THE LANCET
Child & Adolescent Health

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Medvedev MM, Brotherton H, Gai A, et al. Development and validation of a simplified score to predict neonatal mortality risk among neonates weighing 2000 g or less (NMR-2000): an analysis using data from the UK and The Gambia. *Lancet Child Adolesc Health* 2020; published online Feb 28. https://doi.org/10.1016/S2352-4642(20)30021-3.

Methods appendix with supplementary tables and figures

Selection of candidate variables

Studies describing existing scoring systems for assessing neonatal mortality risk, illness severity, and clinical instability were reviewed to generate a list of potential parameters (Table S1). Parameters that are typically unavailable (e.g., oxygenation index), infrequently obtained (e.g., haematocrit), or unreliably measured (e.g., urine output) in low-resource settings were excluded (Table S2). Remaining parameters were evaluated using the following exclusion criteria: low prevalence in the NNRD (<0·1%); high prevalence of missing data in the development dataset (≥20%); not predictive of mortality in preterm or low birthweight neonates (e.g., thermoregulated environment); low prevalence within 24 hours (h) of birth (e.g., phototherapy); limited evidence to support validity (e.g., black race); concept better represented by an alternative variable (Table S2). Selection of candidate variables was conducted by members of the research team, which includes three neonatologists (one from US, two from UK), two of whom have extensive experience working in neonatal care in East Africa; a UK paediatrician based in The Gambia working in neonatal care; Gambian and Ugandan medical doctors with experience in neonatal care; and a UK paediatrician who is a global expert on newborn care and has an extensive background in neonatal care in LMICs, including throughout sub-Saharan Africa.

Study participants and data acquisition

UK samples

This study utilised data from 187 neonatal units in the UK National Neonatal Research Database (NNRD). The NNRD holds electronic patient-level data, recorded by health professionals as part of routine clinical care, from admissions to National Health Service (NHS) neonatal units in England from 2008, and Wales and Scotland from 2012. Data in the NNRD are de-identified and critical data items are fed back to and validated by treating clinicians. A formal comparison of NNRD data items against those recorded in case report forms of a randomised controlled trial demonstrated a high level of agreement (>95%).[1] Items in the National Neonatal Data Set are held within the NHS Data Dictionary.[2] This study included data on newborns admitted to neonatal units in England and Wales between January 2010 and December 2017.

The NNRD includes a total of approximately 140000 neonates born weighing ≤2000 grams (g) who were admitted to participating neonatal units in England and Wales from January 2010 to December 2017. For model development, 5 to 10 outcome events per predictor variable are required to obtain accurate and clinically useful results.[3] Using this guidance, the dataset was divided into the following samples:

- Development sample – all neonates ≤2000g admitted to a random sample of participating neonatal units in England and Wales from January 2010 to December 2016
- Random validation sample – all neonates ≤2000g admitted to the remaining participating neonatal units in England and Wales from January 2010 to December 2016
- Temporal validation sample – all neonates ≤2000g admitted to all neonatal units in England and Wales from January to December 2017

The following exclusion criteria were applied: birthweight >2000g; admitted at >6h of age or following discharge home; stillborn; died in the delivery room; moribund (received only comfort care prior to in-hospital death). Comfort care was defined as not receiving intubation, mechanical ventilation, vasopressors, or chest compressions.

Gambian sample

The Gambian validation sample included all neonates <2000g who were admitted to the neonatal unit at Edward Francis Small Teaching Hospital in Banjul between May 2018 and September 2019, and who were screened but not enrolled in the 'Early KMC' (eKMC) trial (NCT03555981).

Routine data, including mode of delivery and treatments administered during the first 24h post-birth, were collected from routine medical charts and recorded in an Excel spreadsheet by trained study personnel. Other routine and non-routine data, collected as part of the eKMC trial screening process, were exported from the trial database and transferred to the spreadsheet. These data included birthweight, sex, birth plurality (singleton or multiple), referral status (inborn or outborn), and oxygen saturation (SpO_2) at admission (%). Due to the stepwise nature of the eKMC screening process, SpO_2 was not required for those neonates who were deemed ineligible. Accordingly, if a neonate's SPO_2 measurement was not recorded in the trial database, study personnel attempted to collect this data from routine medical charts.

Outcome measure

The primary outcome was in-hospital mortality. Mortality has been utilised as the outcome against which most neonatal intensive care risk scores have been designed and validated.[4,5,14,6–13] Mortality is clearly and directly related to illness severity, objectively measured, and reliably ascertainable.[9]

Missing data

Missing data were excluded from the analysis for continuous variables. In this study, categorical variables include therapy-based risk factors (e.g., IV fluids) and clinical signs that are non-continuous (e.g., convulsions). Recording the absence of categorical variables is not mandatory in the data sources from which the NNRD is extracted; thus, these fields are often left blank to indicate absence. Therefore, there are several reasons why a neonate may not have such a variable recorded.[15] First is the possibility that the neonate was healthy and, thus, did not require the therapy or have the clinical sign. This is the assumption that was made in the development of this score; the same assumption was made in the development of the widely used SNAP and CRIB scores. The second possibility is that the therapy was given or the clinical sign was present, but this information was not documented in the medical record. Given that a comparison of NNRD data items against those recorded for a randomised trial demonstrated >95% agreement,[1] this was considered to be unlikely. Other possibilities include that the clinician should have ordered the therapy or noted the clinical sign, but failed to do so as a result of inadequate knowledge or skill. This was thought to be unlikely in UK neonatal units, which are staffed by neonatologists and/or paediatricians experienced in neonatal care. The data sources from which the NNRD data are drawn are summary data describing the treatments received or the signs detected on a particular day; therefore, treatments 'ordered' but not administered will not be recorded and, thus, will not be included in the NNRD. This was confirmed in the aforementioned NNRD validation study.[1]

Model development

Continuous data were presented using means and standard deviations (SD) for parametric data and medians and interquartile ranges (IQR) for non-parametric data. Categorical data were presented as counts and proportions. Logistic regression models were derived to model the risk of in-hospital mortality. Robust standard errors were used to allow for clustering within neonatal units. All candidate variables were included in a complete multivariable model, which was progressively simplified using reverse stepwise selection, with the least statistically significant variable being removed at each step. Model discrimination was assessed with the c-index,[16] equivalent to the area under the receiver operating characteristic curve, which ranges from 0·5 (no predictive ability) to 1 (perfect discrimination).[17] Overall goodness-of-fit was assessed with the Brier score, which ranges from 0 (perfect fit) to a maximum value dependent upon outcome incidence (0·25 for 50% incidence).[18,19] Calibration was evaluated using graphical plots of observed versus predicted risk; perfect predictions lie on the 45 degree line.[19] Locally weighted scatterplot smoothing was used to estimate the relationship between observed and predicted probabilities.[20] A sensitivity analysis excluding neonates whose admission age was uncertain (anonymised data derived from calculated difference between time of birth and time of admission) was conducted to reassess performance, as admission at >6h of birth was an exclusion criterion. Model performance was additionally reassessed following exclusion of neonates who were transferred for any reason since outcome data were not available for these babies. Performance for predicting mortality within 24h of birth was evaluated in a secondary analysis, as 36% of neonatal deaths globally occur within this timeframe.[21]

Multiple imputation

Methods

We employed multiple imputation (MI) with chained equations to impute missing values for incomplete predictor variables in the development sample. The imputation model included the primary outcome, predictor variables, and candidate variables believed to be associated with missing data values and/or patterns of missingness. Candidate variables were added in a stepwise process to assess model convergence and compatibility; those resulting in convergence failure were excluded. Stata version 15 was used to perform all imputation analyses (*mi impute chained, mi estimate*). Continuous variables were imputed using predictive mean matching (k=10) due to non-normality and restricted range; categorical variables were imputed using logistic regression.[22] Variables were analysed in sequence from the most observed to the least observed. Fifteen imputed datasets were generated, with 10 iterations per dataset. The logistic regression model was executed across the 15 datasets and results were combined to create a single set of inferential statistics, using Rubin's rules.[23] MI estimates of the β coefficients and c-index were compared to original estimates. Monte Carlo errors were examined to assess statistical reproducibility.[23]

Results

Following imputation of missing values for incomplete predictor variables (n=54956), β coefficient estimates were nearly identical to original estimates (Table S3), thus no adjustments were made to the model coefficients. Discriminatory performance of the model was unchanged, with a c-index of 0·8894 (95% CI: 0·8818-0·8969). Estimates of Monte Carlo errors for β coefficients, standard errors, and p-values suggested that the MI process could be statistically reproduced. The average relative variance increase was 0·0457 and the largest fraction of missing information was 0·0902.

Risk score development

We assigned the parameters in the final model points proportional to their β regression coefficient values.[24–26] Whole numbers were used in order to generate an easily calculable score.

Logistic regression equation relating the risk model to in-hospital mortality:

$$\text{Log odds of mortality} = 2.6142 - (0.0032*\text{birthweight}) + (0.3167*\text{nasal cannula/headbox}) + (1.6214*\text{CPAP/mechanical ventilation}) - (0.0390*\text{admission } SpO_2)$$

Logistic regression equation relating the risk score to in-hospital mortality:

$$\text{Log odds of mortality} = 0.1901 - (0.3256*\text{NMR-2000})$$

A risk score was calculated for each patient and predictive margins with 95% confidence intervals (CI) were computed across a broad range of risk score percentiles (Table S4). Using these margins as a guide, the development sample was arbitrarily divided into three groups: neonates at low risk, medium risk, and high risk for mortality. To assess the calibration of the integer score to the model using regression coefficients, observed risks in risk groups and population deciles of the risk score were derived and compared with the mean predicted risks in each group or population decile (Figure S1). We assessed overall predictive ability of the integer score using the c-index.

Internal validation

Internal validity refers to the reproducibility of a risk prediction model for the underlying population from which the data originated.[27] Bootstrap resampling, with 1000 samples within the development sample, was used to internally validate the final model at the two time-points and to estimate optimism-adjusted measures of model discrimination and overall fit in each bootstrap sample.[28] Overfitting is a phenomenon whereby the process of generating a model that has optimal fit for the development dataset results in reduced fit when the model is applied to other datasets and, thus, provides an optimistic evaluation of its predictive ability.[29] Performance of the refitted model in each bootstrap sample was compared to that of the refitted model in the original development sample. Estimates of optimism for the c-index and Brier score were averaged and subtracted to provide optimism-adjusted measures.

External validation

External validity refers to the generalisability of a model's performance to related populations.[27] The final model was evaluated in three external validation samples, each selected to assess distinctive features of performance. The random validation sample, drawn from the neonatal units withheld from the development sample, tested the performance of the model when applied to different neonatal care settings in the UK within the same timeframe. The temporal validation sample, drawn from all units in the development and random validation samples during the final twelve months of data collection, tested performance in the same neonatal units over time. The Gambian sample tested performance in a LMIC neonatal care setting. We assessed model performance in each validation sample separately and in the UK full (combined) validation sample. Discrimination and overall goodness-of-fit were evaluated using the c-index and Brier score, respectively. Calibration was assessed using graphical plots of observed versus predicted risk. Sensitivity, specificity, positive predictive value, and negative predictive value were calculated across a wide range of possible cut-points in the UK development and full validation samples (Table S5).

We assessed overall performance of the simplified integer score using the c-index and Brier score in all external validation samples (Table S6). In the Gambian sample, we re-defined low-, medium-, and high-risk groups to account for increased case fatality in this setting. To assess the calibration of the integer score to the model using regression coefficients, observed risks in groups and population deciles of risk scores were derived and compared with mean predicted risks in each group or population decile of the Gambian sample (Figure S2). A risk score was calculated for each neonate and predictive margins with 95% CIs were computed across a broad range of score percentiles (Table S7). Using these margins as a guide, the Gambian sample was arbitrarily divided into three groups: neonates at low-, medium-, and high-risk for mortality.

Comparison with the CRIB-II score

The Clinical Risk Index for Babies (CRIB, CRIB-II),[5,8] the Score for Neonatal Acute Physiology (SNAP, SNAP-II),[6,9] and the SNAP Perinatal Extension (SNAPPE, SNAPPE-II)[7,9] are the most widely used neonatal intensive care risk scores. The Transport Risk Index of Physiologic Stability (TRIPS, TRIPS-II) is a related physiology-based approach that can be assessed at any point within the first 24h and repeated as a baby's clinical condition changes.[13,14] The NNRD did not include all of the variables required for calculation of CRIB, SNAP, SNAP-II, SNAPPE-II, TRIPS, or TRIPS-II (Table S2); hence, CRIB-II was selected for comparison. CRIB-II includes 5 variables (sex, birthweight, gestational age, temperature, base excess), all collected within 1h of admission.[8] As CRIB-II has only been validated for use in neonates up to 32 weeks gestational age, we compared the c-index and Brier score for CRIB-II and NMR-2000 amongst neonates ≤32 weeks in the full validation sample. All statistical analyses for this study were completed using Stata version 15 (College Station, Texas, United States).

References

1. Battersby C, Statnikov Y, Santhakumaran S, Gray D, Modi N, Costeloe K. The United Kingdom National Neonatal Research Database: A validation study. PLoS One 2018;**13:**e0201815.
2. National Health Service. National Neonatal Data Sets Menu. Available from: https://www.datadictionary.nhs.uk/data_dictionary/messages/clinical_data_sets/data_sets/national_neonatal_data_set/national_neonatal_data_set_-_episodic_and_daily_care_fr.asp?shownav=1. Accessed 2018 July 28.
3. Wasson J, Sox H, Neff R, Goldman L. Clinical prediction rules. Applications and methodological standards. N Engl J Med 1985;**313:**793–9.
4. Gray J, Richardson D, McCormick M, Workman-Daniels K, Goldmann D. Neonatal therapeutic intervention scoring system: a therapy-based severity-of-illness index. Pediatr 1992;**90:**561–7.
5. International Neonatal Network. The CRIB (clinical risk index for babies) score: a tool for assessing initial neonatal risk and comparing preformance of neonatal intensive care units. Lancet 1993;**342:**193–8.
6. Richardson D, Gray J, McCormick M, Workman K, Goldmann D. Score for neonatal acute physiology: a physiologic severity index for neonatal intensive care. Pediatr 1993;**91:**617–23.
7. Richardson K, McCormick C, Gray E. Birth weight and illness severity: independent predictors of neonatal mortality. Pediatr 1993;**91:**969–75.
8. Parry G, Tucker J, Tarnow-Mordi W, for the UK Neonatal Staffing Study Collaborative Group. CRIB II: an update of the clinical risk index for babies score. Lancet 2003;**361:**1789–91.
9. Richardson D, Corcoran J, Escobar G, Lee S. SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores. J Pediatr 2001;**138:**92–100.
10. Zupancic J, Richardson D, Horbar J, Carpenter J, Lee S, Escobar G. Revalidation of the Score for Neonatal Acute Physiology in the Vermont Oxford Network. Pediatr 2007;**119:**e156–63.
11. Broughton S, Berry A, Jacobe S, Cheeseman P, Tarnow-Mordi W. The Mortality Index for Neonatal Transportation Score: A New Mortality Prediction Model for Retrieved Neonates. Pediatr 2004;**114:**e424–8.
12. Sutcuoglu S, Celik T, Alkan S, Ilhan O, Ozer E. Comparison of neonatal transport scoring systems and transport-related mortality score for predicting neonatal mortality risk. Pediatr Emerg Care 2015;**31:**113–6.
13. Lee S, Zupancic J, Pendray M, Thiessen P, Schmidt B, Whyte R, et al. Transport risk index of physiologic stability: A practical system for assessing infant transport care. J Pediatr 2001;**139:**220–6.
14. Lee S, 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. Am J Perinatol 2013;**30:**395–400.
15. Richardson D, Tarnow-Mordi W, Lee S. Risk adjustment for quality improvement. Pediatr 1999;**103:**255–65.
16. Harrell F, Califf R, Pryor D, Lee K, Rosati R. Evaluating the yield of medical tests. JAMA 1982;**247:**2543–6.
17. Hanley A, McNeil J. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology 1982;**143:**29–36.
18. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;**75:**1–3.
19. Steyerberg E, Vickers A, Cook N, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology 2010;**21:**128–38.
20. Harrell F. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
21. Lawn J, Blencowe H, Oza S, You D, Lee A, Waiswa P, et al. Progress, priorities, and potential beyond survival. Lancet 2014;**384:**189–205.
22. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007;**16:**219–42.
23. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 2011;**30:**377–99.
24. Mehta H, Mehta V, Girman C, Adhikari D, Johnson M. Regression coefficient–based scoring system should be used to assign weights to the risk index. J Clin Epidemiol 2016;**79:**22–8.
25. Moons K, Harrell F, Steyerberg E. Should scoring rules be based on odds ratios or regression coefficients? J Clin Epidemiol 2002;**55:**1054–5.
26. Harrell F. Regression coefficients and scoring rules. J Clin Epidemiol 1996;**49:**819.
27. Steyerberg E, Vergouwe Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. Eur Heart J 2014;**35:**1925–31.
28. Steyerberg E, Moons K, van der Windt D, Hayden J, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Med 2013;**10:**e1001381.
29. Smith G, Seaman S, Wood A, Royston P, White I. Correcting for optimistic prediction in small data sets. Am J Epidemiol 2014;**180:**318–24.
30. Horbar J, Onstad L, Wright E, Yaffe S, Catz C, Wright L, et al. Predicting mortality risk for infants weighing 501 to 1500 grams at birth: A National Institutes of Health Neonatal Research Network report. Crit Care Med 1993;**21:**12–8.

31. Maier R, Rey M, Metze B, Obladen M. Comparison of mortality risk: A score for very low birthweight infants. Arch Dis Child Fetal Neonatal Ed 1997;**76:**F146-151.
32. Fischer C, Sontheimer D, Scheffer F, Bauer J, Linderkamp O. Cardiorespiratory stability of premature boys and girls during kangaroo care. Early Hum Dev 1998;**52:**145–53.
33. Rosenberg R, Ahmed S, Saha S, Ahmed N, Chowdhury M, Law P, et al. Simplified Age-Weight Mortality Risk Classification for Very Low Birth Weight Infants in Low-Resource Settings. J Pediatr 2008;**153:**519–24.
34. Shah P, Mirea L, Ng E, Solimano A, Lee S. Association of unit size, resource utilization and occupancy with outcomes of preterm infants. J Perinatol 2015;**35:**522–9.
35. Rathod D, Adhisivam B, Vishnu Bhat B. Sick Neonate Score - A Simple Clinical Score for Predicting Mortality of Sick Neonates in Resource Restricted Settings. Indian J Pediatr 2016;**83:**103–6.
36. Morgan M, Nambuya H, Waiswa P, Tann C, Elbourne D, Seeley J, et al. Kangaroo Mother Care for clinically unstable neonates: is it feasible at a hospital in Uganda? J Glob Health 2018;**8:**010701.
37. Suresh GK, Horbar JD, Kenny M, Carpenter JH. Major birth defects in very low birth weight infants in the Vermont Oxford Network. J Pediatr 2001;**139:**366–73.
38. Youden WJ. Index for rating diagnostic tests. Cancer 1950;**3:**32–5.

Table S1. Characteristics of development studies reviewed to generate list of potential parameters

	Model	Approach	Dates	Setting	Sample size[a]	In-hospital mortality[a]
Gray, 1992[4]	NTISS[b]	Therapy-based	1989-90	3 NICU[c], USA	1768	114
Horbar, 1993[30]	NICHHD[d]	Perinatal factors	1987-89	7 NICU[c], USA	3603	890
International Neonatal Network, 1993[5]	CRIB[e]	Physiology-based + perinatal factors	1988-90	4 NICU[c], UK	812	201
Richardson, 1993[6]	SNAP[f]	Physiology-based	1989-90	3 NICU[c], USA	1643	114
Richardson, 1993[7]	SNAPPE[g]	Physiology-based + perinatal factors	1989-90	3 NICU[c], USA	1089	59
Maier, 1997[31]	Unnamed	Physiology-based + perinatal factors + 1 therapy measure	1978-87	1 NICU[c], Germany	396	106
Fischer, 1998[32]	SCRIP[h]	Physiology-based	Not reported	1 NICU[c], Germany	20	Not reported
Richardson, 2001[9]	SNAP-II[f], SNAPPE-II[g]	Physiology-based (SNAP) + perinatal factors (SNAPPE)	1996-97	17 NICU[c], Canada	10819	418
Lee, 2001[13]	TRIPS[i]	Physiology-based	1996-97	8 NICU[c], Canada (transport service)	1115	Not reported
Parry, 2003[8]	CRIB-II[e]	Physiology-based + perinatal factors	1998-99	54 NICU[c], UK	3027	240
Broughton, 2004[11]	MINT[j]	Physiology-based + perinatal factors	1992-2001	Neonatal transport service in Australia	1252	138
Zupancic, 2007[10]	VON-RA[k]	Physiology-based + perinatal factors	2002	>500 NICU[c], Vermont Oxford Network	10439	1072
Rosenberg, 2008[33]	SAWS[l]	Perinatal factors	1998-2003	2 NICU[c], Egypt and Bangladesh	428[j]	262
Lee, 2013[14]	TRIPS-II[i]	Physiology-based	2006-08	15 NICU[c], Canada	11383	411
Sutcuoglu, 2015[12]	TREMS[m]	Physiology-based	2011	1 NICU[c], Turkey (transport service)	306	56
Shah, 2015[34]	Unnamed	Therapy-based + perinatal factors	2010-12	23 NICU[c], Canada	9978	650
Rathod, 2016[35]	SNS[n]	Physiology-based	2012-13	1 NICU[c], India (transport service)	303	60
Morgan, 2018[36]	Unnamed	Therapy-based	2015-16	1 neonatal unit (regional referral hospital), Uganda	264[o]	2[o]

[a] Total sample size and number of in-hospital deaths in the development cohort.
[b] Neonatal Therapeutic Intervention Scoring System (NTISS)
[c] Neonatal intensive care unit (NICU)
[d] National Institute of Child Health and Human Development (NICHHD)
[e] Clinical Risk Index for Babies (CRIB, CRIB-II)
[f] Score for Neonatal Acute Physiology (SNAP, SNAP-II)
[g] Score for Neonatal Acute Physiology Perinatal Extension (SNAPPE, SNAPPE-II)
[h] Stability of the Cardio-Respiratory System in Premature Infants (SCRIP)
[i] Transport Risk Index of Physiologic Stability (TRIPS, TRIPS-II)
[j] Mortality Index for Neonatal Transportation (MINT)
[k] Vermont Oxford Network-Risk Adjustment (VON-RA)
[l] Simplified age-weight-sex (SAWS); sample comprised of neonates enrolled in clinical trials of topical emollient therapy at two tertiary care hospitals (one in Egypt, one in Bangladesh).
[m] Transport Related Mortality Score (TREMS)
[n] Sick Neonate Score (SNS)
[o] Sample comprised of 254 neonates in a retrospective audit and 10 in a prospective study evaluating the feasibility of kangaroo mother care for clinically unstable neonates; in-hospital mortality is reported for the feasibility study.

Table S2. Parameters evaluated for potential inclusion in the modelling process

	Model(s)	Inclusion/exclusion
Perinatal factors	NA[a]	NA[a]
Birthweight	CRIB[b], SNAPPE-II[c], CRIB-II[b], MINT[d], SAWS[e], NICHHD[f], Maier[g]	Included
Gestational age	CRIB, CRIB-II[b], SAWS[e], VON-RA[h]	Included
Sex	CRIB-II[b], SAWS[e], NICHHD[f], VON-RA[h]	Included
Postnatal age	MINT[d]	Excluded- only validated as a binary risk factor (0-1 vs. >1 hour) among neonates transported within 72 hours (h)[11]
Small-for-gestational age	SNAPPE-II[c], NICHHD[f]	Included
Apgar score at 1 minute	MINT[d], NICHHD[f], VON-RA[h]	Excluded- often unavailable in LMIC facilities, especially for babies born at home or transferred from another facility
Apgar score <7 at 5 minutes, Apgar score at 5 minutes	SNAPPE-II[c], Maier[g], VON-RA[h]	Excluded- as above
Congenital anomaly[i]	CRIB[b], MINT[d], VON-RA[h]	Unreliable diagnosis in LMIC settings- modified to 'presence of visually recognisable anomaly at birth' using a predefined list of conditions
Black race	NICHHD[f]	Excluded- limited evidence; only validated in 1 study (published in 1993)
Outborn status	VON-RA[h]	Excluded- only validated in combination with SNAP-II[10]
Multiple gestation	VON-RA[h]	Excluded- as above
Caesarean delivery	VON-RA[h]	Excluded- as above, plus not available in many LMIC facilities
Therapy-based	NA[a]	NA[a]
Cardiopulmonary resuscitation in first 24h	NTISS[j]	Modified to 'bag-mask resuscitation at time of delivery'
Surfactant administration in first 24h	NTISS[j]	Excluded- infeasible for LMIC settings
Oxygen therapy in first 24/48h	NTISS[j], KMC[k]	Included as 'oxygen therapy within 24h of birth'
Continuous positive airway pressure in first 24/48h	NTISS[j], KMC[k]	Modified to 'highest level of respiratory support within 24h of birth'
Mechanical/high frequency ventilation in first 24h, at admission	NTISS[j], Maier[g]	Excluded- better represented by alternative variable ('highest level of respiratory support within 24h of birth')
Tracheostomy in first 24h	NTISS[j]	Excluded- infeasible for LMIC settings
Extracorporeal membrane oxygenation in first 24h	NTISS[j]	Excluded- infeasible for LMIC settings
Endotracheal intubation in first 24h	NTISS[j]	Excluded- better represented by alternative variable ('highest level of respiratory support within 24h of birth'), low data completeness in NNRD
PaO_2/FiO_2 ratio in first 12/24h	SNAP-II/SNAP[l]	Excluded- not included in NNRD, infeasible for LMIC settings
Oxygenation index in first 24h	SNAP[l]	Excluded- not included in NNRD, infeasible for LMIC settings
Minimum/maximum FiO_2 in first 12h	CRIB[b]	Excluded- not included in NNRD, infeasible for LMIC settings
Indomethacin administration in first 24h	NTISS[j]	Excluded- infeasible for LMIC settings
Vasopressor administration in first 24h	NTISS[j]	Excluded- infeasible for LMIC settings
Pacemaker therapy in first 24h	NTISS[j]	Excluded- not included in NNRD, infeasible for LMIC settings
Antibiotic therapy in first 24/48h	NTISS[j], KMC[k]	Included as 'antibiotic therapy within 24h of birth'
Diuretic therapy in first 24h	NTISS[j]	Excluded- infeasible for LMIC settings
Steroid administration in first 24h	NTISS[j]	Excluded- infeasible for LMIC settings
Anticonvulsant therapy in first 24/48h	NTISS[j], KMC[k]	Included as 'anticonvulsant therapy within 24h of birth'
Caffeine (or aminophylline) in first 24/48h	NTISS[j], KMC[k]	Included as 'caffeine or aminophylline within 24h of birth'
Treatment of metabolic acidosis in first 24h	NTISS[j]	Excluded- infeasible for LMIC settings
Potassium binding resin administration in first 24h	NTISS[j]	Excluded- infeasible for LMIC settings
Frequent vital signs/cardiorespiratory monitoring in first 24h	NTISS[j]	Excluded- unreliable, infeasible for LMIC settings
Frequent phlebotomy in first 24h	NTISS[j]	Excluded- infeasible for LMIC settings
Thermoregulated environment in first 24h	NTISS[j]	Excluded- not useful to predict mortality risk amongst neonates ≤2000g, as all require some form of thermal support (KMC, incubator, or radiant warmer)
Arterial, central venous pressure monitoring in first 24h	NTISS[j]	Excluded- not included in NNRD, infeasible for LMIC settings
Urinary catheter in first 24 hours	NTISS[j]	Excluded- not included in NNRD, infeasible for LMIC settings
Gavage feeding	NTISS[j]	Excluded- not useful to predict mortality risk amongst neonates ≤2000g, as those born at <35 weeks may require gavage feeding until coordinated suck and swallow develops (typically around 32 to 34 weeks)

Intravenous (IV) amino acid administration, IV potassium infusion within first 24h	NTISS[j]	Modified to 'IV fluids within 24h of birth'
IV fluids within 48h of birth	KMC[k]	Excluded- better represented by alternative variable ('IV fluids within 24h of birth')
Insulin administration in first 24h	NTISS[j]	Excluded- low prevalence, infeasible for LMIC settings
Phototherapy in first 24/48h	NTISS[j], KMC[k]	Excluded- low prevalence within 24h of birth
Blood product transfusion in first 24h	NTISS[j]	Excluded- infeasible for LMIC settings
Exchange transfusion in first 24h	NTISS[j]	Excluded- infeasible for LMIC settings
Patient transport in first 24h	NTISS[j]	Excluded- infeasible in many LMIC settings
Chest tube in first 24h	NTISS[j]	Excluded- low prevalence in NNRD, infeasible for LMIC settings
Pericardial tube in first 24h	NTISS[j]	Excluded- low prevalence in NNRD, infeasible for LMIC settings
Operation in first 24h	NTISS[j]	Excluded- low prevalence in NNRD, infeasible for LMIC settings
Thoracentesis in first 24h	NTISS[j]	Excluded- low prevalence in NNRD, infeasible for LMIC settings
Pericardiocentesis in first 24h	NTISS[j]	Excluded- low prevalence in NNRD, infeasible for LMIC settings
Dialysis in first 24h	NTISS[j]	Excluded- low prevalence in NNRD, infeasible for LMIC settings
Vascular access in first 24h[m]	NTISS[j]	Excluded- low prevalence in NNRD, arterial and central venous access infeasible for LMIC settings
Clinical signs/observations	NA[a]	NA[a]
Blood pressure in first 12/24h, at admission	SNAP-II/SNAP[l], TRIPS, TRIPS-II[n], TREMS[o], SNS[p]	Excluded- high proportion of missing data (30·3%) in development set
Heart rate in first 24h, at admission	SNAP[l], MINT[d], SNS[p]	Included as 'Heart rate at admission'
Respiratory rate in first 24h, at admission	SNAP[l], SNS[p]	Modified to 'Respiratory rate at admission'
Temperature at admission (within first hour)	CRIB-II[b], TREMS[o], TRIPS, TRIPS-II[n], SNS[p]	Included as 'temperature at admission'
Temperature in first 12/24h	SNAP-II/SNAP[l]	Excluded- better represented by alternative variable ('temperature at admission')
Oxygen saturation in first 24h, at admission	NTISS[j], TREMS[o], SNS[p]	Included as 'Oxygen saturation at admission'
Urine output in first 12/24h, quantitative intake and output in first 24h	SNAP-II/SNAP[l], NTISS[j]	Excluded- unreliable measure, infeasible for LMIC settings
Number of seizures in first 12/24h	SNAP-II/SNAP[l]	Number not included in NNRD- modified to 'any seizures within 24h of birth'
Apnoeic episodes in first 24h	SNAP[l]	Unreliable measure in present form- modified to 'clinically relevant increase in apnoea/bradycardia episodes, oxygen requirement, or ventilatory support'
Respiratory status/effort, severity of respiratory distress at admission	TRIPS, TRIPS-II[n], Maier[g], SNS[p]	Excluded- better represented by alternative variables ('RR at admission,' 'SpO_2 at admission' 'clinically relevant increase in apnoea/bradycardia episodes, oxygen requirement, ventilatory support, or respiratory rate')
Response to noxious stimuli	TRIPS, TRIPS-II[n]	Excluded- not included in NNRD
Capillary refill time at admission	SNS[p]	Excluded- prevalence <0.1% in NNRD
Episodes of apnoea, bradycardia, or oxygen desaturation measured over 5-minute periods 13 times throughout the first 6h	SCRIP[q]	Excluded- infeasible for routine clinical use; better represented by alternative variable ('clinically relevant increase in apnoea/bradycardia episodes, oxygen requirement, ventilatory support, or respiratory rate')
Laboratory measures	NA[a]	NA[a]
Serum pH in first 12/24h, at admission	SNAP-II/SNAP[l], MINT[d]	Excluded- infeasible for routine use in LMIC settings
PaO_2 in first 24h, at admission	SNAP[l], MINT[d]	Excluded- not included in NNRD, infeasible for LMIC settings
pCO_2 in first 24h, at admission	SNAP[l], TREMS[o]	Excluded- infeasible for routine use in LMIC settings
Base excess in first 12h, within 1h, on admission	CRIB, CRIB-II[b], Maier[g]	Excluded- infeasible for routine use in LMIC settings
Haematocrit in first 24h	SNAP[l]	Excluded- not included in NNRD
White blood cell count in first 24h	SNAP[l]	Excluded- not included in NNRD
Immature total ratio in first 24h	SNAP[l]	Excluded- not included in NNRD, infeasible for LMIC settings
Absolute neutrophil count in first 24h	SNAP[l]	Excluded- not included in NNRD, infeasible for LMIC settings
Platelet count in first 24h	SNAP[l]	Excluded- not included in NNRD, infeasible for routine use in LMIC settings
Blood urea nitrogen in first 24h	SNAP[l]	Excluded- not included in NNRD, infeasible for routine use in LMIC settings
Creatinine in first 24h	SNAP[l]	Excluded- not included in NNRD, infeasible for routine use in LMIC settings
Bilirubin in first 24h	SNAP[l]	Excluded- not included in NNRD, infeasible for routine use in LMIC settings

Sodium in first 24h	SNAP[l]	Excluded- not included in NNRD, infeasible for routine use in LMIC settings
Potassium in first 24h	SNAP[l]	Excluded- not included in NNRD, infeasible for routine use in LMIC settings
Calcium in first 24h	SNAP[l]	Excluded- not included in NNRD, infeasible for routine use in LMIC settings
Blood glucose in first 24h, on admission	SNAP[l], TREMS[o], SNS[p]	Excluded- high proportion of missing data (23·9%) in development set
Serum bicarbonate in first 24h	SNAP[l]	Excluded- not included in NNRD, infeasible for routine use in LMIC settings
Stool guaiac in first 24h	SNAP[l]	Excluded- not included in NNRD

[a] Not applicable (NA).
[b] Clinical Risk Index for Babies (CRIB, CRIB-II).
[c] Score for Neonatal Acute Physiology Perinatal Extension (SNAPPE, SNAPPE-II).
[d] Mortality Index for Neonatal Transportation (MINT).
[e] Simplified age-weight-sex (SAWS).
[f] National Institute of Child Health and Human Development (NICHHD).
[g] Unnamed mortality risk score for VLBW neonates, published by Maier et al.
[h] Vermont Oxford Network-Risk Adjustment (VON-RA).
[i] The CRIB score stratified the risk of congenital anomalies into 3 categories: 1) none; 2) non-acutely life threatening; 3) acutely life threatening.[5] The MINT score categorized this variable solely by its presence or absence, as recorded at the time of the referral call.[11] The VON-RA score defined congenital anomalies using a predefined list of conditions.[37]
[j] Neonatal Therapeutic Intervention Scoring System (NTISS).
[k] Therapy-based clinical instability criterion used in study exploring KMC feasibility amongst unstable neonates in Uganda.[36]
[l] Score for Neonatal Acute Physiology (SNAP, SNAP-II).
[m] The NTISS defined vascular access to include peripheral IV, arterial, and central venous lines, with higher therapeutic intensity weights assigned to arterial and central venous access (subscore: 2) than peripheral IV access (subscore: 1).[4]
[n] Transport Risk Index of Physiologic Stability (TRIPS, TRIPS-II).
[o] Transport Related Mortality Score (TREMS).
[p] Sick Neonate Score (SNS).
[q] Stability of the Cardio-Respiratory System in Premature Infants (SCRIP).

Table S3. NMR-2000 logistic model following multiple imputation versus original estimates in development sample

	Multiple imputation (n=54956)		Original estimates (n=46108)	
	β Coefficient	**95% Confidence interval**	**β Coefficient**	**95% Confidence interval**
Birthweight	-0.0032	-.0035 to -.00289[a]	-0·0032	-0·0035 to -0·0029[a]
Highest respiratory support within first 24h	··, NA[b]	··, NA[b]	··, NA[b]	··, NA[b]
Nasal cannula or headbox	0.3896	0.0014 to 0.7778[c]	0·3167	-0·1055 to 0·7389[a]
CPAP, Bi/SiPAP, or invasive ventilation	1.4977	1.1909 to 1.8045[a]	1·6214	1·2682 to 1·9746[a]
SpO_2 at admission (%)	-0.0386	-0.0449 to -0.0322[a]	-0·0390	-0·0455 to -0·0326[a]
Constant	2.8229	1.9410 to 3.7047[a]	2·6142[f]	1·7655 to 3·4629[a]

[a] Estimate significant to p-value <0·0001.
[b] Not applicable.
[c] Estimate significant to p-value <0·05.

Table S4. Predicted mortality risk across score percentiles in the development sample (n=46108)

Percentile	Score	Mean predicted mortality risk[a]	95% Confidence interval[a]
1%	3·9	34·1	29·1 - 39·0
5%	6·1	18·5	15·6 - 21·4
10%	7·9	11·1	9·3 - 12·8
25%	12·0	4·7	3·9 - 5·4
50%	17·2	1·1	0·9 - 1·3
75%	21·1	0·2	0·2 - 0·3
90%	22·9	0·1	0·06 - 0·1

[a] All predictions significant to p-value <0.0001.

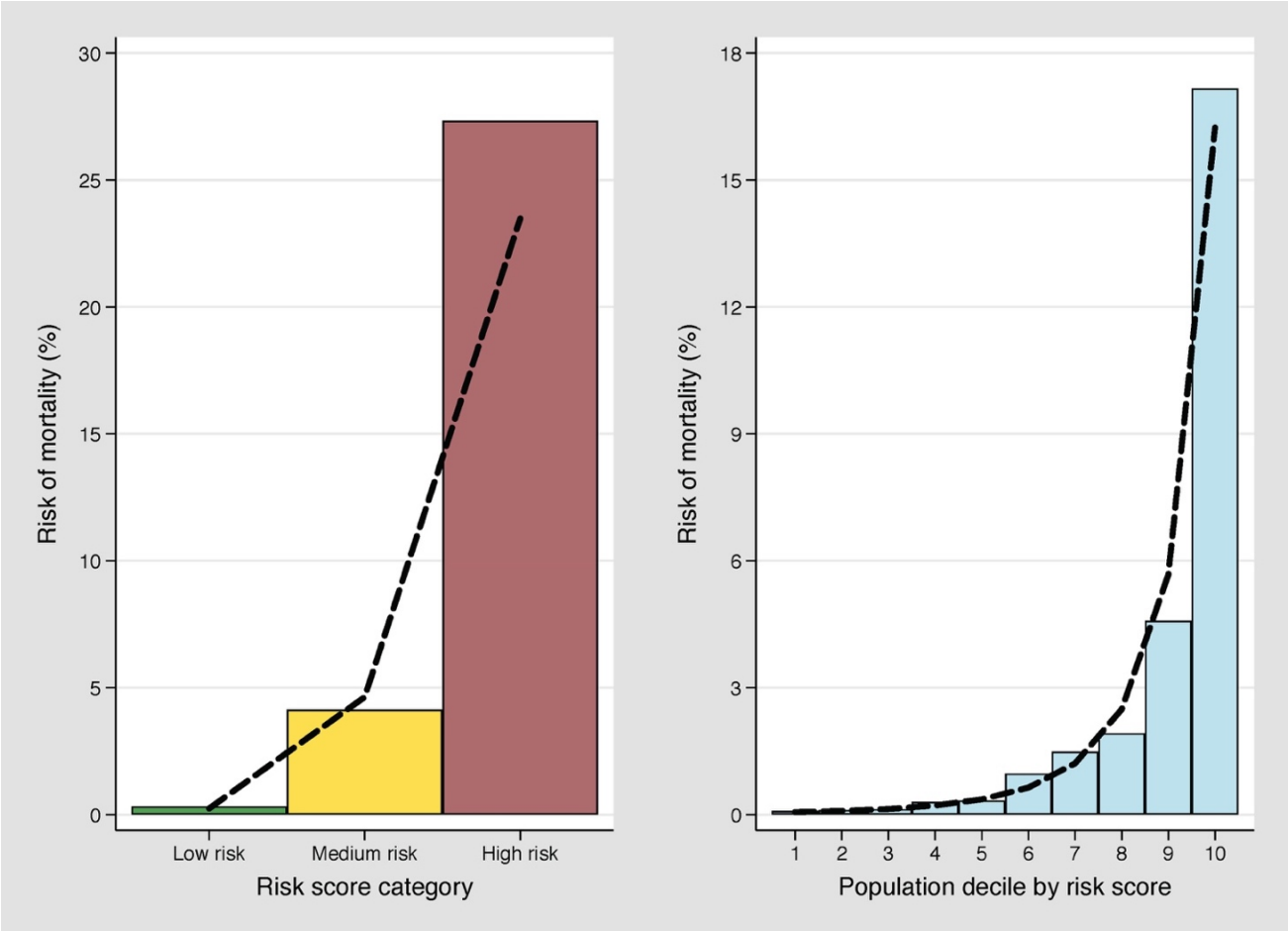

Figure S1. Predicted versus observed risk of death for pre-defined categories and population deciles by risk score in the development sample (n=46108). Predicted risk of death derived from exact regression model (dotted lines) versus observed risk of death (solid bars).

Table S5. Sensitivity and specificity based on predicted mortality risk in the development (n=46108) and full validation samples (n=47846)

	Sensitivity (95% CI)	Specificity (95% CI)	PPV[a] (95% CI)	NPV[b] (95% CI)
Development	NA[c]	NA[c]	NA[c]	NA[c]
0·5%	96·1 (94·8-97·1)	52·9 (52·4-53·3)	5·4 (5·1-5·7)	99·8 (99·7-99·8)
1%	91·9 (90·2-93·4)	64·4 (63·9-64·8)	6·7 (6·3-7·1)	99·7 (99·6-99·7)
3·9%[d]	79·1 (76·7-81·3)	82·9 (82·5-83·2)	11·4 (10·7-12·1)	99·3 (99·2-99·4)
5%	75·3 (72·8-77·7)	85·7 (85·4-86·0)	12·8 (12·0-13·5)	99·2 (99·1-99·3)
10%	60·1 (57·4-62·9)	93·1 (92·9-93·3)	19·5 (18·2-20·8)	98·8 (98·7-98·9)
20%	26·1 (23·7-28·7)	98·4 (98·3-98·5)	31·3 (28·5-34·3)	98·0 (97·8-98·1)
Full validation	NA[c]	NA[c]	NA[c]	NA[c]
0·5%	96·5 (95·5-97·4)	52·0 (51·5-52·4)	6·1 (5·8-6·4)	99·8 (99·7-99·8)
1%	91·7 (90·2-93·1)	63·4 (63·0-63·9)	7·5 (7·1-7·9)	99·6 (99·5-99·7)
3·9%[d]	81·6 (79·6-83·6)	81·0 (80·7-81·4)	12·2 (11·6-12·9)	99·3 (99·2-99·4)
5%	78·4 (76·2-80·4)	83·9 (83·5-84·2)	13·6 (12·9-14·4)	99·2 (99·1-99·3)
10%	65·6 (63·1-68·0)	91·3 (91·1-91·6)	19·7 (18·6-20·9)	98·8 (98·7-98·9)
20%	34·6 (32·2-37·1)	97·7 (97·6-97·8)	32·8 (30·5-35·2)	97·9 (97·7-98·0)

[a] Positive predictive value (PPV).
[b] Negative predictive value (NPV).
[c] Not applicable (NA).
[d] Empirical optimal cutpoint based on the Youden Index.[38]

Table S6. Risk score performance in the external validation samples

	Random validation n=35193	Temporal validation n=12653	Full validation n=47846	Gambian validation n=457
Brier score	0·0272	0·0237	0·0263	0·1715
C-index	0·8910	0·8872	0·8903	0·8082

Table S7. Predicted mortality risk across score percentiles in the Gambian validation sample (n=457)

Percentile	Score	Mean predicted mortality risk[a]	95% Confidence interval[a]
1%	4·2	96·8	94·0 - 99·6
5%	7·0	92·5	88·3 - 96·7
10%	9·0	89·2	84·3 - 94·2
25%	12·5	73·8	67·3 - 80·4
50%	17·0	48·1	42·7 - 53·5
75%	20·0	25·3	20·3 - 30·2
90%	22·0	14·5	10·0 - 18·9

[a] All predictions significant to p-value <0.0001.

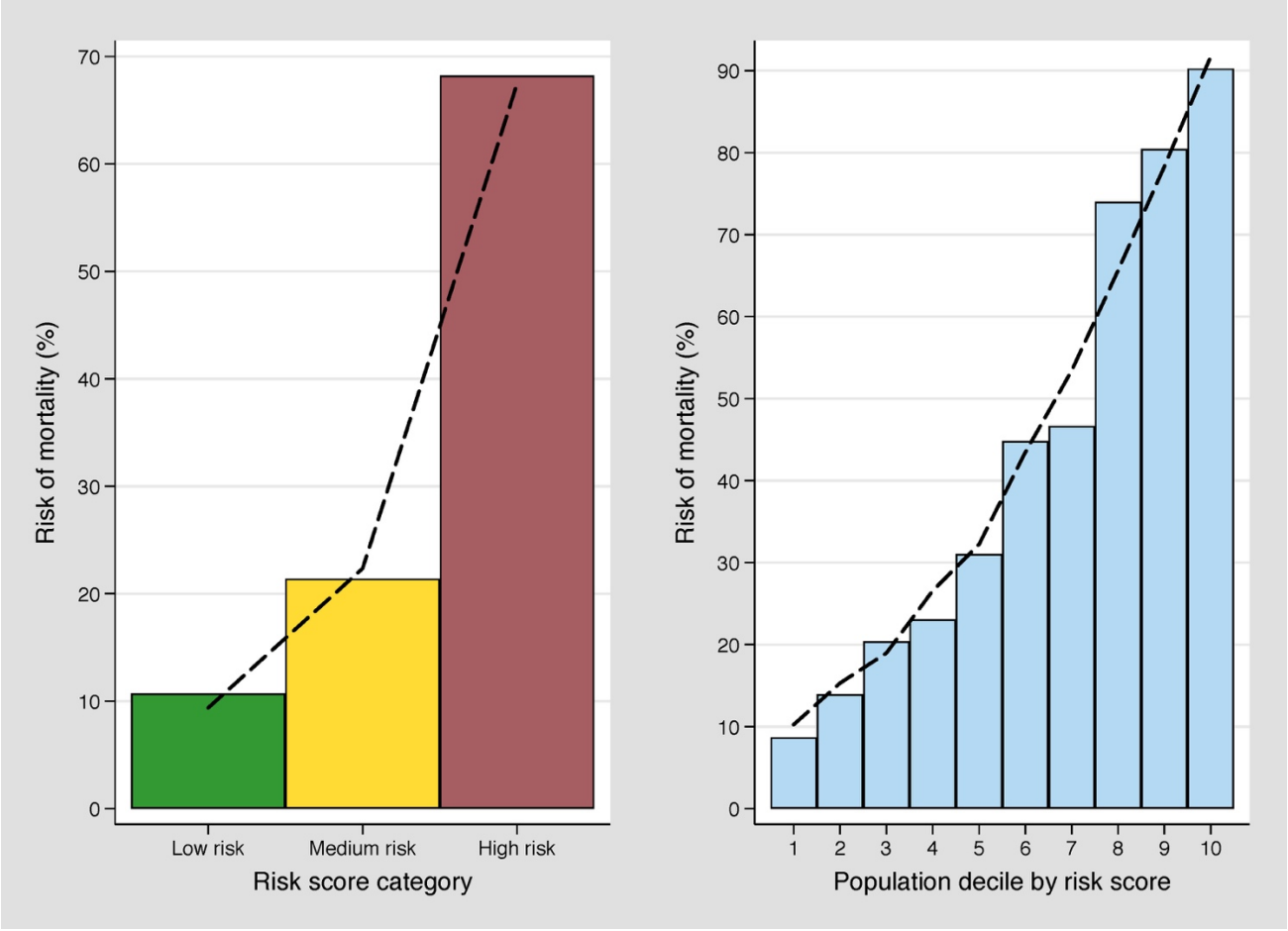

Figure S2. Predicted versus observed risk of death for pre-defined categories and population deciles by risk score in the Gambian validation sample (n=457). Predicted risk of death derived from exact regression model (dotted lines) versus observed risk of death (solid bars).