

RESEARCH

Open Access



A systematic review of neonatal treatment intensity scores and their potential application in low-resource setting hospitals for predicting mortality, morbidity and estimating resource use

Jalemba Aluvaala^{1,2,3*} , Gary S. Collins⁵, Michuki Maina¹, James A. Berkley^{1,3,4} and Mike English^{1,3}

Abstract

Background: Treatment intensity scores can predict mortality and estimate resource use. They may therefore be of interest for essential neonatal care in low resource settings where neonatal mortality remains high. We sought to systematically review neonatal treatment intensity scores to (1) assess the level of evidence on predictive performance in predicting clinical outcomes and estimating resource utilisation and (2) assess the applicability of the identified models to decision making for neonatal care in low resource settings.

Methods: We conducted a systematic search of PubMed, EMBASE (OVID), CINAHL, Global Health Library (Global index, WHO) and Google Scholar to identify studies published up until 21 December 2016. Included were all articles that used treatments as predictors in neonatal models. Individual studies were appraised using the CHECKlist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS). In addition, Grading of Recommendations Assessment, Development, and Evaluation (GRADE) was used as a guiding framework to assess certainty in the evidence for predicting outcomes across studies.

Results: Three thousand two hundred forty-nine articles were screened, of which ten articles were included in the review. All of the studies were conducted in neonatal intensive care units with sample sizes ranging from 22 to 9978, with a median of 163. Two articles reported model development, while eight reported external application of existing models to new populations. Meta-analysis was not possible due to heterogeneity in the conduct and reporting of the identified studies. Discrimination as assessed by area under receiver operating characteristic curve was reported for in-hospital mortality, median 0.84 (range 0.75–0.96, three studies), early adverse outcome and late adverse outcome (0.78 and 0.59, respectively, one study).

(Continued on next page)

* Correspondence: jaluvaala@kemri-wellcome.org

¹KEMRI-Wellcome Trust Research Programme, P.O. Box 43640 – 00100, Nairobi, Kenya

²Department of Paediatrics and Child Health, College of Health Sciences, University of Nairobi, Kenyatta National Hospital, P. O. Box 19676-00202, Nairobi, Kenya

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



© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

(Continued from previous page)

Conclusion: Existing neonatal treatment intensity models show promise in predicting mortality and morbidity. There is however low certainty in the evidence on their performance in essential neonatal care in low resource settings as all studies had methodological limitations and were conducted in intensive care. The approach may however be developed further for low resource settings like Kenya because treatment data may be easier to obtain compared to measures of physiological status.

Systematic review registration: PROSPERO CRD42016034205

Keywords: Neonatal prognosis, Treatment intensity, Prediction model, CHARMS

Background

Improving neonatal care is now a global concern, and tools to examine system performance and guide service planning in low and middle income countries (LMIC) are needed. Higher quality of care across a broader range of settings is necessary if LMICs are to realise the substantial reduction in neonatal mortality expected from the delivery of essential interventions at scale in facilities [1, 2]. Clinical prediction models are typically used to support shared decision making at individual patient level, for risk stratification in clinical trials or for case-mix adjustment in quality of care assessments [3, 4]. They have also been developed to estimate resource use and thereby inform service delivery planning [3, 5]. By facilitating better decision making, prediction models could contribute to the improvement of the quality of hospital care for neonates in LMICs ultimately improving neonatal survival.

In considering the use of prediction models to support decision making in hospital-based essential neonatal care in LMICs, we may either develop a new model or choose from amongst existing models. Collins and colleagues recommend the latter approach before developing new models to avoid waste of resources [6, 7]. Selecting from amongst existing models should however be guided by the evaluation of existing models for performance and suitability to the context [8]. Prediction models have variably been termed as prediction rules, probability assessments, prediction models, decision rules and risk scores [9]. Existing models that predict in-hospital mortality were identified from published reviews of neonatal models and summarised in Additional file 1 [10–18]. This overview revealed that the neonatal therapeutic intervention scoring system (NTISS) is unique amongst neonatal prediction models as it uses treatments rather than clinical and pathophysiological factors as predictors [10]. The NTISS predicts in-hospital mortality and morbidity in addition to estimating resource utilisation, particularly nursing workload [19]. The latter may help identify service delivery bottlenecks in essential neonatal care providing information to guide strategic planning [20].

Existing neonatal prediction models have typically been developed for settings offering advanced neonatal intensive care including mechanical ventilation and other expensive interventions. These are not directly applicable to settings offering only essential neonatal care where respiratory support is limited to oxygen via nasal cannula without monitoring such as pulse oximetry [21, 22]. Amongst the models included in the published reviews, only one was developed specifically for a low-resource setting, the simplified age-weight-sex (SAWS) [23]. This was, however, not considered further as it was developed for very low birth weight neonates only. We therefore conducted a systematic review to systematically identify and characterise prediction model research that has used treatments as predictors (treatment intensity models) in neonatal care specifically to (1) assess the certainty of evidence on predictive performance in predicting clinical outcomes (primarily in-hospital mortality) and estimating resource utilisation and (2) assess the applicability of the identified models to neonatal care in LMIC.

Methods

The systematic review was conducted following the recently published guidance from the Cochrane Prognosis Methods group; the CHECKlist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) [24]. In addition, the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) was used as a guiding framework to assess the quality of the retrieved articles [25]. Reporting of the review was done using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) recommendations (Additional file 2) [26]. Details of the protocol for this systematic review were registered on the international prospective register of systematic reviews (PROSPERO): Jalemba Aluvaala, Gary Collins, Michuki Maina, James Berkley, Mike English. Neonatal treatment intensity scores and their potential application for low resource settings: a systematic

review. PROSPERO 2016: CRD42016034205. Available from: http://www.crd.york.ac.uk/PROSPERO/display_record.asp?ID=CRD42016034205.

Registration of the protocol was done after the initial screening of articles.

Search strategy

The primary electronic database used was PubMed with supplementary searches conducted in EMBASE (OVID), CINAHL, Global Health Library (Global index, WHO) and Google Scholar. The last search conducted was on 21 December 2016. Bibliographies of identified papers were also hand-searched for additional papers.

The following key search terms were used “neonate” or “neonatal or newborn”, “treatment or therapeutic” or “therapy”, “intensity” and “score or scoring”. In PubMed, the search strategy was implemented using medical subject headlines (MeSH) where applicable and the appropriate Boolean terms, (((“Neonatology”[Mesh] OR “Infant, Newborn”[Mesh]) AND (“Therapeutics”[Mesh] OR “therapy”[Subheading])) AND intensity) AND scor*. A similar approach was used in all the other electronic databases. The search strategy was developed with input from an expert medical librarian. No language restrictions were applied on the selection of the articles.

The primary search strategy was augmented by substituting a validated search string for prediction models in PubMed for the scor* term used in the first search string; ((Validat* OR Predict* OR Rule*) OR (Predict* AND (Outcome* OR Risk* OR Model*)) OR ((History OR Variable* OR Criteria OR Scor* OR Characteristic* OR Finding* OR Factor*) AND (Predict* OR Model* OR Decision* OR Identif* OR Prognos*)) OR (Decision* AND (Model* OR Clinical* OR Logistic Models/)) OR (Prognostic AND (History OR Variable* OR Criteria OR

Scor* OR Characteristic* OR Finding* OR Factor* OR Model*)) [27].

Screening process

All articles identified by the search were initially screened for eligibility on title independently by two reviewers (JA, ME) with disagreements resolved by discussion. The search results were exported to the reference management software EndNote X7 (Thomas Reuters, Philadelphia, USA). Duplicate articles were removed and the remaining titles and abstracts screened. Full-text articles were retrieved and assessed for eligibility using predefined criteria (Table 1) for inclusion in the review. The target population was neonates defined as babies aged 0–28 days (Table 1).

Data extraction and critical appraisal of individual studies

Two independent reviewers (JA, MM) extracted data using a standardised form based on the CHARMS checklist, and any disagreements were resolved by discussion. Data elements extracted were study design, participants, geographical location, outcomes predicted, description of model development (type of model, e.g. logistic regression), number and type of predictors, number of study participants and number of outcome events, handling of missing data and model performance (calibration, discrimination). Critical appraisal of individual studies was by applying the CHARMS guidance on each of the elements of data extracted to assess potential limitations [24].

Descriptive analyses

A quantitative meta-analysis was not conducted due to heterogeneity in the conduct and reporting of the

Table 1 Eligibility criteria for inclusion in the review

Criteria	Inclusion	Exclusion
Model type	Prognostic models	Diagnostic models*
Intended scope of the review	Inform decision making at individual level (e.g. using risk of in-hospital mortality) and planning service delivery (e.g. nursing staffing)	
Types of modelling studies	Development and/or validation	
Target population	Neonates† admitted to a neonatal unit in any country	Studies limited to neonates with congenital anomalies older children or adults
Predictors	Any use of treatments or interventions as predictors	Non-therapeutic intervention, e.g. radiological imaging intensity Treatment intensity not measured by enumeration of therapeutic interventions
Outcomes	Any outcome	
Time of prediction	No restriction	
Intended moment of use	No restriction	

*Diagnostic models estimate probability that a particular disease is currently present in an individual in contrast to prognostic models that estimate probability of future events

†Neonate defined as a baby aged 0–28 days and all the articles included adhered to this definition

identified studies. Results were therefore summarised descriptively and synthesised using a narrative approach.

Certainty of the evidence across studies

In the absence of a specific tool for assessing risk of bias across studies for clinical prediction models, the GRADE approach was used as a guiding framework for this purpose [28]. GRADE assesses the certainty of the evidence in the estimates of effects for given outcomes across studies. In general, this is achieved using explicit criteria including study design, risk of bias, imprecision, inconsistency, indirectness and magnitude of effect [28]. For this review, the GRADE approach for diagnostic studies was used as a guide (Additional file 3) [29–31]. Certainty in the evidence for each outcome across the identified articles was initially rated as high quality if the studies were prospective cohort as recommended by CHARMS [24]. Subsequently, certainty was downgraded if there were serious limitations in the conduct of the studies as defined by CHARMS and if there was inconsistency, indirectness, imprecision and publication bias as defined by GRADE (Additional file 4). Certainty of the evidence

on predictive performance was thus rated as high, moderate, low or very low for each outcome.

Results

Study selection

A total of 3249 unique articles were identified by the search strategy, of which 3229 were excluded based on the title and abstract. The full texts of 20 articles were screened, of which 10 articles met the inclusion criteria and were included in this review (Fig. 1). Articles were excluded for the following reasons: a population other than neonates was studied, intensity referred to a non-therapeutic intervention (e.g. radiological imaging intensity), treatment intensity was measured by means other than enumeration of therapeutic interventions provided (e.g. proportion of days that hospital intensive care was required) and studies limited to neonates with congenital anomalies considered to be lethal.

Study characteristics

Table 2 provides information on the general characteristics of the included studies. All of the studies were conducted in tertiary neonatal intensive care units (NICU)

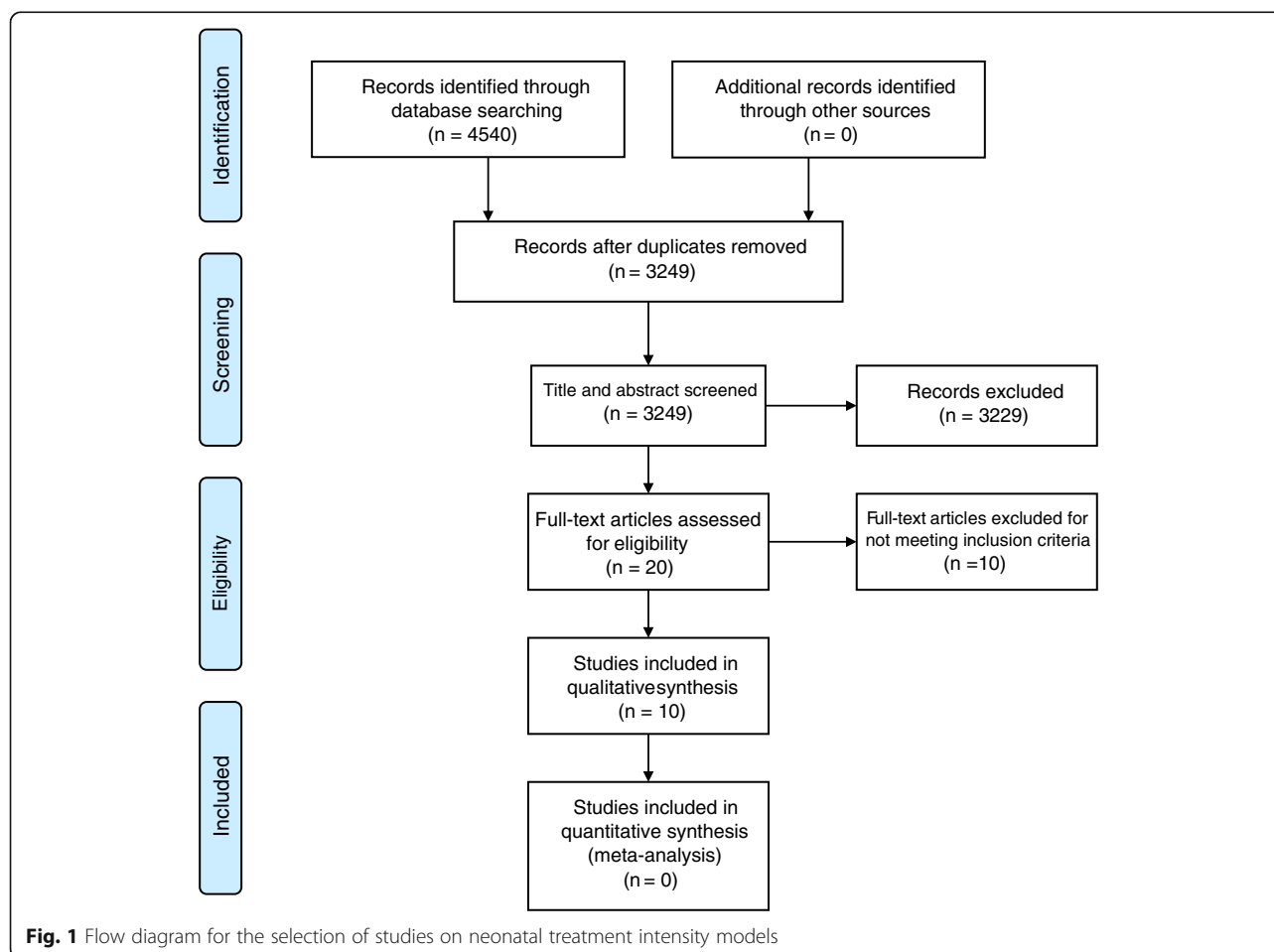


Table 2 Characteristics of studies included in the review

Study	Study dates	Objective*	Setting	Sample size	In-hospital mortality
Georgieff, 1989 [40]	1987	Validation (TISS) [†]	1 NICU [‡] , USA	55	0
Gray, 1992 [19]	1989–1990	Development (NTISS) [§]	3 NICU [‡] , USA	1768	114
Davies, 1995 [33]	Not reported	Validation (NTISS) [§]	1 NICU [‡] , South Africa	50	8
Eriksson, 2002 [34]	1991–1995	Validation (NTISS) [§]	2 NICU [‡] , Sweden	240	39
Zupancic, 2002 [35]	1998 & 1999	Validation (NTISS) [§]	1 NICU [‡] , USA	154	Not reported
Mendes, 2006 [35]	2004	Validation (NTISS) [§]	2 NICU [‡] , Brazil	96	9
Rojas, 2011 [37]	2007	Validation (NTISS) [§]	1 NICU [‡] , 1 intermediate unit, Colombia	22	Not reported
Oygur, 2012 [38]	2006–2010	Validation (NTISS) [§]	1 NICU [‡] , Turkey	364	103
Shah, 2015 [32]	2010–2012	Development (unnamed)	23 NICU [‡] s, Canada	9978	650
Wu, 2015 [39]	2007–2011	Validation (NTISS) [§]	1 NICU [‡] , Taiwan	172	18

*Study objective, model development (creation of a new model) or validation (application of an existing score/model to an external population)

[†]Neonatal intensive care unit

[‡]Therapeutic intervention scoring system

[§]Neonatal therapeutic intervention scoring system

^{||}Primary outcome for the review. Data was however extracted on all outcomes reported by the authors

with five single centre and five multicentre studies and included 12,899 neonates. Sample sizes ranged from 22 to 9978 with a median of 163. Two out of the ten articles reported model development [19, 32]. Eight were reports of external application of existing models to new populations. Gray and colleagues developed the NTISS while Shah et al. developed a score using selected variables derived from the NTISS [19, 32]. Seven of the articles reporting external application to new populations used the NTISS [33–39]. One study used a therapeutic score originally developed for adults, the therapeutic intervention scoring system (TISS) [40].

Critical appraisal of study design and statistical analysis of individual studies

All studies

All of the included studies were individually critically appraised based on the limitations in the design of the study (Table 3) and limitations in the statistical analysis (Table 4). With regard to study design, four were prospective studies, two combined prospective and retrospective data, while four were retrospective (Table 3). With regard to the selection of participants, the majority of the studies (8/10) recruited all eligible infants and therefore were deemed to have no selection bias. None of the articles explicitly reported blinding during collection of data on predictors and outcomes (Table 3).

However, this does not constitute a risk of bias for mortality in all instances and where a prospective study design was used for all outcomes. Missing data may also give rise to bias during analysis, and only one study was judged to have no risk of bias due to missingness [32].

Score development studies

Two studies described the development of a treatment intensity score [19, 32]. The NTISS was developed to serve as a “therapy-based severity-of-illness tool for use in intensive care” by modification of the adult TISS [19, 41]. Included in the NTISS were 63 therapeutic interventions delivered in neonatal intensive care, e.g. surfactant and mechanical ventilation [19]. Shah and colleagues on the other hand used 24 NICU therapeutic interventions (many that are included in the NTISS) to develop a score to measure intensity of NICU resource use [32]. In both of these studies, treatment predictor selection and their respective predictor weights were determined by consensus amongst experts. With respect to prediction of mortality, there were a total number of 114 deaths (events) in Gray et al. and 650 deaths in Shah et al. (Table 3) [19, 32]. From a predictive modelling perspective, it is recommended that there should be at least ten outcome events (deaths in this case) for each predictor variable in the regression model [24]. It was however not

Table 3 Limitations in individual studies with respect to study design and data collection

Study	Study type*	Participants [†]	Outcome(s) [‡]	Predictors [§]	Sample size	Missing data [¶]
Georgieff,1989 [40]	-	-	+	+	+	?
Gray,1992 [19]	-	-	+	+	?	+
Davies,1995 [33]	+	+	+	+	+	?
Eriksson,2002 [34]	-	-	+	+	+	+
Zupancic,2002 [35]	-	+	?	+	+	?
Mendes,2006 [36]	-	-	+	+	+	+
Rojas,2011 [37]	-	-	+	+	+	?
Oygur,2012 [38]	+	-	+	+	+	-
Shah,2015 [32]	+	-	+	+	?	+
Wu,2015 [39]	+	-	+	+	+	?

+ Limitation present

- No limitation

? Not reported therefore unclear risk of bias

Limitation present if (based on CHARMS criteria):

* Data collection not prospective

† Not all eligible neonates recruited resulting in risk of selection bias

‡ Risk of measurement error in determining outcome status

§ Risk of measurement error in determining predictor status

|| Sample size less than the recommended

¶ Missing data causing risk of bias

possible to assess for risk of overfitting on the basis of the sample size since none of the two studies published the final regression models (Table 4), and therefore, it is not clear what the number of events per predictor (EPV) were.

Neither of the two score development studies specified a regression model that included the therapeutic interventions as individual predictors. Rather, the therapeutic interventions were assigned sub-scores by an expert

panel which were then summated to give a total score for each patient. The statistical relationship between the total scores and the outcomes was then examined. Nonetheless, Shah et al. reported model discrimination with a c-statistic of a regression model (Table 4) [32]. The study describing the development of the NTISS did not report model discrimination but assessed calibration using the Hosmer-Lemeshow (H-L) technique (Table 4) [19]. Finally, neither of these two studies presented a

Table 4 Limitations in individual studies with respect to statistical analysis

Study	Study type	Performance [†]	Validation [‡]	Presentation [§]
Georgieff,1989 [40]	Evaluation of existing score*	?	?	?
Gray, 1992 [19]	Development of new score	+	?	?
Davies, 1995 [33]	Evaluation of existing score*	?	?	?
Eriksson, 2002 [34]	Evaluation of existing score*	+	?	?
Zupancic, 2002 [35]	Evaluation of existing score*	?	?	?
Mendes, 2006 [36]	Evaluation of existing score*	?	?	?
Rojas, 2011 [37]	Evaluation of existing score*	?	?	?
Oygur,2012 [38]	Evaluation of existing score*	+	?	?
Shah et al., 2015 [32]	Development of new score	+	?	?
Wu, 2015 [39]	Evaluation of existing score*	+	?	?

+ Limitation present (based on CHARMS criteria)

- No limitation

? Not reported

* None of these applied a regression formula from the original score development study to the new population but instead specified new models thus were model re-development rather than external validation studies

† Limitation present if (based on CHARMS criteria) either score discrimination OR calibration only was reported

‡ Internal validation (to quantify model overfitting) OR external validation (model performance in new population)

§ Presentation of final model as either a regression formula or a score chart

final model or score chart to demonstrate how to determine predicted probabilities of the outcomes for individual patients.

Application of existing scores to external populations

Eight studies described the evaluation of existing therapeutic intensity models in a different geographical setting but still in neonatal intensive care. Seven studies evaluated the NTISS [33–39]. A single study evaluated a version of the adult TISS in neonates [40]. For external validation studies using regression approaches, recommended sample sizes of between 100 and 250 outcome events have been suggested [42–44]. With respect to the primary outcome, (in-hospital mortality) only three studies had at least 100 outcome events (Table 2) [32, 38, 45].

Zupancic et al. used individual treatments as separate parameters in a linear regression model with personnel time as outcome [35]. Oygur et al. also had individual treatments as separate parameters but did not specify the type of model used to predict in-hospital mortality [38]. In contrast, the other six studies summated the treatments to give a single score which was then used in subsequent analyses as a single parameter [33, 34, 36, 37, 39, 40]. This is the same approach used by Gray et al. in developing the NTISS [19]. In terms of model performance, Eriksson et al., Oygur et al. and Wu et al. computed discrimination using area under the receiver operating characteristic curve (AUROC) but did not report model calibration [34, 38, 39]. None of these eight studies applied the exact model formula as obtained from the original score development work but instead conducted fresh analyses; these were therefore all in

effect model re-development studies in a new population rather than external validation [24, 46, 47].

Synthesis of results of estimated performance across studies

The primary outcomes of interest were mortality and nursing workload. However, for the narrative synthesis of the results, we determined that outcomes were reported in three main categories: mortality, morbidity and resource utilisation (including nursing workload) across the ten articles. There are two distinct aspects of performance for statistical models, predictive (measures include discrimination and calibration) or explanatory (testing causal relationships) [48, 49]. However, the two approaches are often conflated creating ambiguity, and this distinction is therefore made in the narrative synthesis [49].

Performance with respect to mortality

Five studies reported on in-hospital mortality as an outcome (Table 5, Additional file 5). No other mortality outcome measure was reported. Three of these computed discriminatory performance using AUROC analysis [34, 38, 39]. Calibration was reported in only one article [19]. Explanatory rather than predictive performance for in-hospital mortality was reported in three articles [19, 33, 39].

In Eriksson et al., AUROC for in-hospital mortality (39 deaths) was 0.82(SE 0.04) [34]. Oygur et al. reported performance with 63 treatment variables and with sensitive treatments only (based on significant association with in-hospital mortality by chi-square test). The AUROC

Table 5 Summary of certainty of evidence in predicting outcome and resource use using GRADE

Outcome	No. of studies	Factors that may decrease certainty of evidence*					Overall† certainty	Importance ‡
		Limitations	Indirectness	Inconsistency	Imprecision	Reporting bias		
Mortality	5 studies (n = 2594)	Serious	Serious	None	None	Unlikely	⊕⊕⊕⊕ low	Critical
Morbidity	2 studies (n = 295)	Serious	Serious	None	None	Unlikely	⊕⊕⊕⊕ low	Critical
Composite (Morbidity and Mortality)	2 studies (n = 10,218)	Serious	Serious	None	None	Unlikely	⊕⊕⊕⊕ low	Important
Nursing workload	3 studies (n = 317)	Serious	Serious	None	None	Unlikely	⊕⊕⊕⊕ low	Critical
Hospital Costs	1 study (n = 1768)	Serious	Serious	None	None	Unlikely	⊕⊕⊕⊕ low	Important
Length of stay	1 study (n = 1768)	Serious	Serious	None	None	Unlikely	⊕⊕⊕⊕ low	Important
Time inputs	1 study (n = 154)	Serious	Serious	None	None	Unlikely	⊕⊕⊕⊕ low	Critical
Comparison of resource use	1 study (n = 96)	Serious	Serious	None	None	Unlikely	⊕⊕⊕⊕ low	Important
Rehabilitation	1 study (n = 240)	Serious	Serious	None	None	Unlikely	⊕⊕⊕⊕ low	Not important

* From GRADE, Grading of Recommendations Applicability, Development and Evaluation

†Certainty rating scale; high (⊕⊕⊕⊕), moderate, low, very low (⊕⊕⊕⊕)

‡Importance of outcomes (from GRADE)

for all treatment variables (63 treatments) by birth weight categories were 500–1499 g (103 deaths), 0.851 (95% CI 0.809–0.885); 1000–1499 g (33 deaths), 0.834 (95% CI 0.781–0.878); and 500–999 g (70 deaths), 0.749 (95% CI 0.662–0.822). Using Student's *t* test to compare AUROCs for all variables (63) versus sensitive variables only (18 treatments) resulted in a significant change for the 500–999 g (70 deaths) category only, 0.749 vs 0.823 ($P = 0.02$) [38].

Performance of NTISS 24 h after admission was examined by Wu et al. who compared serial scores in preterm infants weighing < 1500 g (18 deaths) and found AUROC at 24 h, 0.913; 48 h, 0.955; and 72 h, 0.958. Confidence intervals for the AUROC were not provided, but using Tukey's honest significant difference (HSD) test, there was significant difference between the 24 h average score when compared to both 48 and 72 h average scores ($p < 0.001$). There was no difference between the 48 and 72 h scores ($p = 0.391$) [39].

With regard to explanatory performance, Gray et al. found that NTISS correlated with in-hospital mortality [19]. Davies et al. found no difference between predicted mortality with actual mortality [33], while Wu and colleagues found that average NTISS scores at 24, 48, and 72 after admission were higher in the mortality group than in those who survived [39].

Performance with respect to morbidity

Predictive performance for morbidity was reported by Eriksson et al. only; here NTISS poorly predicted morbidity at 4 years of age (growth retardation, neurodevelopmental impairment, pulmonary problems) with AUROC of 0.59 (SE 0.05) but calibration was not reported (Table 5, Additional file 5) [34]. Explanatory performance was reported by Georgieff and colleagues who found a linear relationship between severity of physiologic instability and the TISS. In addition, the mean TISS was significantly higher in infants discharged after 14 days and those with severe lung disease (hyaline membrane disease (HMD)) [40].

Performance with respect to composite of mortality and morbidity

Predictive performance for the composite outcome of morbidity and mortality was reported in two articles (Table 5, Additional file 5). In the first, Eriksson et al. found that for early adverse outcome (all-cause mortality or oxygen dependence, intraventricular haemorrhage/periventricular leukomalacia or retinopathy of prematurity), the AUROC curve for NTISS was 0.78 (SE 0.03) [34]. Shah et al. reported a *c*-statistic of 0.86 for a composite outcome (all-cause mortality, \geq grade 3 IVH, PVL, stage \geq 3 ROP,

oxygen dependence or stage \geq necrotizing enterocolitis) from a logistic regression model that included therapeutic intensity, NICU size, NICU occupancy rate, gestational age, small for gestational age, multiple births, outborn, caesarean delivery, SNAP II score > 20 and mechanical ventilation).

Performance with respect to estimating resource utilisation

Four studies estimated the extent to which resource use (measured in different ways) was explained by treatment intensity (Table 5, Additional file 5). There was statistical association between nursing workload and both TISS and NTISS [19, 35, 37, 40]. In addition, the NTISS correlated with hospital costs and length of stay [19].

Critical appraisal of certainty of the evidence in model performance across all studies

GRADE criteria (Additional file 3 and Additional file 4) were used to assess the certainty in the evidence of predictive performance for each outcome *across all* studies where it was reported and tabulated (Table 5). The outcomes were reported in three main categories: mortality, morbidity and resource utilisation (nursing workload, length of stay, time inputs, comparison of resource use). Table 5 shows that there is low certainty in the evidence of predictive performance for these three categories of outcomes for low-resource settings as all the studies had serious limitations in their conduct (as determined by CHARMS) in addition to indirectness as they were all conducted in high-resource settings (GRADE).

Discussion

Performance of neonatal treatment intensity models in model development studies

Existing neonatal treatment intensity models can predict mortality and morbidity. Discriminatory performance as measured by the area under the receiver operating curve is poorest for long-term morbidity (0.59) and highest for in-hospital mortality in infants weighing 1000–1499 g for NTISS measured at 72 h post admission (0.958) [34, 39]. Calibration was reported by Gray et al. only who found a close agreement between observed and predicted in-hospital mortality by the Hosmer-Lemeshow test [19]. Using a variety of tests of statistical association rather than predictive performance, treatment intensity was found also to be associated with mortality, morbidity and resource utilisation [19, 35, 37, 39, 40].

Discriminatory performance of neonatal treatment intensity models for predicting in-hospital mortality measured by AUROC ranges from 0.749 to 0.958 [38, 39]. The recommended threshold for good discrimination is 0.8 [11]. The reported performance therefore suggests that the treatment scoring approach may perform well in distinguishing between neonates who die and those

who do not in neonatal intensive care. This discriminatory ability is comparable to that of more commonly used neonatal predictive models. Discriminatory performance at model development in the prediction of in-hospital mortality measured by AUROC ranges from 0.87 (SE 0.33) for the Transport Risk Index for Physiological Stability version 2; TRIPS-II [50] to 0.92 (SE 0.01) for the Clinical Risk Index for Babies version 2; CRIB II [51]. Despite the difference in predictors, there is thus consistency in the evidence that clinical prediction modelling approaches may be useful in predicting in-hospital mortality in neonatal medicine. The application of the treatment intensity scoring approach in adult and paediatric intensive care preceded use in the neonatal population. However, none of these adult or paediatric studies reported the discriminatory performance of the scores [41, 52–55].

With regard to calibration, Gray et al. reported good calibration in predicting for in-hospital mortality by the NTISS using the H-L test. This means that the predicted deaths closely matched the observed deaths. Similarly, good calibration in predicting in-hospital mortality as measured by the H-L statistic has been reported for the physiologic scores Score for Neonatal Acute Physiology – Perinatal Extension version II (SNAP-PE II) [56], CRIB II [51] and TRIPS II [50]. These results must however be interpreted with caution given the limitations of the H-L test [57, 58]. None of the included studies reported calibration using the preferred method, calibration plots [57].

There is however low certainty in this evidence of performance with reference to essential neonatal care in low resource settings for two reasons. One, there were serious limitations identified in the design and conduct of these studies such that even for high-income countries, the certainty in score performance would be low as well. Secondly, these results are not directly generalisable to essential neonatal care in low-resource settings as the studies were all conducted in neonatal intensive care in high-income settings given the differences in range of treatments, staffing and case-mix.

Whereas the focus of this work was to investigate predictive performance of these scores, the retrieved articles also reported on explanatory performance. Treatment intensity was found to be associated with mortality, morbidity and resource utilisation [19, 35, 37, 39, 40]. This is consistent with findings in adult and paediatric intensive care for mortality [41, 54]. Similarly, treatment intensity was also found to be associated with nursing workload and could be used to determine the workload a typical nurse is capable of per shift in adult intensive care [41, 55]. Finally, treatment intensity also correlates with costs in adult intensive care [41]. The frequent reporting of such analyses alongside predictive analyses underscores

the importance of clearly distinguishing predictive from explanatory aspects of models [49, 59].

Generalisability to essential neonatal care in low-resource settings

The second objective of this review was concerned with generalisability of the identified treatment intensity scores to essential neonatal care in low-resource settings. Generalisability is assessed by evaluating the performance of previously developed models in a new population, i.e. external validation [58]. There were no studies of the treatment intensity models in low-resource settings identified. The eight studies which applied previously developed treatment intensity models to new populations were potentially external validation studies [33–40]. However, none of these applied model coefficients derived from the derivation studies to the new populations to compute the predicted probabilities of the outcomes to allow comparison with the observed outcomes. None of the studies therefore qualified as an external validation.

The use of treatments as predictors presents three potential limitations. To begin with, the treatments prescribed are dependent on the availability of resources and may not necessarily be an accurate reflection of the patient's actual requirements. Secondly, there may be a variation in clinical practice which will be reflected in differences in treatments and thus potentially in the model performance. Thirdly, the treatment predictors used as predictors in the identified studies reflect the neonatal intensive care study sites where mechanical ventilation and exogenous surfactant amongst other expensive and invasive interventions are available. These cannot be applied directly to populations receiving essential neonatal care in low-resource settings in LMICs, for example with respiratory support limited to oxygen via nasal cannula. However, the treatment intensity approach has potential in the Kenyan context since there is a relatively high degree of standardisation of essential neonatal care in the first referral level facilities which mitigates these limitations to some extent [60].

Limitations of the review

The systematic review was guided by the CHARMS recommendations which provide a clear guide on extracting data and identifying limitations in individual studies. However, it does not provide guidance on how to assess the quality of evidence on individual outcomes across all the identified studies. As a result, the GRADE approach was used as a guiding framework for this purpose. While there is no GRADE guideline specifically for clinical prediction models,

GRADE for diagnostic models was judged sufficiently similar to and used for this review. A key strength of the GRADE approach is that the judgements on the quality of evidence are made on explicit criteria. Nonetheless, the PROBAST (prediction study risk of bias assessment tool) tool which is specific for this purpose is under development [61].

Conclusion

Existing neonatal treatment intensity scores show promise in predicting mortality and morbidity, but there is low certainty in the evidence on their performance in essential neonatal care in low resource settings in LMICs. The limitations in the included studies mirror those reported in other systematic reviews including unclear study designs, failure to report follow-up times, sample size calculations and performance measures [5, 62–66]. However, over 12,000 patients were included in this review of ten studies compared to 467 patients in the only model developed to date for neonates in low-income countries (SAW) [23]. In addition, increase in treatment intensity is associated with higher nursing workload, hospital costs and length of stay. The approach may therefore be usefully developed further for low-resource settings, e.g. Kenya because treatment data may be easier to obtain compared to other parameters like measures of physiological status [67]. This will entail addressing the limitations identified by applying appropriate methods for model development, validation and most importantly the evaluation of clinical utility and impact [48, 57, 68–70]. Prognostic information obtained in this way can support decision making in the planning and organisation of quality essential neonatal care.

Additional files

Additional file 1: Summary of prognostic models for predicting in-hospital neonatal mortality. Comparison of predictors and outcomes of neonatal prognostic models included in published reviews. (DOCX 15 kb)

Additional file 2: PRISMA checklist. Completed PRISMA checklist. (DOC 63 kb)

Additional file 3: Application of GRADE as a guide in assessing certainty of evidence in prediction modelling studies. GRADE* criteria that were used as a guiding framework in assessing the certainty in the evidence for outcomes across the identified studies. *Grading of Recommendations Assessment, Development and Evaluation. (DOCX 14 kb)

Additional file 4: Using GRADE* as a guiding framework in rating of certainty of evidence in predictive model predictive performance. Description of the four categories in GRADE for rating certainty of evidence. *Grading of Recommendations Assessment, Development and Evaluation. (DOCX 13 kb)

Additional file 5: Summary of results from individual studies. Summary of data extracted from each eligible article. (DOCX 17 kb)

Abbreviations

AUROC: Area under the receiver operating characteristic curve;
CHARMS: CHecklist for critical Appraisal and data extraction for systematic

Reviews of prediction Modelling Studies; CRIB: Clinical risk index for babies; GRADE: Grading of Recommendations Assessment, Development, and Evaluation; H-L: Hosmer-Lemeshow; HMD: Hyaline membrane disease; IVH: Intraventricular haemorrhage; LMIC: Low and middle income countries; NICU: Neonatal intensive care unit; NTISS: Neonatal therapeutic intervention scoring system; NUMIS: Nursing utilisation management information system; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PROBAST: Prediction study risk of bias assessment tool; PROSPERO: International prospective register of systematic reviews; PVL: Periventricular leukomalacia; ROP: Retinopathy of prematurity; SAW: Sex age weight score; SNAP: Score for neonatal acute physiology; SNAP-PE: Score for neonatal acute physiology – perinatal extension; TISS: Therapeutic intervention scoring system; TRIPS: Transport Risk Index for Physiological Stability; TTN: Transient tachypnoea of the newborn

Acknowledgements

The search strategy was developed with the help of an expert medical librarian, Nia Roberts, Knowledge Centre, Bodleian Health Care Libraries, University of Oxford, UK. This work is published with the permission of the director of KEMRI.

Funding

This work was supported in part by a Health Systems Research Initiative joint grant provided by the Department for International Development, UK (DFID), Economic and Social Research Council (ESRC), Medical Research Council (MRC) and Wellcome Trust, grant number MR/M015386/1. Funds from a Wellcome Trust Senior Fellowship (#097170) supported ME while additional funds from a Wellcome Trust core grant awarded to the KEMRI-Wellcome Trust Research Programme (#092654) also supported this work. The funding body played no role in the design of the study and collection, analysis and interpretation of data and in writing the manuscript.

Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Authors' contributions

JA and ME conceived the idea for the review. JA and ME conducted the literature searches. JA and ME screened the records for eligibility. JA and MM extracted the data and assessed its quality. JA and GSC synthesised the results. JA drafted the final review. All the authors (JA, GSC, MM, JAB and ME) reviewed and approved the final draft.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹KEMRI-Wellcome Trust Research Programme, P.O Box 43640 – 00100, Nairobi, Kenya. ²Department of Paediatrics and Child Health, College of Health Sciences, University of Nairobi, Kenyatta National Hospital, P. O. Box 19676-00202, Nairobi, Kenya. ³Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7FZ, UK. ⁴The Childhood Acute Illness & Nutrition (CHAIN) Network, P.O Box 43640 – 00100, Nairobi, Kenya. ⁵Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Oxford OX3 7LD, UK.

Received: 25 July 2017 Accepted: 28 November 2017

Published online: 07 December 2017

References

- Bhutta ZA, Das JK, Bahl R, Lawn JE, Salam RA, Paul VK, et al. Can available interventions end preventable deaths in mothers, newborn babies, and stillbirths, and at what cost? *Lancet*. 2014;384(9940):347–70.
- Dickson KE, Simen-Kapeu A, Kinney MV, Huicho L, Vesel L, Lackritz E. Lancet Every Newborn Study Group. Every Newborn: health-systems bottlenecks and strategies to accelerate scale-up in countries. *Lancet* 2014;384.
- Vincent J-L, Moreno R. Clinical review: scoring systems in the critically ill. *Crit Care*. 2010;14(2):207.
- Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10(2):e1001381.
- Kleinrouweler CE, Cheong-See FM, Collins GS, Kwee A, Thangaratnam S, Khan KS, et al. Prognostic models in obstetrics: available, but far from applicable. *Am J Obstet Gynecol*. 2015;
- Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14(1):40.
- Collins GS, Le Manach Y. Statistical inefficiencies in the development of a prediction model. *Anesth Analg*. 2017;124(3):1011–2.
- Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130(6):515–24.
- Adams ST, Leveson SH. Clinical prediction rules. *BMJ*. 2012;344:d8312.
- Dorling JS, Field DJ, Manktelow B. Neonatal disease severity scoring systems. *Arch Dis Child Fetal Neonatal Ed*. 2005;90(1):F11–F6.
- Dorling JS, Field DJ. Value and validity of neonatal disease severity scoring systems. *Arch Dis Child Fetal Neonatal Ed*. 2008;93(2):F80–F2.
- Fleisher BE, Murthy L, Lee S, Constantinou JC, Benitz WE, Stevenson DK. Neonatal severity of illness scoring systems: a comparison. *Clin Pediatr*. 1997;36(4):223–7.
- Marcin JP, Pollack MM. Review of the methodologies and applications of scoring systems in neonatal and pediatric intensive care. *Pediatr Crit Care Med*. 2000;1(1):20–7.
- Patrick SW, Schumacher RE, Davis MM. Methods of mortality risk adjustment in the NICU: a 20-year review. *Pediatrics*. 2013;131(Supplement 1):S68–74.
- Sacco Casamassima MG, Salazar JH, Papandria D, Fackler J, Chrouser K, Boss EF, et al. Use of risk stratification indices to predict mortality in critically ill children. *Eur J Pediatr*. 2014;173(1):1–13.
- Tarnow-Mordi WO. What is the role of neonatal organ dysfunction and illness severity scores in therapeutic studies in sepsis? *Pediatr Crit Care Med*. 2005;6(3):S135–S7.
- Pollack MM, Koch MA, Bartel DA, Rapoport I, Dhanireddy R, El-Mohandes AAE, et al. A comparison of neonatal mortality risk prediction models in very low birth weight infants. *Pediatrics*. 2000;105(5):1051–7.
- Medlock S, Ravelli ACJ, Taminga P, Mol BWI, Abu-Hanna A. Prediction of mortality in very premature infants: a systematic review of prediction models. *PLoS One*. 2011;6(9):e23441.
- Gray JE, Richardson DK, McCormick MC, Workman-Daniels K, Goldmann DA. Neonatal therapeutic intervention scoring system: a therapy-based severity-of-illness index. *Pediatrics*. 1992;90(4):561–7.
- Dickson KE, Simen-Kapeu A, Kinney MV, Huicho L, Vesel L, Lackritz E, et al. Every Newborn: health-systems bottlenecks and strategies to accelerate scale-up in countries. *Lancet*. 2014;384(9941):438–54.
- World Health Organization. Pregnancy, childbirth, postpartum and newborn care: a guide for essential practice –3rd ed. 2015.
- Aluvaala J, Nyamai R, Were F, Wasunna A, Kosgei R, Karumbi J, et al. Assessment of neonatal care in clinical training facilities in Kenya. *Arch Dis Child*. 2015;100(1):42–7.
- Rosenberg RE, Ahmed S, Saha SK, ASMNU A, MAK A, Law PA, et al. Simplified age-weight mortality risk classification for very low birth weight infants in low-resource settings. *J Pediatr*. 2008;153(4):519–24.e3.
- Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11(10):e1001744.
- Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924–6.
- Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J Clin Epidemiol*. 2009;62(10):1006–12.
- Geersing GJ, Bouwmeester W, Zuihthoff P, Spijker R, Leeflang M, Moons K. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. *PLoS One*. 2012;7
- Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol*. 2011;64(4):383–94.
- Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ*. 2008;336(7653):1106–10.
- Brozek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines: part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. *Allergy*. 2009;64(8):1109–16.
- Gopalakrishna G, Davenport C, Scholten RJP, Hyde C, Brozek J, et al. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. *J Clin Epidemiol*. 2014;67(7):760–8.
- Shah PS, Mirea L, Ng E, Solimano A, Lee SK. Association of unit size, resource utilization and occupancy with outcomes of preterm infants. *J Perinatol*. 2015;35(7):522–9.
- Davies MRQ. The need for a universal method of quantifying severity of illness to allow accurate analysis of the results of treatment in neonatal surgical cases. *Pediatr Surg Int*. 1995;10(5–6):305–8.
- Eriksson M, Bodin L, Finnström O, Schollin J. Can severity-of-illness indices for neonatal intensive care predict outcome at 4 years of age? *Acta Paediatr*. 2002;91(10):1093–100.
- Zupancic JA, Richardson DK. Characterization of neonatal personnel time inputs and prediction from clinical variables—a time and motion study. *J Perinatol*. 2002;22(8):658–63.
- Mendes I, Carvalho M, Almeida RT, Moreira ME. Use of technology as an evaluation tool of clinical care in preterm newborns. *J Pediatr*. 2006;82(5):371–6.
- Rojas JG, Henao-Murillo NA, Quirós-Jaramillo A. A tool for calculating the nursing staff at neonatal intensive care units. *Aquichán*. 2011;11:126–39.
- Oygur N, Ongun H, Saka O. Risk prediction using a neonatal therapeutic intervention scoring system in VLBW and ELBW preterm infants. *Pediatr Int*. 2012;54(4):496–500.
- Wu P-L, Lee W-T, Lee P-L, Chen H-L. Predictive power of serial neonatal therapeutic intervention scoring system scores for short-term mortality in very-low-birth-weight infants. *Pediatrics Neonatol*. 2015;56(2):108–13.
- Georgieff MKM, Mills MMR, Bhatt PB. Validation of two scoring systems which assess the degree of physiologic instability in critically ill newborn infants. *Crit Care Med*. 1989;17(1):17–21.
- Cullen DJM, Civetta JMM, Briggs BAM, Ferrara LCR. Therapeutic intervention scoring system: a method for quantitative comparison of patient care. *Crit Care Med*. 1974;2(2):57–60.
- Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol*. 2005;58(5):475–83.
- Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med*. 2016;35(2):214–26.
- Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167–76.
- Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 74:167–76.
- Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ*. 2009;338
- Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med*. 2000;19(4):453–73.
- Hendriksen JMT, Geersing GJ, Moons KGM, de Groot JAH. Diagnostic and prognostic prediction models. *J Thromb Haemost*. 2013;11:129–41.
- Shmueli G. To explain or to predict? *Stat Sci*. 2010:289–310.
- Lee SK, Aziz K, Dunn M, Clarke M, Kovacs L, Ojah C, et al. Transport Risk Index of Physiologic Stability, Version II (TRIPS-II): a simple and practical neonatal illness severity score. *Amer J Perinatol*. 2013;30(05):395–400.

51. Parry G, Tucker J, Tarnow-Mordi W. CRIB II: an update of the clinical risk index for babies score. *Lancet*. 2003;361(9371):1789–91.
52. Keene ARR, Cullen DJM. Therapeutic intervention scoring system: update 1983. *Crit Care Med* 1983;11(1):1–3.
53. Cullen DJMM, Nemeskal ARR, Zaslavsky AMP. Intermediate TISS: a new therapeutic intervention scoring system for non-ICU patients. *Crit Care Med*. 1994;22(9):1406–11.
54. Gemke RJ, Bonsel GJ, McDonnell J, van Vught AJ. Patient characteristics and resource utilisation in paediatric intensive care. *Arch Dis Child*. 1994;71(4):291–6.
55. Reis Miranda DMD, de Rijk AB, Schaufeli WP. Simplified Therapeutic Intervention Scoring System: the TISS-28 items—results from a multicenter study. *Crit Care Med*. 1996;24(1):64–73.
56. Richardson DK, Corcoran JD, Escobar GJ, Lee SK. SNAP-II and SNAPPE-II: simplified newborn illness severity and mortality risk scores. *J Pediatr*. 2001;138(1):92–100.
57. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.
58. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925–31.
59. Sainani KL. Explanatory versus predictive modeling. *PM&R*. 2014;6(9):841–4.
60. Ministry of Health, Government of Kenya. Basic Paediatric Protocols for ages up to 5 years. 2016.
61. Kleijnen Systematic Reviews Ltd. PROBAST [Available from: <http://s371539711.initial-website.co.uk/probast/>].
62. Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med*. 2012;9(5):e1001221.
63. Collins GS, Omar O, Shanyinde M, Yu L-M. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol*. 2013;66(3):268–77.
64. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. *BMC Med*. 2010;8.
65. Collins GS, Mallett S, Omar O, Yu L-M. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med*. 2011;9(1):103.
66. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353.
67. Aluvaala J, Nyamai R, Were F, Wasunna A, Kosgei R, Karumbi J, et al. Assessment of neonatal care in clinical training facilities in Kenya. *Arch Dis Child*. 2014;
68. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352.
69. Traeger AC, Hübscher M, McAuley JH. Understanding the usefulness of prognostic models in clinical decision-making. *J Phys*. 2017;63(2):121–5.
70. Hingorani AD, Windt DA, Riley RD, Abrams K, KGM M, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ*. 2013;346.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

