917.11.383).

Author details

¹Department of Public Health, Erasmus MC, Dr. Molewaterplein 50, Rotterdam 3015 GE, The Netherlands. ²Department of Population Health,



London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.

Received: 7 October 2013 Accepted: 8 January 2014
Published: 15 January 2014

References

1. Steyerberg EW, Vickers AJ, Cook NR, Gerdts T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW: **Assessing the performance of prediction models: a framework for traditional and novel measures.** *Epidemiology* 2010, **21**(1):128–138.
2. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA: **Evaluating the yield of medical tests.** *JAMA* 1982, **247**(18):2543–2546.
3. Pencina MJ, D'Agostino RB: **Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation.** *Stat Med* 2004, **23**(13):2109–2123.
4. Bouwmeester W, Twisk JW, Kappen TH, van Klei WA, Moons KG, Vergouwe Y: **Prediction models for clustered data: comparison of a random intercept and standard regression model.** *BMC Med Res Methodol* 2013, **13**:19.
5. Gelman A, Hill J: *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press; 2007.
6. Duchateau L, Janssen P: *The Frailty Model.* New York: Springer; 2008.
7. Van Oirbeek R, Lesaffre E: **An application of Harrell's C-index to PH frailty models.** *Stat Med* 2010, **29**(30):3160–3171.
8. Van Oirbeek R, Lesaffre E: **Assessing the predictive ability of a multilevel binary regression model.** *Comput Stat Data Anal* 2012, **56**(6):1966–1980.
9. Collaborators MCT, Perel P, Arango M, Clayton T, Edwards P, Komolafe E, Pocock S, Roberts I, Shakur H, Steyerberg E, et al: **Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients.** *BMJ* 2008, **336**(7641):425–429.
10. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JD, et al: **Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics.** *PLoS Med* 2008, **5**(8):e165.
11. Edwards P, Arango M, Balica L, Cottingham R, El-Sayed H, Farrell B, Fernandes J, Gogichavili T, Golden N, Hartzenberg B, et al: **Final results of MRC CRASH, a randomised placebo-controlled trial of intravenous corticosteroid in adults with head injury-outcomes at 6 months.** *Lancet* 2005, **365**(9475):1957–1959.
12. Teasdale G, Jennett B: **Assessment of coma and impaired consciousness. A practical scale.** *Lancet* 1974, **2**(7872):81–84.
13. Jennett B, Bond M: **Assessment of outcome after severe brain damage.** *Lancet* 1975, **1**(7905):480–484.
14. R Development Core Team: *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2011. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
15. Therneau T, original Splus->R port by Lumley T: **survival: Survival analysis, including penalised likelihood. R package version 2.36-9.** 2011. <http://CRAN.R-project.org/package=survival>.
16. Quade D: *Nonparametric partial correlation. Volume No. 526, Volume 526.* North Carolina: Institute of Statistics Mimeo; 1967.
17. Higgins JP, Thompson SG: **Quantifying heterogeneity in a meta-analysis.** *Stat Med* 2002, **21**(11):1539–1558.
18. DerSimonian R, Laird N: **Meta-analysis in clinical trials.** *Control Clin Trials* 1986, **7**(3):177–188.
19. DerSimonian R, Kacker R: **Random-effects model for meta-analysis of clinical trials: an update.** *Contemp Clin Trials* 2007, **28**(2):105–114.
20. Higgins JPT, Thompson SG, Spiegelhalter DJ: **A re-evaluation of random-effects meta-analysis.** *J R Soc Health Series A* 2009, **172**(1):137–159.
21. Riley RD, Higgins JPT, Deeks JJ: **Interpretation of random effects meta-analyses.** *BMJ* 2011, **342**:d549.
22. Hardy RJ, Thompson SG: **Detecting and describing heterogeneity in meta-analysis.** *Stat Med* 1998, **17**(8):841–856.
23. Lumley T: **rmeta: Meta-analysis. R package version 2.16.** 2009. <http://CRAN.R-project.org/package=rmeta>.
24. Vergouwe Y, Moons KG, Steyerberg EW: **External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients.** *Am J Epidemiol* 2010, **172**(8):971–980.
25. Steyerberg EW: *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* New York: Springer; 2009.
26. Gönen M, Heller G: **Concordance probability and discriminatory power in proportional hazards regression.** *Biometrika* 2005, **92**(4):965–970.
27. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ: **On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data.** *Stat Med* 2011, **30**(10):1105–1117.
28. Hanley JA, McNeil BJ: **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** *Radiology* 1982, **143**(1):29–36.
29. DeLong ER, DeLong DM, Clarke-Pearson DL: **Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.** *Biometrics* 1988, **44**(3):837–845.
30. Nam BH, D'Agostino RB: **Discrimination Index, the Area under the ROC Curve.** In *Goodness-of-Fit Tests and Model Validity.* Boston: Birkhauser; 2002:267–279.
31. Efron B, Tibshirani R: *An Introduction to the Bootstrap.* Boca Raton, FL: CRC press; 1993.

doi:10.1186/1471-2288-14-5

Cite this article as: van Klaveren et al.: **Assessing discriminative ability of risk models in clustered data.** *BMC Medical Research Methodology* 2014 **14**:5.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

