Section 12: Appendix B of SAP for RSV correlates study

Imputation of missing biomarker data

If the received biomarker data has more than 5% missingness, multiple imputation strategies will be used to impute missing data. Such strategies will depend on the correlation between biomarker data from different assays and different time points.

Risk Score analysis

Covariates with missing values

The risk score analysis will be carried out upon scaling of all covariates (binary, count, and continuous variables) to have mean of 0 and standard deviation of 1. Covariates with a high number of missing values ($> 5\%$) will be dropped from the analysis, while missing data for covariates having $< 5\%$ missing values will be imputed. One exception to this was the child5 variable (Indicator other children < 5 years of age in home) that had missing data for 70/784 (8.9%) of records which was included in the analysis upon finding a significant association (univar_logistic_pval < 0.01) with endpoint 1 in the placebo group.

Exclusion of vaccination to birth variable from Risk Score analysis

The vaccination to birth variable (continuous or indicator) will not be considered in the risk score analysis. This is justified as this variable will be used as a covariate in the logistic regression modeling and allow comparison between different marker levels for mother-infant pairs with similar times from vaccination to birth.

Exclusion of SL.glmnet learner when k=1

The risk score models with k=1 for maternal and pediatric variables will not include SL.glmnet learner because glmnet package doesn't support a data matrix with a single variable.

Replace Random Forest learner with cForest

Since the random forest learner does not accept observation weight, it will be replaced by cForest, a tree-based algorithm that accepts weights.

Changes to Maternal enrollment and birth/delivery variable sets

Following preliminary analysis, 4 variables (child5, season, smoker and daycare) were moved from the birth/delivery variable set to the maternal enrollment variable set.

Risk scores

Risk scores will be derived as the fitted probability of the outcome based on the selected superlearner for each variable set and endpoint. For vaccine group, fitted probability of outcome (risk score) will be predicted using the selected model (trained on data from the placebo group) upon averaging over 10 random seeds. For placebo group, fitted probability of outcome (risk score) will be derived from the 5 outer folds conducted to estimate superlearner performance (CV-AUC) upon averaging over 10 random seeds.

Usage of risk scores in correlates analyses

The risk scores will be converted to the logit-linear risk score before being used in correlates analyses.