

A prognostic model and simplified risk score for predicting poor treatment response in children aged 2–59 months hospitalized with severe community-acquired pneumonia: Development and validation study in Hakim Gizaw Hospital, Ethiopia

Mesfin Wudu Kassaw¹

mesfine12a@gmail.com

Asrat Woldeyes Health Science Campus, Debre Berhan University

Tewodrose Mulugeta Nigussie¹

Asrat Woldeyes Health Science Campus, Debre Berhan University

Asaye Worku Agegn

Asrat Woldeyes Health Science Campus, Debre Berhan University

Worku Abemie Seyoum¹

Asrat Woldeyes Health Science Campus, Debre Berhan University

Wegayehu Zeneb Teklehiamanot¹

Asrat Woldeyes Health Science Campus, Debre Berhan University

Bekaghign Girma Negie¹

Asrat Woldeyes Health Science Campus, Debre Berhan University

Yemisrach Tamir Tessema

Asrat Woldeyes Health Science Campus, Debre Berhan University

Selam Tadele Markos¹

Asrat Woldeyes Health Science Campus, Debre Berhan University

Bogale Chekole Fetene¹

Asrat Woldeyes Health Science Campus, Debre Berhan University

Research Article

Keywords: Severe pneumonia, children, response to treatment, prognostic model, risk score, calibration, discrimination, bootstrap validation, decision-curve analysis

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10909348/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

A prognostic model and simplified risk score for predicting poor treatment response in children aged 2–59 months hospitalized with severe community-acquired pneumonia:

Development and validation study in Hakim Gizaw Hospital, Ethiopia

Mesfin Wudu Kassaw¹, Tewodrose Mulugeta Nigussie¹, Asaye Worku Agegn¹, Worku Abemie Seyoum¹, Wegayehu Zeneb Teklehiamanot¹, Bekaghign Girma Negie¹, Yemisrach Tamir Tessema²,
Selam Tadele Markos¹, Bogale Chekole Fetene¹

¹Department of Pediatrics and Child Health Nursing, School of Nursing and Midwifery, Asrat Woldeyes Health Science Campus, Debre Berhan University, Debre Birhan, Ethiopia

²Department of Pediatrics and Child Health, School of Medicine, Asrat Woldeyes Health Science Campus, Debre Berhan University, Debre Birhan, Ethiopia

*Corresponding author: mesfine12a@gmail.com

Abstract

Background: Pneumonia remains a major cause of illness and death among children, particularly in low- and middle-income countries. Severe pneumonia is clinically heterogeneous, and early identification of children at increased risk of poor treatment response may help prioritize monitoring and supportive care. Existing pediatric pneumonia prediction tools have shown variable performance when transported across settings, supporting the development and external validation of locally relevant models.

Methods: We developed and internally validated a multivariable logistic-regression model using data from 630 children aged 2–59 months who were hospitalized with severe pneumonia. The primary outcome was poor response to treatment within 72 hours, defined as prolonged admission beyond 72 hours, death during admission, and/or a change in antibiotics after 72 hours. Four laboratory variables with substantial missingness were imputed using multiple imputation by chained equations with predictive mean matching. Candidate predictors were screened in exploratory univariable analyses, and candidate multivariable models were compared using Akaike's information criterion. The final model contained 16 predictor parameters. Discrimination was assessed using the area under the receiver operating characteristic curve (AUC), and calibration was assessed graphically and using the calibration intercept, slope, maximum absolute error, and Brier score. Internal validation used 1000 bootstrap resamples. A simplified score was derived from selected predictors and assessed for discrimination, calibration, and decision-curve net benefit.

Results: Poor response to treatment occurred in 267 (42.4%) children. The final model included antibiotic treatment at admission, stridor, WBC, cough, oxygen saturation, vaccination status, age,

cyanosis, inability to feed, mental status change, vomiting, and diarrhea. Apparent discrimination was good (AUC 0.875, 95% CI 0.849–0.902). The mean predicted risk was 0.424. Bootstrap validation yielded an optimism-corrected Dxy of 0.713 (95% bootstrap interval 0.659–0.771), corresponding to an approximate corrected AUC of 0.857. The corrected calibration slope was 0.927, and the calibration intercept was 0.021.

The simplified score ranged from 0 to 19 in the observed data. Poor response was 7.7% in the low-risk group (0–3 points), 22.5% in the moderate-risk group (4–7 points), and 70.3% in the high-risk group (≥ 8 points). The score AUC was 0.823 (95% CI 0.791–0.856), lower than that of the full model (DeLong $p < 0.001$). At a classification threshold of ≥ 8 points, sensitivity was 76.4%, specificity 76.3%, positive predictive value 70.3%, negative predictive value 81.5%, and accuracy 76.4%. Bootstrap calibration of the score gave a mean absolute error of 0.010 and mean squared error of 0.00018. Decision-curve analysis showed positive net benefit over the treat-none strategy across a broad range of threshold probabilities.

Conclusions: The full model showed good discrimination and reasonable internal calibration after bootstrap correction. The simplified score retained useful discrimination but performed less well than the regression model. Both instruments require external validation and assessment of clinical impact before they can be considered for routine use.

Keywords: Severe pneumonia; children; response to treatment; prognostic model; risk score; calibration; discrimination; bootstrap validation; decision-curve analysis

Introduction

Pneumonia remains a leading cause of illness and death among children, with a disproportionate burden in low- and middle-income countries [1-3]. Severe pneumonia is clinically heterogeneous, and children who meet the same clinical definition may have substantially different risks of treatment failure and other adverse outcomes [3, 4]. Hypoxaemia, impaired consciousness, feeding difficulty, and other danger signs have been associated with poor outcomes in children with severe pneumonia [4-8].

Clinical prediction models combine multiple prognostic factors to estimate an individual's risk of an outcome. Their value depends not only on discrimination but also on calibration, overall prediction error, and clinical usefulness. Performance can deteriorate when a model is transported to a different population because of differences in case mix, outcome prevalence, clinical management, and measurement practices [9-11].

Several pediatric pneumonia prediction tools have been developed, including RISC and RISC-Malawi. External validation studies have shown that discrimination and calibration can decline

substantially outside the derivation setting [9, 10]. An Ethiopian study also developed and internally validated a model for a composite poor outcome among children with severe community-acquired pneumonia [12]. Its predictor set differs from that of the present study, so direct performance comparisons are not appropriate. The clinical definitions and management context are also informed by WHO guidance for childhood pneumonia [13, 14].

Contemporary guidance recommends transparent reporting of predictor definitions, outcome definitions, model development, internal validation, and performance measures, with particular attention to calibration and the risk of overfitting [15-18]. Bootstrap validation is generally preferable to a simple split-sample approach when developing a prediction model because it uses the available data more efficiently [19].

This study aimed to develop and internally validate a multivariable logistic-regression model for poor response to treatment among children hospitalized with severe pneumonia. We quantified discrimination and calibration, assessed optimism using bootstrap resampling, evaluated the potential clinical utility of a simplified score using decision-curve analysis, and reported the model in a form suitable for external validation.

Methods

Analyses were conducted in RStudio using the mice, pROC, rms, PredictABEL, ResourceSelection, caret, dcurves, gtsummary, and dplyr packages.

We conducted a prognostic model development and internal-validation analysis using inpatient data from Hakim Gizaw Teaching Hospital. The analytic dataset contained 630 children and 33 recorded variables extracted from the electronic medical record. Variables included age, vaccination status, respiratory and gastrointestinal symptoms, neurological findings, vital signs, oxygen saturation, laboratory measurements, antibiotic treatment, oxygen administration, length of stay, and treatment response. Children aged 2–59 months who were hospitalized with severe pneumonia were eligible. Children with severe pneumonia and additional underlying conditions expected to prolong hospitalization were excluded. Children whose diagnosis changed after admission were also excluded.

Poor response to pneumonia treatment was the recorded outcome status in the study data. It was operationalized as prolonged hospitalization beyond 72 hours, death within 72 hours of admission, and/or a change in antibiotics after 72 hours. The outcome was classified as poor response or good response.

Severe pneumonia was defined as cough or difficulty breathing accompanied by at least one danger sign: inability to drink or breastfeed, persistent vomiting, convulsions, lethargy or unconsciousness, stridor in a calm child, or severe malnutrition.

The original sample-size calculation used 10 events per candidate predictor and an anticipated poor-outcome prevalence of 28%, based on the Ethiopian study used as the planning source [12]. With 16 candidate predictors, the calculation yielded 571 participants; allowing 10% for incomplete records gave a target of 629 children with severe pneumonia. The final analytic dataset contained 630 children. Contemporary guidance indicates that sample size for prediction modelling should not be justified by an events-per-variable rule alone; adequacy also depends on outcome prevalence, the number of predictor parameters, anticipated model performance, shrinkage, and precision [17, 18].

The primary outcome was poor response to treatment within 72 hours of admission. Among 630 children, 267 (42.4%) were classified as having poor response and 363 (57.6%) as having good response.

Candidate predictors were selected from variables available at admission or during the initial clinical assessment. The final model included antibiotic treatment at admission, stridor, WBC, cough, oxygen saturation, vaccination status, age, cyanosis, inability to feed, mental status change, vomiting, and diarrhea. Reference categories are defined in the model specification reported in the Results section. Variables occurring after the prediction time point, including length of stay and subsequent antibiotic changes, were not entered as predictors.

The prediction time point was admission or the initial clinical assessment. Predictors were therefore restricted to information plausibly available at that time and were not selected on the basis of subsequent clinical course, consistent with recommendations for prognostic model development [15-18, 20-22].

Missingness was concentrated in four laboratory variables: WBC, 113/630 (17.9%); neutrophils, 116/630 (18.4%); lymphocytes, 113/630 (17.9%); and haemoglobin, 113/630 (17.9%). Missingness in all other variables, including the outcome, was below 5%. Complete-case analysis was not used for the laboratory variables.

Multiple imputation by chained equations was performed using the mice package. Five imputed datasets were generated with 20 iterations, using predictive mean matching for WBC, neutrophils, haemoglobin, and lymphocytes. Convergence and observed-versus-imputed density diagnostics were examined. Regression coefficients, standard errors, and confidence intervals were combined

across the five imputed datasets using Rubin's rules. The pooled imputed analyses were used for model estimation and performance assessment. This approach preserves information that would be lost through complete-case analysis and is consistent with established guidance for missing covariate data [23-27].

Continuous variables were summarized as median (interquartile range) and categorical variables as number (percentage). Wilcoxon rank-sum and Pearson chi-square tests were used for descriptive comparisons. These tests were descriptive and were not used as evidence of predictive performance.

Exploratory univariable logistic-regression models were fitted for candidate predictors. Variables meeting a screening threshold of $p < 0.25$ were entered into the initial multivariable model. Reduced candidate models were then compared using Akaike's information criterion, and the model with the lowest AIC among the sequential candidate models was retained. The final model contained 16 predictor parameters. The selection process was retained as part of the original analysis and is acknowledged as a source of model uncertainty [15-18, 20, 22].

The final model was a multivariable logistic-regression model. The intercept, regression coefficients, and adjusted odds ratios are reported in the Results section. Adjusted odds ratios were obtained by exponentiating the regression coefficients, with 95% confidence intervals derived from the pooled model estimates.

Discrimination was assessed using the receiver operating characteristic curve and area under the curve (AUC), with 95% confidence intervals estimated using the DeLong method [28,29]. The AUC quantifies the ability of the model to rank children with poor response above those with good response [30, 31].

Calibration was assessed using calibration plots, the calibration intercept, calibration slope, maximum absolute calibration error (Emax), and Brier score. A ten-group Hosmer–Lemeshow test was also calculated as a supplementary goodness-of-fit assessment. Calibration measures were interpreted together rather than relying on a single hypothesis test because calibration is a central component of prediction-model performance [15, 16, 32-34].

Internal validation used 1000 bootstrap resamples. In each resample, the model was refitted, and its performance was evaluated in both the bootstrap sample and the original data; the difference was used to estimate optimism. Optimism-corrected performance was obtained by subtracting mean optimism from apparent performance. The validation output included Dxy, R^2 , calibration intercept, calibration slope, Emax, and Brier score. Dxy was converted to an approximate AUC using

$C=[(Dxy+1)/2]$ [19, 21, 30]. Bootstrap validation was selected because it makes efficient use of the available development sample and provides a direct estimate of optimism [19, 21, 35]. The 11 score components and their weights are shown in Table 3, and observed treatment-response rates across score-defined risk groups are shown in Table 4.

A simplified bedside score was derived from the selected predictors. Predictors were recoded according to the categorical coding used in the final model and assigned integer weights across 11 components. The maximum observed score was 19, although the scoring framework permits a theoretical maximum of 22 points.

Scores of 0–3, 4–7, and ≥ 8 were used to describe low-, moderate-, and high-risk groups, respectively. The ≥ 8 value was treated as a classification threshold for descriptive performance assessment, not as a prespecified treatment threshold. Score discrimination, calibration, and decision-curve net benefit were assessed separately from the full regression model.

At a score threshold of ≥ 8 , sensitivity, specificity, positive predictive value, negative predictive value, accuracy, and balanced accuracy were calculated from the confusion matrix. These measures describe classification performance in the development dataset and are expected to vary with outcome prevalence.

Decision-curve analysis was conducted using the *dcurves* package. A logistic model using total score as the predictor generated score-derived probabilities of poor response. Net benefit was evaluated across threshold probabilities and compared with the treat-all and treat-none strategies [34, 36, 37]. Decision-curve analysis was interpreted as an assessment of potential clinical usefulness rather than evidence of clinical effectiveness [34, 36, 37].

Results

Socio-demographic characteristics

The analytic dataset contained 630 children with severe pneumonia: 267 (42.4%) had poor response to treatment and 363 (57.6%) had good response. Median age was 12.0 months (IQR 16.0) overall, 13.0 months (IQR 16.0) among children with good response, and 11.0 months (IQR 14.0) among those with poor response ($p<0.001$). Baseline demographic, clinical, laboratory, and treatment characteristics are presented in Table 1. Convergence and observed-versus-imputed density diagnostics were examined during the imputation process.

Univariable associations

Exploratory analyses identified differences in several candidate predictors between outcome groups, including diarrhea, vomiting, vaccination status, cyanosis, mental status change, inability to feed, stridor, cough, oxygen saturation, and antibiotic treatment initiated at admission. These comparisons were used for model-building screening and should not be interpreted as independent prognostic effects. The categorical definitions used in the final model were complete versus incomplete vaccination; age 2–11, 12–23, 24–35, 36–47, and 48–59 months; low versus very low oxygen saturation; and vancomycin-containing, ceftriaxone-containing, and ampicillin-gentamicin-containing antibiotic regimens at admission.

Final model coefficients

In the final model, incomplete vaccination was associated with higher odds of poor response (adjusted OR 3.79, 95% CI 2.45–5.96), very low oxygen saturation with approximately twice the odds (AOR 2.23, 95% CI 1.42–3.53), and age 12–23 months with approximately threefold higher odds than age 2–11 months (AOR 3.29, 95% CI 1.72–6.43). Ampicillin-gentamicin-containing treatment at admission was associated with higher odds than vancomycin-containing treatment (AOR 2.31, 95% CI 1.22–4.47). Other prominent associations were observed for increased WBC, diarrhea, cyanosis, vomiting, mental status change, inability to feed, and stridor. Cough remained in the model despite weaker evidence after adjustment ($p=0.092$), reflecting the model-selection process rather than an absence of clinical relevance. Complete adjusted model estimates are presented in Table 2.

Predicted risk distribution and discrimination

The final model generated predicted probabilities of poor response ranging from 0.008 to 0.999; the median was 0.364, the mean was 0.424, and the upper quartile was 0.743. The mean predicted probability matched the observed event proportion of 0.424, as expected in the development sample. The ROC curve showed an apparent AUC of 0.875 (95% CI 0.849–0.902), indicating good discrimination in the development dataset (Figure 1). This is an apparent performance estimate and may be optimistic. The corresponding Dxy was 0.737.

Calibration

The ten-group Hosmer–Lemeshow test yielded $p=0.374$. The calibration plot (Figure 2) showed generally close agreement between predicted and observed risks, with some local departures at intermediate and higher predicted probabilities. The supplementary table (Supplementary Table 3) provides predicted and observed event frequencies across ten groups of predicted risk.

Bootstrap validation

After 1000 bootstrap resamples, the apparent Dxy of 0.737 was reduced by an optimism estimate of 0.024 to a corrected Dxy of 0.713 (95% bootstrap interval 0.659–0.771), corresponding to an approximate corrected AUC of 0.857. Corrected R^2 was 0.471, the calibration intercept was -0.021 , the calibration slope was 0.927, Emax was 0.038, and the Brier score was 0.153. Supplementary Table 4 summarizes the apparent, bootstrap training and test performance, estimated optimism, and optimism-corrected performance.

Bootstrap calibration yielded a mean absolute error of 0.012, a mean squared error of 0.0002, and a 90th-percentile absolute error of 0.027. The corresponding bootstrap-corrected calibration curve is shown in Figure 3.

Simplified risk score

The score ranged from 0 to 19 in the observed dataset. The most frequent scores were 4 (n=70), 6 (n=72), 7 (n=63), 9 (n=67), and 5 (n=44); only one participant had a score of 19, and no participants had scores of 20–22 (Supplementary Table 5).

The total score was calculated as: total score = $2 \times (\text{ampicillin-gentamicin-containing treatment at admission}) + 2 \times (\text{stridor}) + 2 \times (\text{increased WBC}) + 2 \times (\text{very low oxygen saturation}) + 3 \times (\text{incomplete vaccination}) + 2 \times (\text{age 12–23 months}) + 2 \times (\text{cyanosis}) + 1 \times (\text{inability to feed}) + 2 \times (\text{mental status change}) + 2 \times (\text{vomiting}) + 2 \times (\text{diarrhea})$.

Poor response increased from 7.7% in the low-risk group to 22.5% in the moderate-risk group and 70.3% in the high-risk group. This gradient supports the score as a candidate stratification tool, but the observed risks are specific to this development cohort (Table 3).

Score discrimination and comparison with the full model

The simplified score had an AUC of 0.8235 (95% CI 0.791–0.856). The full model therefore had higher discrimination, with an AUC difference of approximately 0.052; the DeLong test indicated a statistically significant difference ($p < 0.001$). Supplementary Figure 1 shows the ROC curve for the full model as the comparator, while Supplementary Figure 2 shows the ROC curve for the simplified score.

The two-by-two classification matrix at the ≥ 8 -point threshold is shown in Table 4; derived classification measures are provided in Supplementary Table 6.

Classification performance at score ≥ 8

At the ≥ 8 -point classification threshold, 204 of 267 children with poor response were classified as high risk, and 86 of 363 children with good response were classified as high risk. Sensitivity was 76.4%, specificity 76.3%, positive predictive value 70.3%, and negative predictive value 81.5%. Because these measures depend on outcome prevalence and the classification threshold was derived and assessed in the development sample, they should not be interpreted as transportable clinical performance estimates.

Score calibration

Bootstrap calibration of the score-derived probability model produced a mean absolute error of 0.010, a mean squared error of 0.00018, and a 90th-percentile absolute error of 0.021. The corresponding bootstrap calibration curve for the simplified score-derived probabilities is shown in Figure 4.

Decision curve analysis

Decision-curve analysis of the score-derived probabilities (Figure 5) showed positive net benefit compared with the treat-none strategy across a broad range of threshold probabilities. The treat-all strategy was less favourable over much of the displayed range, while net benefit approached zero at high thresholds. This result indicates potential usefulness for risk stratification but does not demonstrate clinical effectiveness.

Because the decision threshold and intended clinical action were not prespecified, the decision-curve analysis is best interpreted as exploratory evidence of potential utility rather than proof of clinical effectiveness [34, 36, 37].

Discussion

Among 630 children with severe pneumonia, 267 (42.4%) experienced poor response to treatment. The final multivariable model showed good apparent discrimination (AUC 0.875) and acceptable internal calibration. Bootstrap correction reduced Dxy from 0.737 to 0.713, corresponding to an approximate AUC of 0.857; the corrected calibration slope was 0.927 and the calibration intercept was -0.021 . The simplified score retained moderate-to-good discrimination (AUC 0.823) but was less discriminative than the full model.

The predictors retained in the final model are clinically plausible. Hypoxaemia and cyanosis reflect impaired oxygenation, while stridor, inability to feed, and altered mental status indicate severe physiological compromise. These features have been associated with adverse outcomes in pediatric pneumonia and support the clinical relevance of the model's predictor set [3-5, 13, 14]. Incomplete

vaccination was associated with higher odds of poor response (AOR 3.79, 95% CI 2.45–5.96), very low oxygen saturation with approximately twice the odds (AOR 2.23, 95% CI 1.42–3.53), and age 12–23 months with approximately threefold higher odds than age 2–11 months (AOR 3.29, 95% CI 1.72–6.43). Ampicillin-gentamicin-containing treatment at admission was associated with higher odds than vancomycin-containing treatment (AOR 2.31, 95% CI 1.22–4.47).

Diarrhea and vomiting were also associated with poor response to treatment. These findings may reflect dehydration, systemic illness, or coexisting illness, but the observational design does not permit causal interpretation. The association with WBC may reflect systemic inflammation or disease severity; it should be interpreted as a prognostic association rather than as evidence of a mechanistic biomarker.

The apparent AUC indicates good ranking performance within the development cohort, but it should not be considered sufficient evidence for implementation. Calibration was evaluated using several complementary measures: the corrected calibration slope was close to 1, the corrected intercept was close to 0, and the corrected Emax was 0.038. The optimism-corrected Brier score was 0.153. These findings indicate useful prognostic information with residual prediction error; calibration should be considered alongside discrimination when judging model performance [15, 16, 33, 34].

Bootstrap resampling estimates the optimism introduced when a model is developed and evaluated in the same data and is generally more efficient than random sample splitting [19, 21, 35]. The relatively small reduction in discrimination after bootstrap correction is reassuring, but internal validation cannot establish transportability to other hospitals or populations [11, 30, 38, 39].

The model included 16 predictor parameters and 267 events (poor response to treatment). Sample-size adequacy should not be judged by the traditional 10-events-per-variable heuristic alone. Modern guidance considers the number of predictor parameters, outcome prevalence, anticipated model performance, shrinkage, and precision [17, 40, 41]. The use of data-driven screening and AIC-based selection introduces additional model uncertainty that may not be fully captured by the reported bootstrap procedure [20–22, 35].

The simplified score provides a more accessible representation of the model but loses information through categorisation and integer weighting. Its AUC of 0.823 was lower than the full model's AUC of 0.875, and the difference was statistically significant. The score should therefore be regarded as a pragmatic secondary model rather than an equivalent substitute for the regression equation [20, 35].

The score separated children into groups with markedly different observed event rates: 7.7%, 22.5%, and 70.3% across the low-, moderate-, and high-risk groups, respectively. The ≥ 8 threshold produced balanced sensitivity and specificity of approximately 76% in this cohort. Because the threshold was derived and assessed in the same dataset, these estimates may be optimistic and will vary with outcome prevalence.

Decision-curve analysis suggested positive net benefit for the score across a broad range of threshold probabilities. However, net benefit depends on the relative consequences of false-positive and false-negative decisions and on the clinical action triggered by a positive prediction [34, 36, 37]. No prospective impact evaluation was performed; therefore, the analysis does not establish that using the score improves patient outcomes or resource allocation [38, 39].

The model should be interpreted in the context of existing pediatric pneumonia tools. RISC and RISC-Malawi have demonstrated useful discrimination in derivation settings, but external validation has shown attenuation of performance across populations [9, 10]. The Ethiopian model by Anteneh and colleagues addressed a similar composite outcome using different methods, study designs, and eligibility criteria and therefore provides context rather than a direct benchmark [12]. The present model's apparent AUC should not be compared numerically with those studies without accounting for differences in outcome definition, predictor availability, case mix, and validation design. The broader prediction-model literature also emphasizes the need for transparent reporting, appropriate validation, and assessment of transportability before implementation [15, 16, 38, 39, 42-44].

The study has several strengths: a clearly defined analytic cohort of 630 children; explicit reporting of the full regression coefficients; pooled analysis across five imputed datasets; assessment of discrimination, calibration, and prediction error; bootstrap internal validation; and separate evaluation of a simplified score and its potential clinical utility.

Important limitations include the single-centre source, absence of external validation, data-driven predictor screening and AIC-based model selection, categorisation of predictors, and reliance on a composite operational outcome. Although multiple imputation was used appropriately across five datasets, the model-selection strategy and categorisation of predictors can still limit transportability. External validation should therefore examine whether predictor effects and calibration are maintained in other hospitals and populations. The outcome is a composite operational measure of poor response and should not be interpreted as a mortality endpoint.

External validation in an independent cohort should be the next step, with evaluation of discrimination, the calibration intercept and slope, Brier score, and clinical utility. Contemporary guidance recommends assessing calibration-in-the-large and calibration slope, considering recalibration or model updating when transportability is imperfect, and conducting prospective impact assessment before implementation [38-44].

In conclusion, the model provides a reproducible estimate of poor response to treatment among children hospitalized with severe pneumonia and showed good apparent discrimination with limited optimism after bootstrap correction. The simplified score retained useful discrimination but was less accurate than the full model. The findings support external validation and prospective impact assessment rather than immediate clinical adoption.

In conclusion, this 16-parameter logistic-regression model developed from 630 children showed good apparent discrimination and acceptable internal calibration for poor response to treatment. The 11-component score retained useful discrimination and separated children into groups with substantially different observed event rates, although it performed less well than the full model. Considering the limitations, external validation is required before either model is used to guide clinical decisions.

Ethics approval and consent to participate

The ethical clearance for this specific secondary analysis was waived by the institutional review board committee of Debre Berhan University, Asrat Woldeyes Health Science Campus with a reference number of AWHSC/IRB-20567/2026. An official permission letter was received from Asrat Woldeyes Health Science Campus's teaching hospital, Hakim Gizaw Hospital. All the data were de-identified, and more importantly, all study methodology was performed in accordance with the ethical principles of the Declaration of Helsinki.

Data sharing

The analytical dataset is available from the principal investigator upon reasonable request.

Funding

No funding was received for this study.

Competing interests

The authors declare that they have no competing interests.

Author contributions

MWK, TMN, AW, WA, WZT, BGN, YTT, STM, and BCF conceptualized the study, extracted and processed the data, conducted the statistical analyses, interpreted the findings, and drafted the manuscript. All authors reviewed and revised the manuscript, approved the final version, and agreed to be accountable for all aspects of the work.

Acknowledgments

We thank Asrat Woldeyes Health Science Campus and Hakim Gizaw Teaching Hospital for allowing access to the study data. We also acknowledge the clinical and data-management teams who supported the underlying records and data extraction.

References

1. World Health Organization. Pneumonia; Geneva: World Health Organization; 2024.
2. McAllister DA, Liu L, Shi T, Chu Y, Reed C, Burrows J, et al. Global, regional, and national estimates of pneumonia morbidity and mortality in children younger than 5 years between 2000 and 2015: a systematic analysis. *Lancet Glob Health*; 2019; 7:e47-e57.
3. O'Brien KL, Baggett HC, Brooks WA, Deloria Knoll M, Saha SK, Abdullah Brooks M, et al; PERCH Study Group. Causes of severe pneumonia requiring hospital admission in children without HIV infection from Africa and Asia: the PERCH multi-country case-control study. *Lancet*; 2019; 394:757-779.
4. Chisti MJ, Tebruegge M, La Vincente S, Graham SM, Duke T. Pneumonia in severely malnourished children in developing countries-mortality risk, aetiology and validity of WHO clinical signs: a systematic review. *Trop Med Int Health*; 2009; 14:1173-1189.
5. Subhi R, Smith K, Duke T, et al. Hypoxaemia as a mortality risk factor in acute lower respiratory infections in children in low and middle-income countries: systematic review and meta-analysis. *PLOS-One*. 2015; 10:e0136160.
6. O'Brien KL, Wolfson LJ, Watt JP, Henkle E, Deloria-Knoll M, McCall N, et al. Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. *Lancet*. 2009;374:893-902.
7. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis. *Lancet*. 2020;396:1204-1222.
8. Black RE, Cousens S, Johnson HL, Lawn JE, Rudan I, Bassani DG, et al. Global, regional, and national causes of child mortality in 2008: a systematic analysis. *Lancet*. 2010;375:1969-1987.

9. Hooli S, Colbourn T, Lufesi N, Costello A, McCollum ED, King C. Predicting hospitalised paediatric pneumonia mortality risk: external validation of RISC and mRISC, and local tool development (RISC-Malawi) from Malawi. *PLoS One*. 2016;11:e0168126.
10. Rees CA, Hooli S, King C, et al. External validation of the RISC, RISC-Malawi, and PERCH clinical prediction rules to identify risk of death in children hospitalized with pneumonia. *J Glob Health*. 2021;11:04062.
11. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*. 2015;68:25-34.
12. Anteneh ZA, Arega HE, Mihretie KM. Validation of risk prediction for outcomes of severe community-acquired pneumonia among under-five children in Amhara region, Northwest Ethiopia. *PLoS One*. 2023;18:e0281209.
13. World Health Organization. Revised WHO classification and treatment of childhood pneumonia at health facilities: evidence summaries. Geneva: World Health Organization; 2014.
14. World Health Organization. Pocket Book of Hospital Care for Children: Guidelines for the Management of Common Childhood Illnesses. 3rd ed. Geneva: World Health Organization; 2022.
15. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162:55-63.
16. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
17. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441.
18. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al; PROBAST Group. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019; 170:51-58.
19. Steyerberg EW, Harrell FE Jr, Borsboom GJJM, Eijkemans MJ, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiology*; 2001; 54:774-781.
20. Steyerberg EW. Clinical Prediction Models; 2nd Ed. New York: Springer; 2019.

21. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014; 35:1925-1931.
22. Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart*; 2012; 98:683-690.
23. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011; 30:377-399.
24. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009; 338:b2393.
25. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011; 45:1-67.
26. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies with multiple imputation. *Stat Med*. 2009; 28:1558-1570.
27. Janssen KJM, Donders ART, Harrell FE Jr, Vergouwe Y, Chen Q, Grobbee DE, et al. Missing covariate data in medical research: to impute is better than to ignore. *Arch Intern Med*. 2010; 170:453-460.
28. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves. a nonparametric approach; *Biometrics*; 1988; 44:837-845.
29. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve; *Radiology*; 1982; 143:29-36.
30. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol*; 2013; 13:33.
31. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*: Oxford: Oxford University Press; 2003.
32. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials: *Stat Med*; 1986; 5:421-433.
33. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019; 17:230.

34. Van Calster B, Wynants L, Verbeek JFM, Verbakel JY, Christodoulou E, Vickers AJ, et al. Reporting and interpreting decision curve analysis: a guide for prediction model development and validation. *Eur Urol*; 2018; 74:796-804.
35. Harrell FE Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. 2nd ed. Cham: Springer; 2015.
36. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*; 2006; 26:565-574.
37. Vickers AJ, Van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis; *Diagn Progn Res*; 2019; 3:18.
38. Riley RD, Archer L, Snell KIE, Ensor J, Debray TPA, Collins GS, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ*; 2024; 384:e074820.
39. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*; 2012; 98:691-698.
40. van Smeden M, Moons KGM, de Groot JAH, Collins GS, Altman DG, Eijkemans MJC, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat Methods Med Res*; 2019; 28:2455-2474.
41. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, van Smeden M, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med*; 2021; 40:4230-4250.
42. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ*; 2020; 369:m1328.
43. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance: *BMJ*; 2017; 356:i6460.
44. Riley RD, Archer L, Snell KIE, Ensor J, Debray TPA, Collins GS, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*; 2024; 384:e074819.

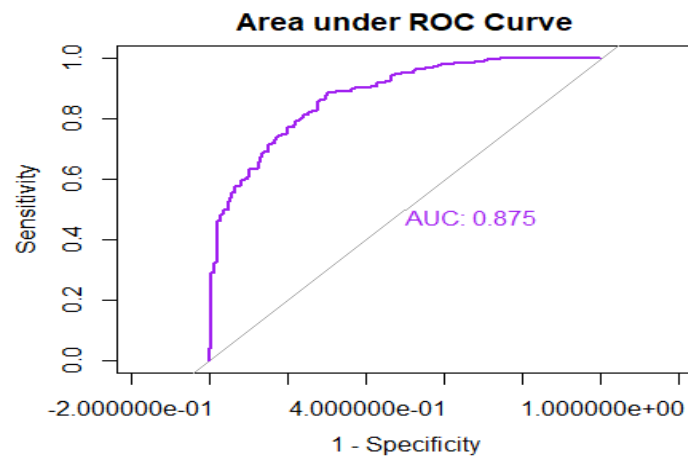


Figure 1: ROC curve for the final prognostic model for poor treatment response status in children with severe pneumonia hospitalized in Hakim Gizaw Hospital, Debre Birhan, Ethiopia

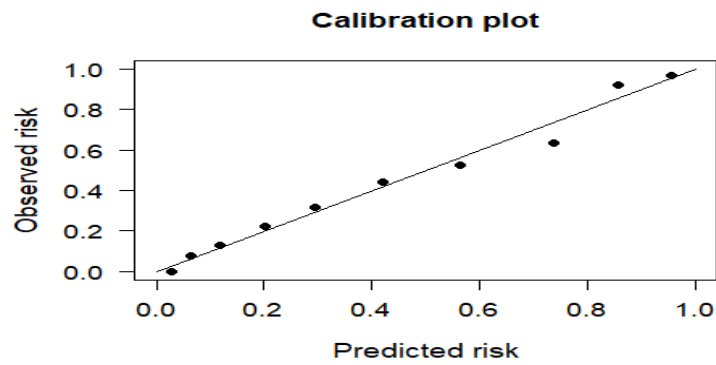


Figure 2: Calibration plot for the final prognostic model for poor treatment response status in children with severe pneumonia hospitalized in Hakim Gizaw Hospital, Debre Birhan, Ethiopia

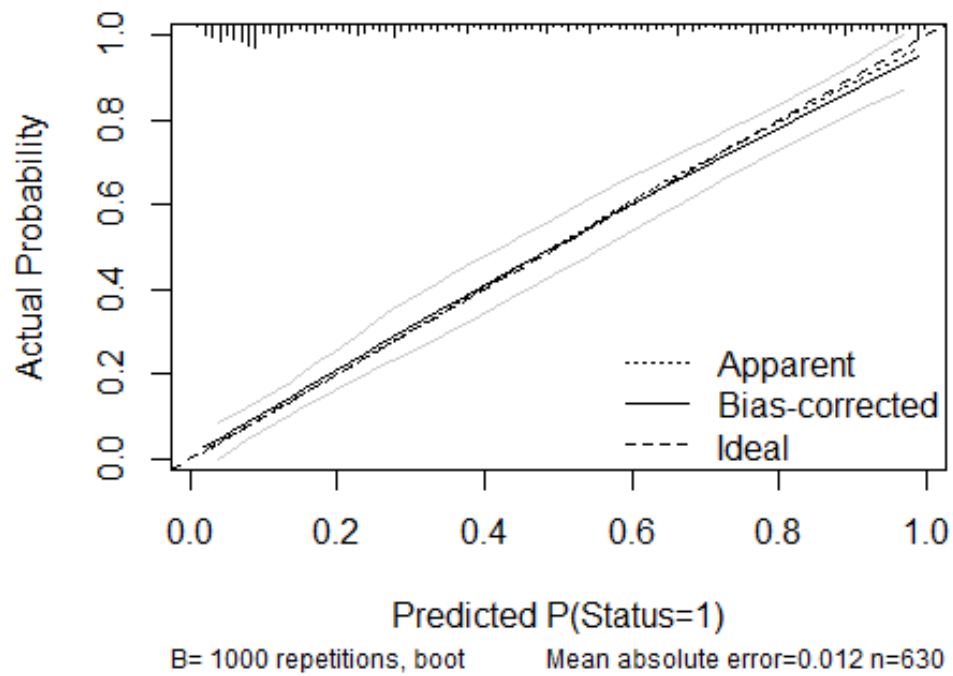


Figure 3: Bootstrap-corrected calibration of the final prognostic model for poor treatment response status in children with severe pneumonia hospitalized in Hakim Gizaw Hospital, Debre Birhan, Ethiopia

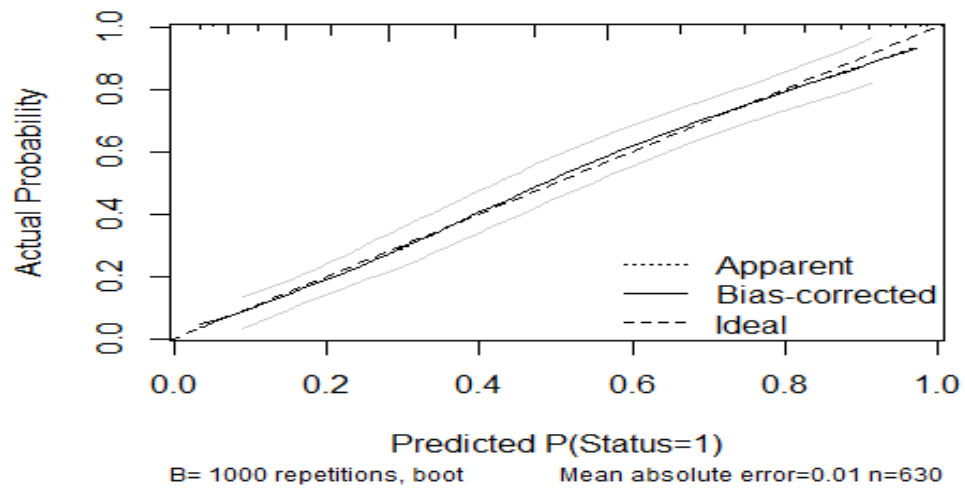


Figure 4: Bootstrap calibration of the simplified score for poor treatment response status in children with severe pneumonia hospitalized in Hakim Gizaw Hospital, Debre Birhan, Ethiopia

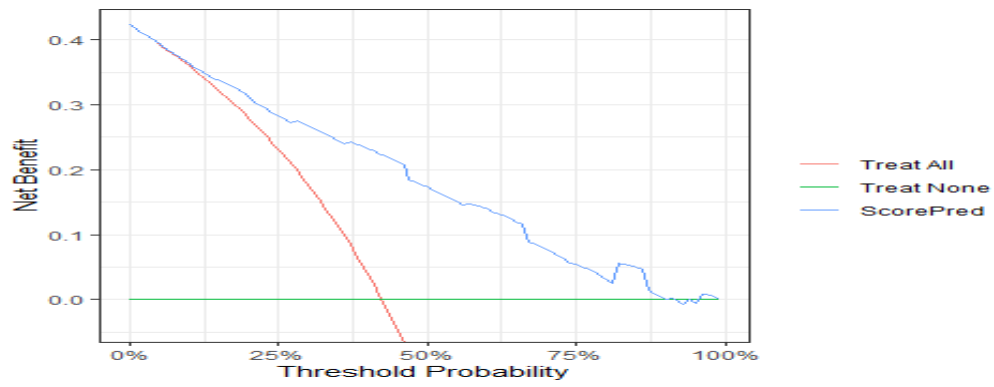


Figure 5: Decision-curve analysis of the simplified score for poor treatment response status in children with severe pneumonia hospitalized in Hakim Gizaw Hospital, Debre Birhan, Ethiopia

Table1. Baseline characteristics of children with severe pneumonia by treatment-response status hospitalized in Hakim Gizaw Hospital, Debre Birhan, Ethiopia

Characteristic	Total N = 630	Good response N = 363	Poor response N = 267	P-value
Age(months)	12.0 (16.0)	13.0 (16.0)	11.0 (14.0)	<0.001
Antibiotic change				<0.001
No	538.0 (85.4%)	346.0 (95.3%)	192.0 (71.9%)	
Yes	92.0 (14.6%)	17.0 (4.7%)	75.0 (28.1%)	
Analgesic				0.060
Yes	397.0 (63.0%)	240.0 (66.1%)	157.0 (58.8%)	
No	233.0 (37.0%)	123.0 (33.9%)	110.0 (41.2%)	
Neutrophil	55.7 (30.1)	56.7 (29.6)	54.4 (30.7)	0.7
Pulse rate (beats/min)	152.0 (29.0)	152.0 (30.0)	153.0 (28.0)	0.2
Stridor				<0.001
No	364.0 (57.8%)	244.0 (67.2%)	120.0 (44.9%)	
Yes	266.0 (42.2%)	119.0 (32.8%)	147.0 (55.1%)	
Mental status change				<0.001
No	441.0 (70.0%)	284.0 (78.2%)	157.0 (58.8%)	
Yes	189.0 (30.0%)	79.0 (21.8%)	110.0 (41.2%)	
Diarrhea				<0.001
No	372.0 (59.0%)	268.0 (73.8%)	104.0 (39.0%)	
Yes	258.0 (41.0%)	95.0 (26.2%)	163.0 (61.0%)	
Antibiotics at admission				<0.001
Vancomycin-containing	94.0 (14.9%)	63.0 (17.4%)	31.0 (11.6%)	
Ceftriaxone-containing	259.0 (41.1%)	167.0 (46.0%)	92.0 (34.5%)	
Ampicillin-gentamicin-containing	277.0 (44.0%)	133.0 (36.6%)	144.0 (53.9%)	
WBC	10.6 (8.6)	9.6 (6.5)	12.6 (13.8)	<0.001
Respiratory rate	56.0 (24.8)	54.0 (26.0)	58.0 (23.0)	0.054
Convulsions				0.3
No	521.0 (82.7%)	305.0 (84.0%)	216.0 (80.9%)	
Yes	109.0 (17.3%)	58.0 (16.0%)	51.0 (19.1%)	
Retractions				0.8
No	220.0 (34.9%)	128.0 (35.3%)	92.0 (34.5%)	
Yes	410.0 (65.1%)	235.0 (64.7%)	175.0 (65.5%)	
Vomiting				<0.001
No	373.0 (59.2%)	251.0 (69.1%)	122.0 (45.7%)	
Yes	257.0 (40.8%)	112.0 (30.9%)	145.0 (54.3%)	
Sex				0.6
Female	255.0 (40.5%)	144.0 (39.7%)	111.0 (41.6%)	
Male	375.0 (59.5%)	219.0 (60.3%)	156.0 (58.4%)	
Length of stay (hours)	3.0 (3.0)	2.0 (1.1)	5.0 (3.0)	<0.001
Hemoglobin	11.1 (2.6)	11.3 (2.2)	10.8 (3.0)	0.087
Oxygen saturation				<0.001
Low (80–90%)	423.0 (67.1%)	275.0 (75.8%)	148.0 (55.4%)	

Characteristic	Total N = 630	Good response N = 363	Poor response N = 267	P-value
Very low (<80%)	207.0 (32.9%)	88.0 (24.2%)	119.0 (44.6%)	
Temperature (°C)	37.5 (1.5)	37.5 (1.5)	37.5 (1.5)	0.7
Lethargy				0.9
No	615.0 (97.6%)	354.0 (97.5%)	261.0 (97.8%)	
Yes	15.0 (2.4%)	9.0 (2.5%)	6.0 (2.2%)	
Cyanosis				<0.001
No	475.0 (75.4%)	315.0 (86.8%)	160.0 (59.9%)	
Yes	155.0 (24.6%)	48.0 (13.2%)	107.0 (40.1%)	
Vaccination				<0.001
Complete	408.0 (64.8%)	276.0 (76.0%)	132.0 (49.4%)	
Incomplete	222.0 (35.2%)	87.0 (24.0%)	135.0 (50.6%)	
Oxygen given (L/min)				0.008
0	92.0 (14.6%)	65.0 (17.9%)	27.0 (10.1%)	
0.5	169.0 (26.8%)	104.0 (28.7%)	65.0 (24.3%)	
1	323.0 (51.3%)	172.0 (47.4%)	151.0 (56.6%)	
2	46.0 (7.3%)	22.0 (6.1%)	24.0 (9.0%)	
Lymphocyte	37.9 (27.1)	36.9 (26.1)	39.2 (29.5)	0.6
Oxygen saturation	84.0 (10.0)	84.0 (7.0)	83.0 (13.0)	0.015
Grunting				0.6
No	227.0 (36.0%)	128.0 (35.3%)	99.0 (37.1%)	
Yes	403.0 (64.0%)	235.0 (64.7%)	168.0 (62.9%)	
Inability to feed				<0.001
No	386.0 (61.3%)	256.0 (70.5%)	130.0 (48.7%)	
Yes	244.0 (38.7%)	107.0 (29.5%)	137.0 (51.3%)	
Cough				<0.001
No	73.0 (11.6%)	57.0 (15.7%)	16.0 (6.0%)	
Yes	557.0 (88.4%)	306.0 (84.3%)	251.0 (94.0%)	
Age (months)				<0.001
2-11	141.0 (22.4%)	89.0 (24.5%)	52.0 (19.5%)	
12-23	131.0 (20.8%)	51.0 (14.0%)	80.0 (30.0%)	
24-35	96.0 (15.2%)	62.0 (17.1%)	34.0 (12.7%)	
36-47	88.0 (14.0%)	53.0 (14.6%)	35.0 (13.1%)	
48-59	174.0 (27.6%)	108.0 (29.8%)	66.0 (24.7%)	

Table2. Final pooled multivariable model for poor treatment response *status in children with severe pneumonia hospitalized in Hakim Gizaw Hospital, Debre Birhan, Ethiopia*

Predictor	β	Adjusted OR	95% CI	P-value
Antibiotics at admission, vancomycin-containing	-	-	-	-
Antibiotics at admission, ceftriaxone-containing	0.352	1.42	0.73–2.80	0.302
Antibiotics at admission, ampicillin-gentamicin-containing	0.839	2.31	1.22–4.47	0.011
Stridor, no	-	-	-	-
Stridor, yes	0.820	2.27	1.48–3.51	<0.001
Within normal range WBC	-	-	-	-
Increase WBC	1.247	3.48	2.04–5.96	<0.001
Cough, no	-	-	-	-
Cough, yes	0.685	1.98	0.92–4.53	0.092
Oxygen-saturation, low	-	-	-	-
Oxygen-saturation, very low	0.804	2.23	1.42–3.53	<0.001
Vaccination, complete	-	-	-	-
Vaccination, incomplete	1.333	3.79	2.45–5.96	<0.001
2–11 months	-	-	-	-
12–23 months	1.192	3.29	1.72–6.43	<0.001
24–35 months	-0.024	0.98	0.48–2.00	0.948
36–47 months	0.026	1.03	0.49–2.14	0.944
48–59 months	0.098	1.10	0.60–2.03	0.753
Cyanosis, no	-	-	-	-
Cyanosis, yes	1.136	3.11	1.90–5.19	<0.001
Inability to feed, no	-	-	-	-
Inability to feed, yes	0.713	2.04	1.33–3.16	0.001
Mental status change, no	-	-	-	-
Mental status change, yes	0.784	2.19	1.37–3.53	0.001
Vomiting, no	-	-	-	-
Vomiting, yes	0.882	2.42	1.57–3.75	<0.001
Diarrhea, no	-	-	-	-
Diarrhea, yes	1.244	3.47	2.27–5.36	<0.001

Table3. Treatment-response rates by simplified score risk group for treatment-response status in children with severe pneumonia hospitalized in Hakim Gizaw Hospital, Debre Birhan, Ethiopia

Risk group	Score	Total	Good response	Poor response	Poor response (%)
Low	0–3	91	84	7	7.7%
Moderate	4–7	249	193	56	22.5%
High	≥ 8	290	86	204	70.3%
Total	0–19	630	363	267	42.4%

Table 4: Classification matrix for the simplified score at ≥ 8 points for treatment-response status in children with severe pneumonia hospitalized in Hakim Gizaw Hospital, Debre Birhan, Ethiopia

Score classification	Good response	Poor response
Predicted low risk	277	63
Predicted high risk	86	204

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [BMCMedicineSupplementary.docx](#)