Supplemental Table 1

TRIPOD Checklist: Prediction Model Development



|  |  |  |  |
| --- | --- | --- | --- |
| **Section/Topic** | **Item** | **Checklist Item** | **Page** |
| **Title and abstract** |  |  |  |
| Title | 1 | Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted. | 1 |
| Abstract | 2 | Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions. | 2 |
| **Introduction** |  |  |  |
| Background and objectives | 3a | Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models. | 3 |
| 3b | Specify the objectives, including whether the study describes the development or validation of the model or both. | 3 |
| **Methods** |  |  |  |
| Source of data | 4a | Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable. | 4 |
| 4b | Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up. | 4 |
| Participants | 5a | Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres. | 4 |
| 5b | Describe eligibility criteria for participants. | 4 |
| 5c | Give details of treatments received, if relevant. | N/A |
| Outcome | 6a | Clearly define the outcome that is predicted by the prediction model, including how and when assessed. | 4 |
| 6b | Report any actions to blind assessment of the outcome to be predicted. | 6 |
| Predictors | 7a | Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured. | 5 |
| 7b | Report any actions to blind assessment of predictors for the outcome and other predictors. | 6 |
| Sample size | 8 | Explain how the study size was arrived at. | 7 |
| Missing data | 9 | Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method. | 7 |
| Statistical analysis methods | 10a | Describe how predictors were handled in the analyses. | 6 |
| 10b | Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation. | 6 |
| 10d | Specify all measures used to assess model performance and, if relevant, to compare multiple models. | 6 |
| Risk groups | 11 | Provide details on how risk groups were created, if done. | N/A |
| **Results** |  |  |  |
| Participants | 13a | Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful. | 7 |
| 13b | Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome. | 7 |
| Model development | 14a | Specify the number of participants and outcome events in each analysis. | 7 |
| 14b | If done, report the unadjusted association between each candidate predictor and outcome. | N/A |
| Model  specification | 15a | Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point). | 7 |
| 15b | Explain how to the use the prediction model. | 7 |
| Model performance | 16 | Report performance measures (with CIs) for the prediction model. | 7 |
| **Discussion** |  |  |  |
| Limitations | 18 | Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data). | 10 |
| Interpretation | 19b | Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence. | 9 |
| Implications | 20 | Discuss the potential clinical use of the model and implications for future research. | 11 |
| **Other information** |  |  |  |
| Supplementary information | 21 | Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets. | 11 |
| Funding | 22 | Give the source of funding and the role of the funders for the present study. | 11 |

We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.

Supplementary Table 2

Table S2: Rank of variable importance by reduction in residual sum of squares (RSS) using random forest regression.

Bacterial Etiology

Variable Name RSS Reduction

Age 73.9

Blood in stool 33.9 Breastfed 29.1

HAZ 25.9 MUAC 24.3

Resp. Rate 20.6 Season 19.9

Wealth Index 19.7

Temperature 18.9

Vomiting 18.0

Supplementary Table 3



Table S3: The table contains all predictors considered by the random forest variable screening as well as the relative order of importance (reduction in mean square error (MSE) by the variable’s inclusion) for viral prediction for the whole dataset.

Supplemental Table 4Table S4: Steyerberg’s A and B (intercept and slope) for both logistic regression and random forest regression models fit with a different number of variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Logistic Regression** | | **Method** | **Random Forest** | |
| **# of Vars.** | **Intercept (A)** | **Slope (B)** | **# of Vars.** | **Intercept (A)** | **Slope (B)** |
| 1 | -0.0207 | 1.06 | 1 | 0.00665 | 0.974 |
| 2 | -0.00154 | 0.994 | 2 | 0.0211 | 0.922 |
| 3 | 0.00796 | 0.976 | 3 | 0.0286 | 0.911 |
| 4 | 0.00666 | 0.97 | 4 | 0.0508 | 0.83 |
| 5 | 0.00584 | 0.981 | 5 | 0.0375 | 0.88 |
| 6 | 0.00542 | 0.972 | 6 | 0.035 | 0.878 |
| 7 | 0.00224 | 0.985 | 7 | 0.0279 | 0.905 |
| 8 | 0.00452 | 0.978 | 8 | 0.0315 | 0.893 |
| 9 | 0.00582 | 0.977 | 9 | 0.0472 | 0.847 |
| 10 | 0.00692 | 0.976 | 10 | 0.0454 | 0.854 |
| 15 | 0.00783 | 0.968 | 15 | 0.0365 | 0.875 |
| 20 | 0.0173 | 0.949 | 20 | 0.0441 | 0.864 |
| 30 | 0.0239 | 0.924 | 30 | 0.0522 | 0.833 |
| 40 | 0.0276 | 0.914 | 40 | 0.054 | 0.83 |
| 50 | 0.0345 | 0.883 | 50 | 0.0562 | 0.814 |

Supplemental Table 5

Table S5: The odds ratios, 95% confidence interval, and p-value from a logistic regression model for the viral only outcome for the top 5 variables.

|  |  |  |
| --- | --- | --- |
|  |  | |
| Variable Name | Odds Ratios (95% CI) | P-value |
| Intercept | 3.046 (2.400 – 3.865) | <0.0001 |
| Age (mo.) | 0.945 (0.937 – 0.954) | <0.0001 |
| Season |  |  |
| Dry/Cold | Reference |  |
| Rainy/Cold | 0.220 (0.165 – 0.294) | <0.0001 |
| Dry/Hot | 0.290 (0.233 – 0.362) | <0.0001 |
| Rainy/Hot | 0.356 (0.284 – 0.446) | <0.0001 |
| Blood in stool | 0.141 (0.105 – 0.188) | <0.0001 |
| HAZ | 1.220 (1.142 – 1.305) | <0.0001 |
| Vomiting | 2.391 (2.007 – 2.849) | <0.0001 |
|  |  |  |

Supplementary Table 6

Any Bacteria

Variable Name

Beta Estimate (95%CI)

P-value

|  |  |  |
| --- | --- | --- |
| Intercept | -1.9330 (-5.3748 - 1.5089) | 0.2710 |
| Age (mo.) | 0.0567 (0.0455 - 0.0679) | <0.0001 |
| Blood in stool | 2.0062 (1.7533 - 2.2591) | <0.0001 |
| Breastfed |  |  |
| None | Reference |  |
| Partially | -0.7296 (-0.9592 - -0.4999) | <0.0001 |
| Exclusively | -0.6245 (-1.0411 - -0.2078) | 0.0033 |
| HAZ | -0.1036 (-0.1768 - -0.0305) | 0.0055 |
| MUAC | -0.0353 (-0.1029 - 0.0322) | 0.3050 |
| Resp. Rate (per min.) | 0.0131 (0.0028 - 0.0234)  Reference | 0.0124 |
| Season |
| Dry/Cold |
| Rainy/Cold | 0.9156 (0.6344 - 1.1969) | <0.0001 |
| Dry/Hot | 0.9725 (0.7595 - 1.1855) | <0.0001 |
| Rainy/Hot | 0.8389 (0.6164 - 1.0614) | <0.0001 |
| Wealth Index | -0.0211 (-0.1053 - 0.0631) | 0.6230 |
| Temperature (◦C) | 0.0299 (-0.0621 - 0.1218) | 0.5241 |
| Vomiting | -0.6766 (-0.846 - -0.5071) | <0.0001 |

Table S6: The estimate, 95% confidence interval, and p-value from a logistic regression model.

Supplemental Table 7

Table S7: The estimate, 95% confidence interval, and p-value from a logistic regression model with rotavirus cases removed.

|  |  |  |
| --- | --- | --- |
|  | Viral Only | |
| Variable Name | Odds Ratios (95% CI) | P-value |
| Intercept | 1.693 (0.006 – 488.504) | 0.855 |
| Age (mo.) | 0.951 (0.933 – 0.969) | <0.0001 |
| Season |  |  |
| Dry/Cold | Reference |  |
| Rainy/Cold | 0.390 (0.246 – 0.617) | <0.0001 |
| Dry/Hot | 0.550 (0.385 – 0.786) | 0.001 |
| Rainy/Hot | 0.600 (0.415 – 0.866) | 0.006 |
| Blood in stool | 0.154 (0.096 – 0.248) | <0.0001 |
| HAZ | 1.085 (0.964 – 1.220) | 0.176 |
| Vomiting | 1.578 (1.194 – 2..85) | 0.001 |
| Breastfed |  |  |
| None | Reference |  |
| Partially | 2.281 (1.520 – 3.424) | <0.0001 |
| Exclusively | 1.827 (0.921 – 3.622) | 0.085 |
| MUAC | 1.061 (0.965 – 1.167) | 0.220 |
| Resp. Rate (per min.) | 0.982 (0.965 – 1.000) | 0.047 |
| Wealth Index | 1.099 (0.958 – 1.260) | 0.178 |
| Temperature (◦C) | 0.954 (0.819 – 1.112) | 0.550 |

Supplementary Table 8

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Africa | |  |  | Asia |  |
| Country  Variable | The Gambia | Mali | Mozambique | Kenya | India | Bangladesh | Pakistan |
| 1 | Age | Age | Age | Age | Age | Age | Age |
| 2 | Blood in stool | MUAC | Breastfed | Breastfed | MUAC | Blood in stool | MUAC |
| 3 | HAZ | Breastfed | HAZ | HAZ | HAZ | Eyes | HAZ |
| 4 | Breastfed | HAZ | MUAC | MUAC | Season | Vomiting | Breastfed |
| 5 | MUAC | Resp. Rate | Season | Resp. Rate | Resp. Rate | Season | Resp. Rate |
| 6 | Temp. | Wealth Index | Resp. Rate | Temp. | Blood in stool | MUAC | Wealth Index |
| 7 | Resp. Rate | Temp. | Temp. | Wealth Index | Wealth Index | Rectal Straining | Temp. |
| 8 | Season | Ppl. of House | Wealth Index | Ppl. of House | # Share Fac. | Temp. | Ppl. of House |
| 9 | Wealth Index | Vomiting | Vomiting | # Share Fac. | Temp. | HAZ | Blood in stool |
| 10 | Ppl. of House | Season | Blood in stool | Water Source | Ppl. of House | Wealth Index | Days of Episode |
| Country AUCs | **0.827** (0.818-0.836)  0.840 (0.832-0.849) | **0.741** (0.731-0.752)  0.756 (0.746-0.767) | **0.802** (0.791-0.813)  0.831 (0.821-0.841) | **0.763** (0.752-0.773)  0.773 (0.763-0.784) | **0.775** (0.767-0.784)  0.807 (0.799-0.815) | **0.892** (0.885-0.898)  0.921 (0.915-0.927) | **0.760** (0.750-0.770)  0.765 (0.755-0.775) |
| Continent AUCs |  | **0.779 (0.774-0.785)**  0.797 (0.792-0.802) | |  |  | **0.826 (0.821-0.830)**  0.850 (0.845-0.854) |  |
| Overall AUC |  |  | | **0.806 (0.803-0.809)**  0.829 (0.825-0.832) |  |  |  |
| Continent External Validation | **0.746** (0.696-0.797)  0.767 (0.719-0.815) | **0.711** (0.657-0.765)  0.790 (0.736-0.842) | **0.773** (0.722-0.823)  0.831 (0.787-0.876) | **0.723** (0.672-0.773)  0.715 (0.664-0.766) | **0.772** (0.735-0.810)  0.811 (0.777-0.846) | **0.897** (0.868-0.926)  0.930 (0.906-0.953) | **0.765** (0.719-0.811)  0.782 (0.738-0.826) |

Table S8: The table contains both site-specific variable importance ordering and a cross-validated average overall AUC, AUC by country, and AUC by continent and confidence intervals from a 5 (bold) and 10 (ital.) variable logistic regression model for predicting a viral etiology with variables based on the overall variable importance. Lastly, it shows the AUC and a 95% confidence interval resulting from testing the logistic regression with variables based on the overall variable importance on each site individually following its training on the other countries in the same continent.

9