Supplemental Figures

The following plots compare performance of the different modeling strategies for the 5 longitudinal measures. In general summary measures perform best, however the best summary measure differs based on predictor. Predictive variability stabilizes by a sample size of 5,000.

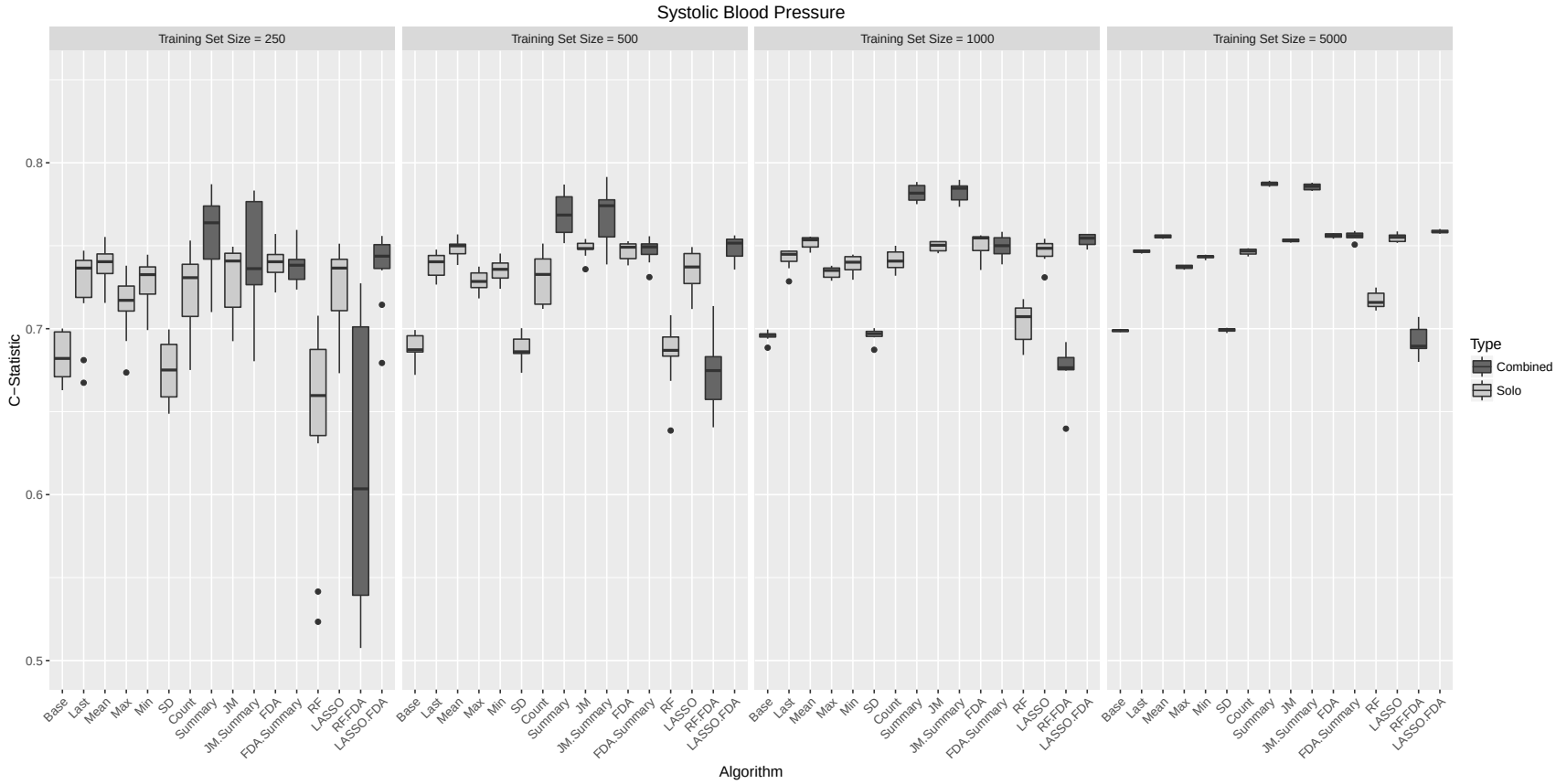

Figure 1: Box plots of model performance (C-statistics) for multiple measurements of *Systolic Blood Pressure*. Each panel refers to different training set sizes ranging from 250 - 5,000 people, with each analysis run 10 times. Each model was evaluated was evaluated on the same test set of 5,000 people. Models are grouped based on whether multiple methods were grouped together.

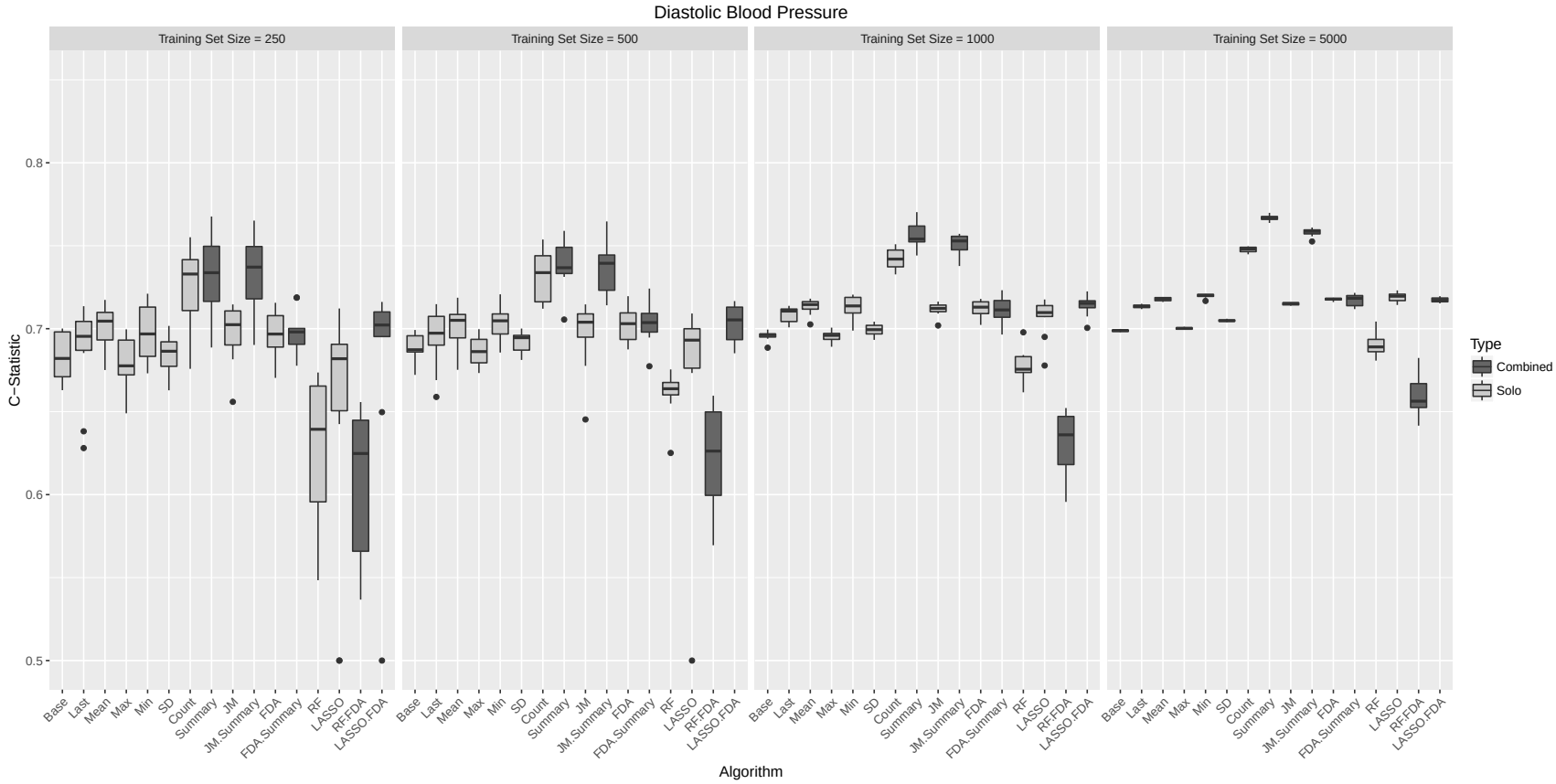

Figure 2: Box plots of model performance (C-statistics) for multiple measurements of *Diastolic Blood Pressure*. Each panel refers to different training set sizes ranging from 250 - 5,000 people, with each analysis run 10 times. Each model was evaluated was evaluated on the same test set of 5,000 people. Models are grouped based on whether multiple methods were grouped together.

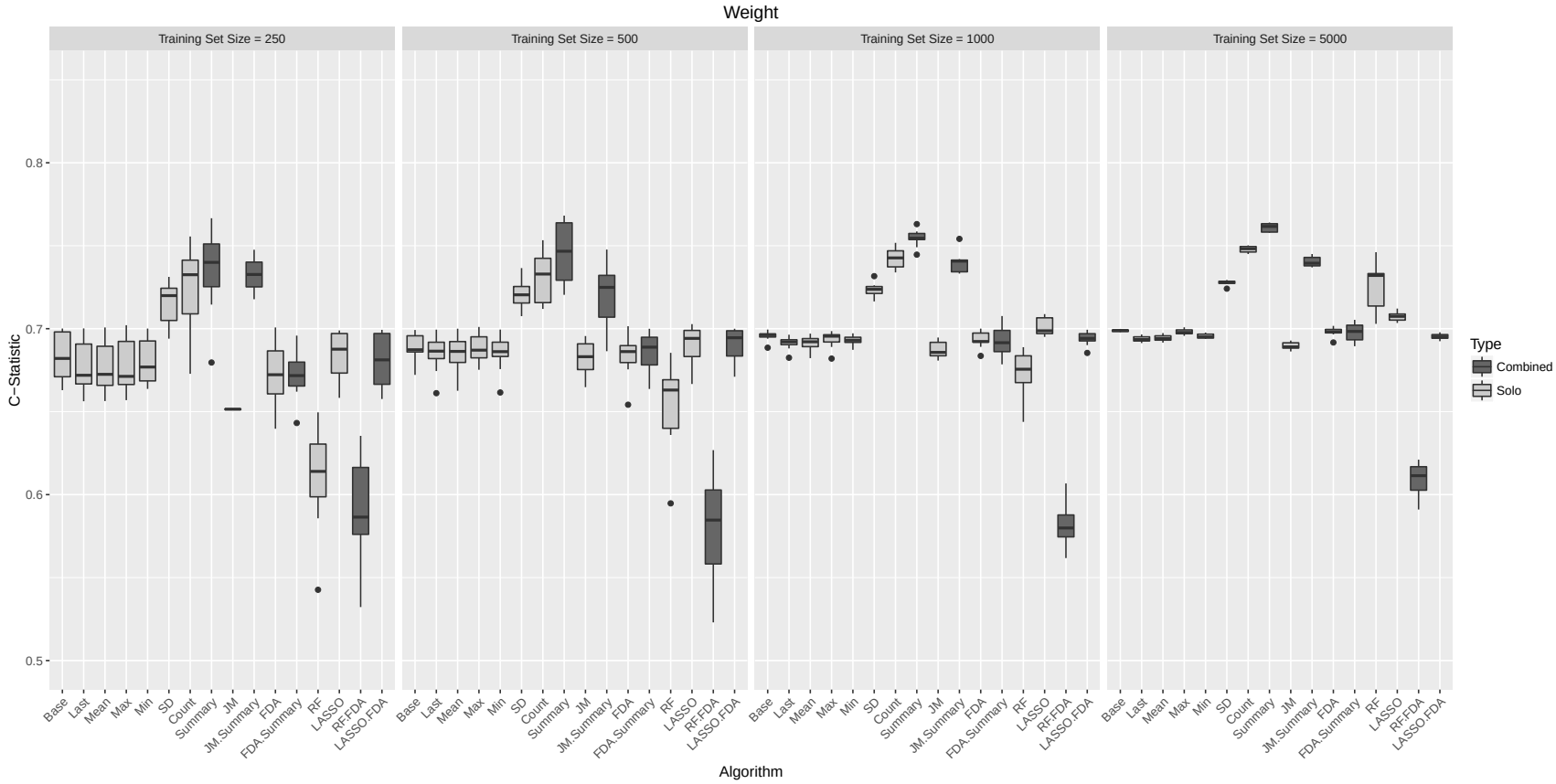

Figure 3: Box plots of model performance (C-statistics) for multiple measurements of *Weight*. Each panel refers to different training set sizes ranging from 250 - 5,000 people, with each analysis run 10 times. Each model was evaluated was evaluated on the same test set of 5,000 people. Models are grouped based on whether multiple methods were grouped together.

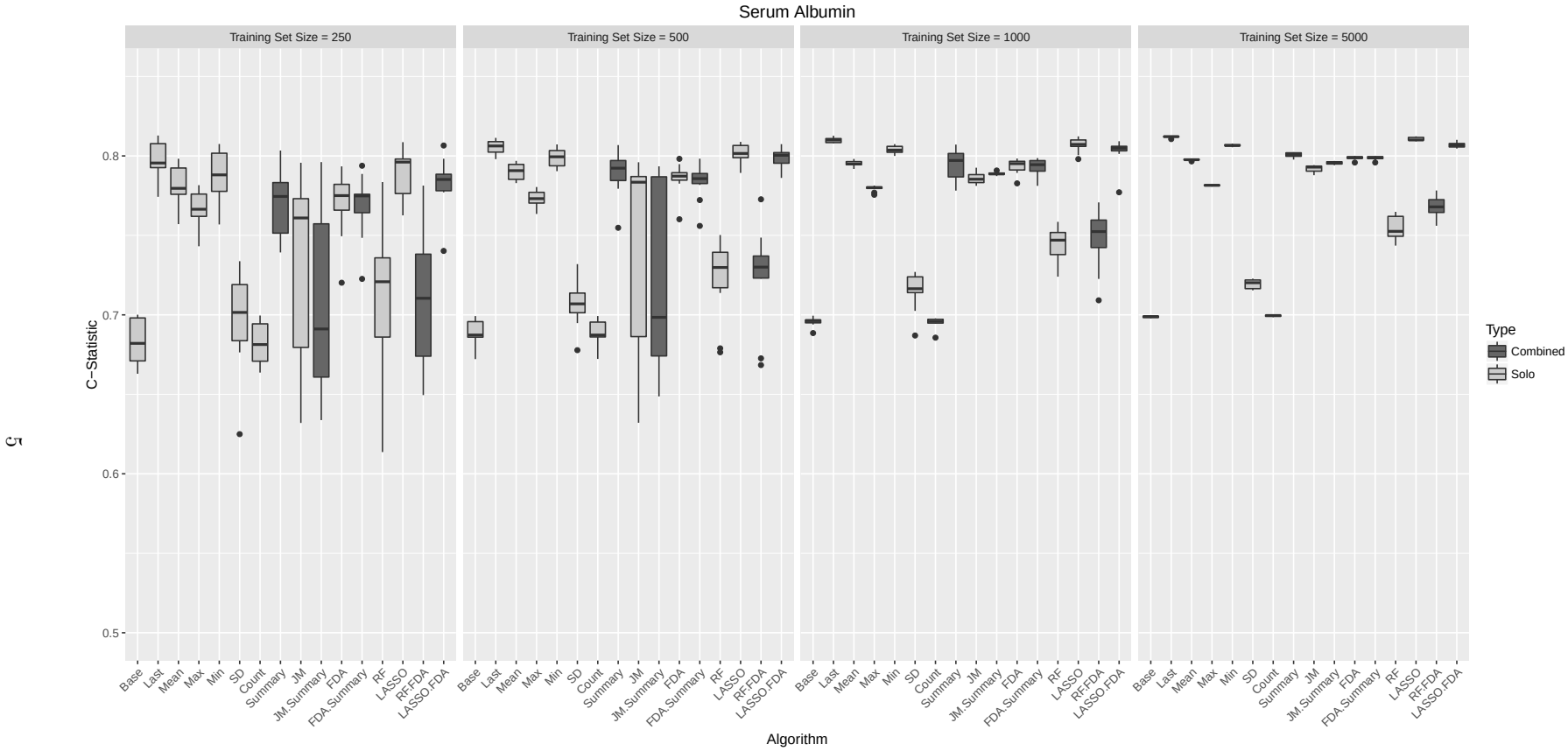

Figure 4: Box plots of model performance (C-statistics) for multiple measurements of *Serum Albumin*. Each panel refers to different training set sizes ranging from 250 - 5,000 people, with each analysis run 10 times. Each model was evaluated was evaluated on the same test set of 5,000 people. Models are grouped based on whether multiple methods were grouped together.

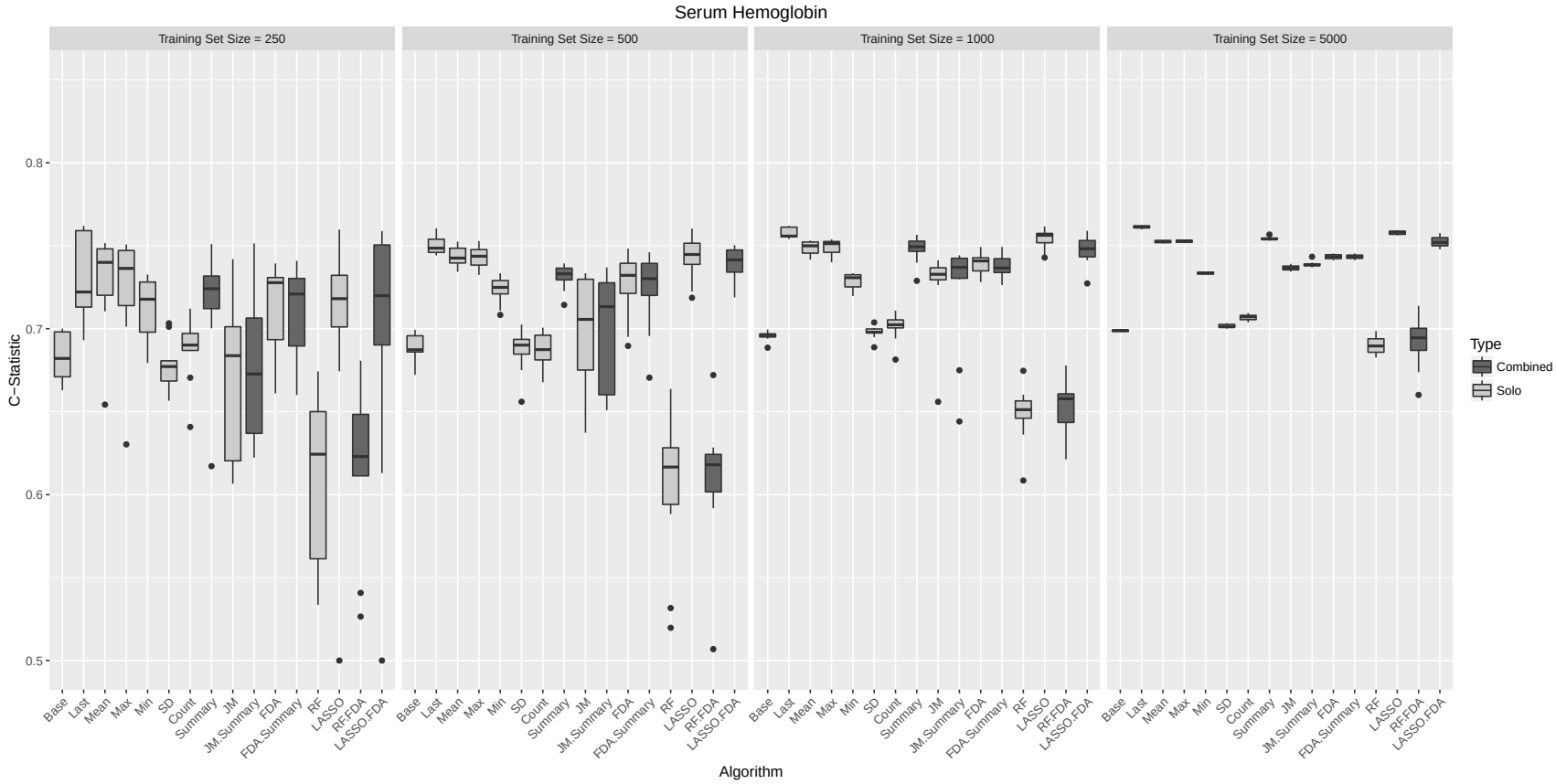

Figure 5: Box plots of model performance (C-statistics) for multiple measurements of *Serum Hemoglobin*. Each panel refers to different training set sizes ranging from 250 - 5,000 people, with each analysis run 10 times. Each model was evaluated was evaluated on the same test set of 5,000 people. Models are grouped based on whether multiple methods were grouped together.