Web Appendix 1: Supplementary Methods

Estimating uncertainty in carriage prevalence

We sought to use serotype-specific carriage prevalence in children as one of the covariates in a model predicting IPD incidence in children and adults. The observed number of carriers (*n)* was assumed to be a sample from a binomial distribution, where the number of trials was equal to the number of swabs (*N*) and the true prevalence is unobserved and estimated from the data. The prior distribution for the true prevalence was uniform between 0.0001 and 0.9999. This was estimated using the R package RJags (1). The posterior mean and standard deviation of the log(true prevalence) were then used in subsequent models of IPD incidence to account for the uncertainty in the true prevalence values.

Estimating uncertainty in invasiveness

Invasiveness is defined here as the incidence of disease (number of cases per 1,000 population per year) divided by the prevalence of carriage (percent of individuals carrying the serotype). In this instance, there is uncertainty in both the numerator and denominator. To estimate the invasiveness for each serotype in each time period and age group, and the uncertainty around these estimates, we again used a Bayesian framework and jointly modeled carriage and IPD incidence. We performed a 2-stage analysis, first estimating invasiveness for each serotype, age group, and time period, and then pooling the estimates. In the first stage, the uncertainty in carriage prevalence was modeled as in the previous section. The number of cases of disease was assumed to be a sample from a Poisson distribution with mean mu. The prior on mu was weakly informative (mean 0, precision: $1x10^{-3}$). Invasiveness was then calculated by dividing the

estimated true incidence by the estimated true carriage prevalence. This first step gives an estimate of invasiveness for each age group, serotype, and time period.

In the second step, we pooled the data between time periods and borrowed information between serotypes to obtain robust estimate of invasiveness for each serotype. This approach effectively shrinks the estimates of invasiveness for each serotype towards the overall mean invasiveness for all serotypes. For serotypes with robust information, the true invasiveness estimate will be close to the observed value from the two stage model. For serotypes with sparse information, the estimate for true invasiveness will be close to the mean across all serotypes. The estimates of invasiveness calculated in the first stage were assumed to be drawn from a normal distribution with a mean equal to the true, unobserved invasiveness for the serotype and a precision equal to the precision estimated in the first stage. This results in a single estimate of invasiveness for each serotype across the time periods. The serotype-specific invasiveness estimates were drawn from a normal distribution with a mean of theta and precision of tau. Theta and tau were given weakly informative priors (Theta: normal prior, mean 0, precision of $1x10^{-6}$, Tau: gamma (0.001,0.001) prior). This second stage model was fit separately to each age group to obtain separate estimate of invasiveness for each age group and serotype. Results are based on 100,000 Markov Chain Monte Carlo iterations of 2 chains (i.e. 2 independent estimates that are combined). The convergence of the models was assessed by visual evaluation of the trace plots.

Evaluation of alternative models of IPD incidence

We also compared out main model (Model 1) with a simpler model that did not include a random intercept for each serotype (Model 2) and a simpler model (Model 3) that did not include the random effect for individual measurements. Model 1 had a Deviance Information Criteria

(DIC) score that was worse (3 points higher) in children <7 and adults 40+ but was slightly better (1 point lower) in adults 18-39 when compared with the simpler models. Model 3 (no observation-level random effect) had a significantly worse DIC score (i.e. increase in DIC of at least 13 points in all age groups). Because the DIC scores were similar for Models 1 and 2, we performed our cross validation approach using both models and then compared the mean-squared errors (MSE; comparison of observed values in the validation dataset and the predicted values). Both models performed similarly among the <7 and 18-39 year olds. However, the MSE of the predictions for the 40+ year olds was substantially reduced when including the serotype-specific random intercept. The MSE improved in three of the four respective cross-validation sets among 40+ year olds when including the serotype-specific random intercept from 57.6►36.5, (23.7►39.8), 123.5►68.7, 8.2►5.9. For the other age groups, MSE estimates were within 1.5 points for all time periods.

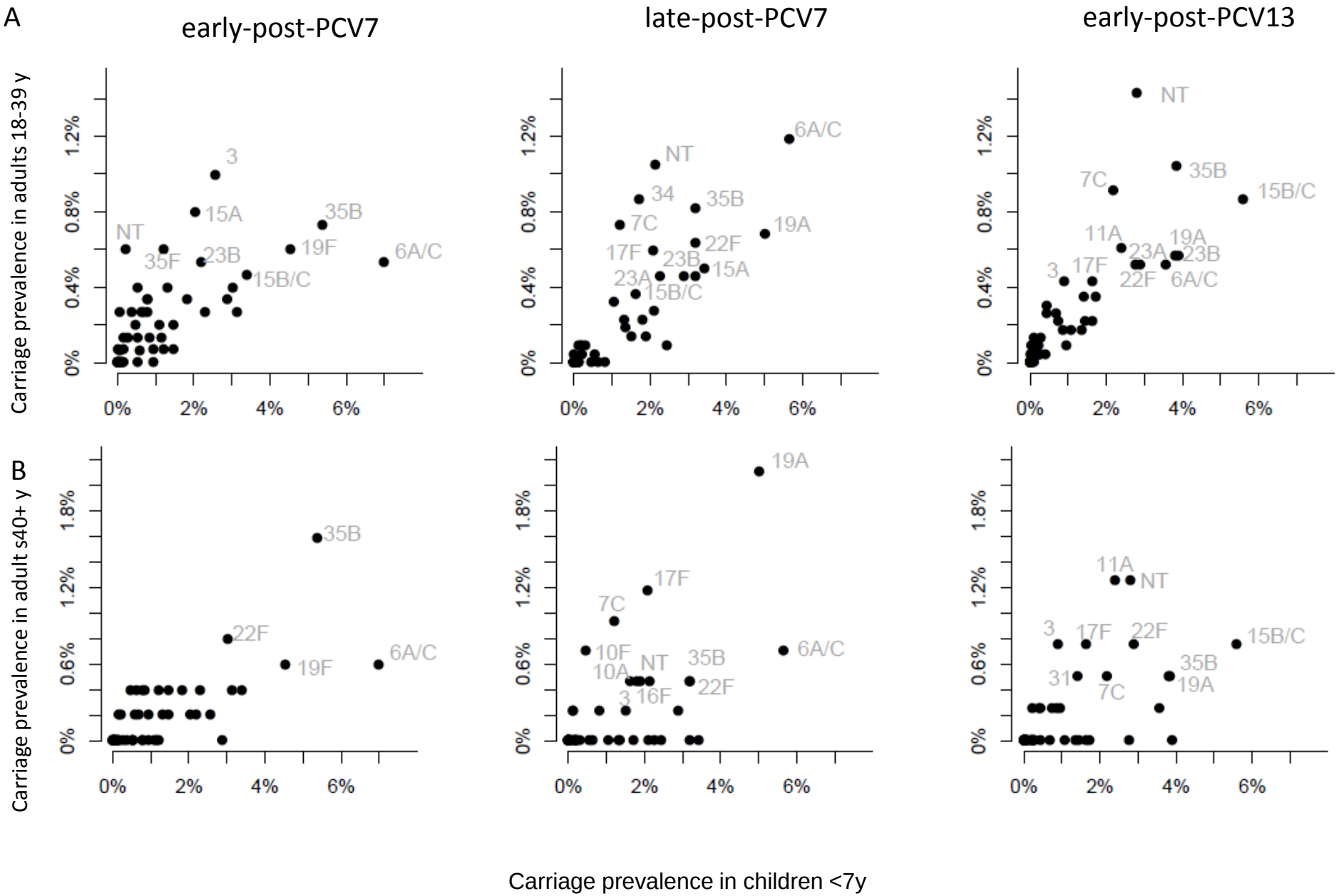
Web Appendix: 2: R Analysis code and data (see .zip files):

1) R code and data for running the cross-validation analysis in JAGS. Random integers were added to the carriage and invasive pneumococcal disease data for privacy reasons
2) R code, sample data (with random integers added), and instructions for generating predictions based on pediatric carriage data

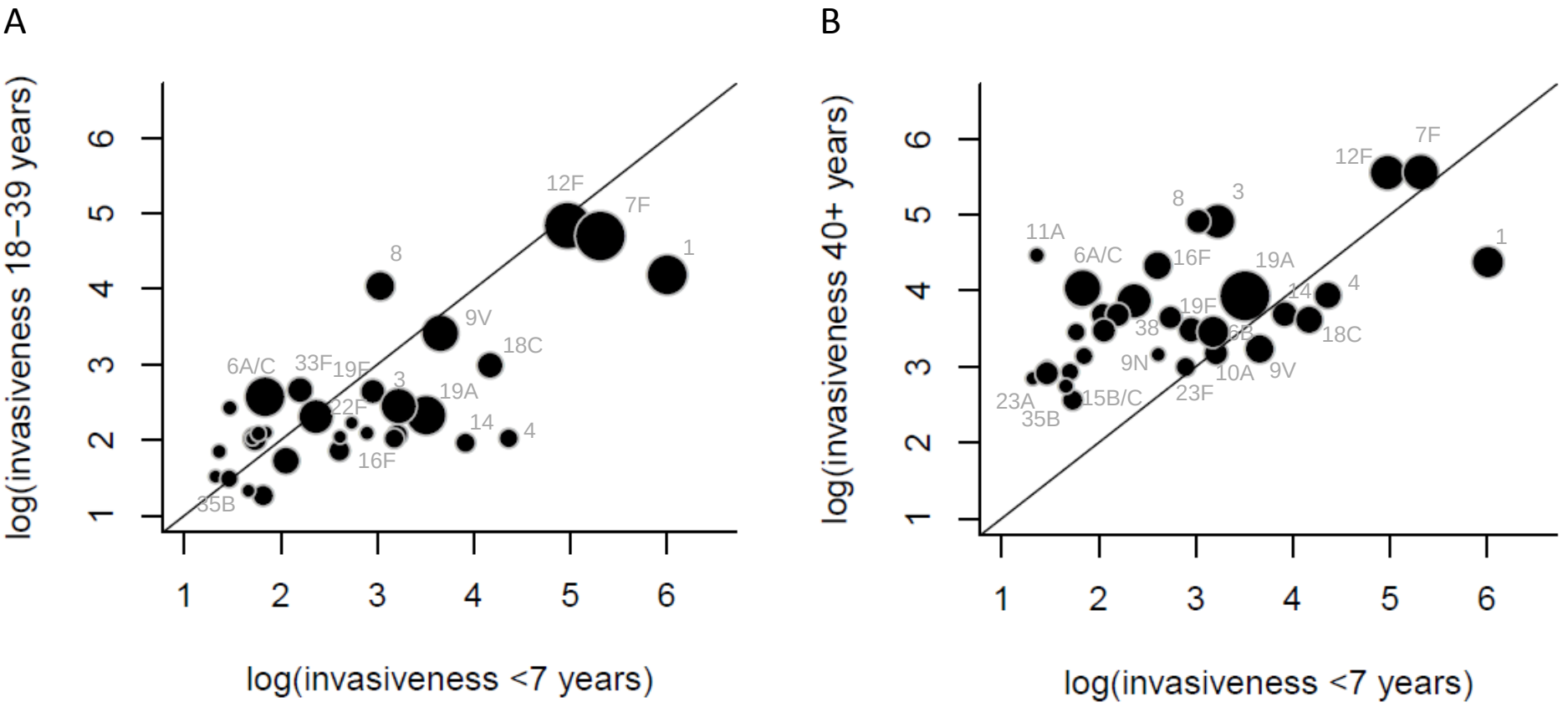
Web Appendix References

1. Plummer M. rjags: Bayesian graphical models using MCMC. R package version 3-14., 2014.

Web Figures

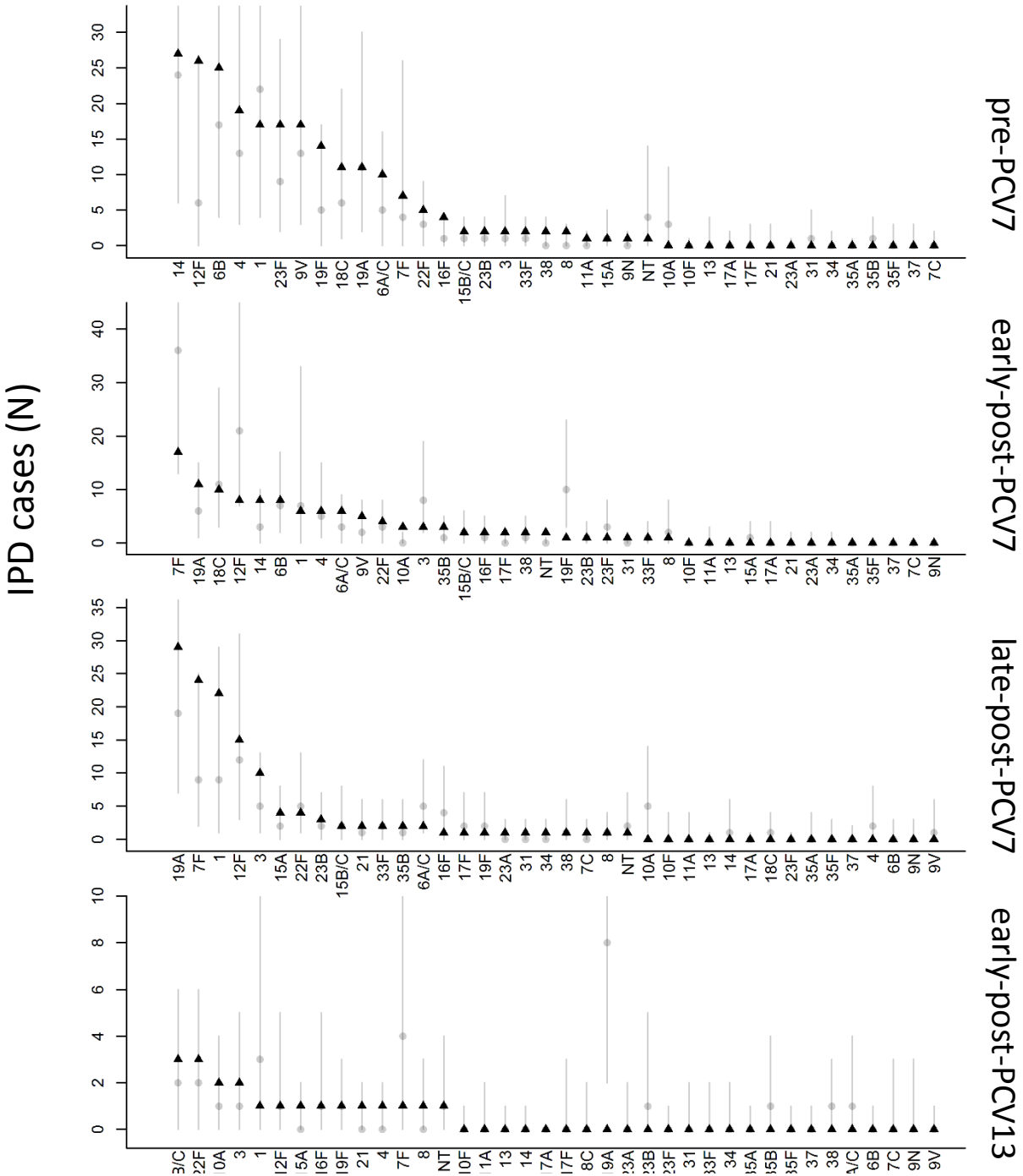


Web Figure 1. Comparison of carriage frequency among children <7 years and (A) adults 18-39 years of age or (B) adults 40+ years of age. Each panel represents a different time period: early-post-pneumococcal conjugate vaccine period (PCV7), late-post-PCV7 period, and early-post-PCV13 period. Carriage data from adults were not available from the pre-PCV7 period.



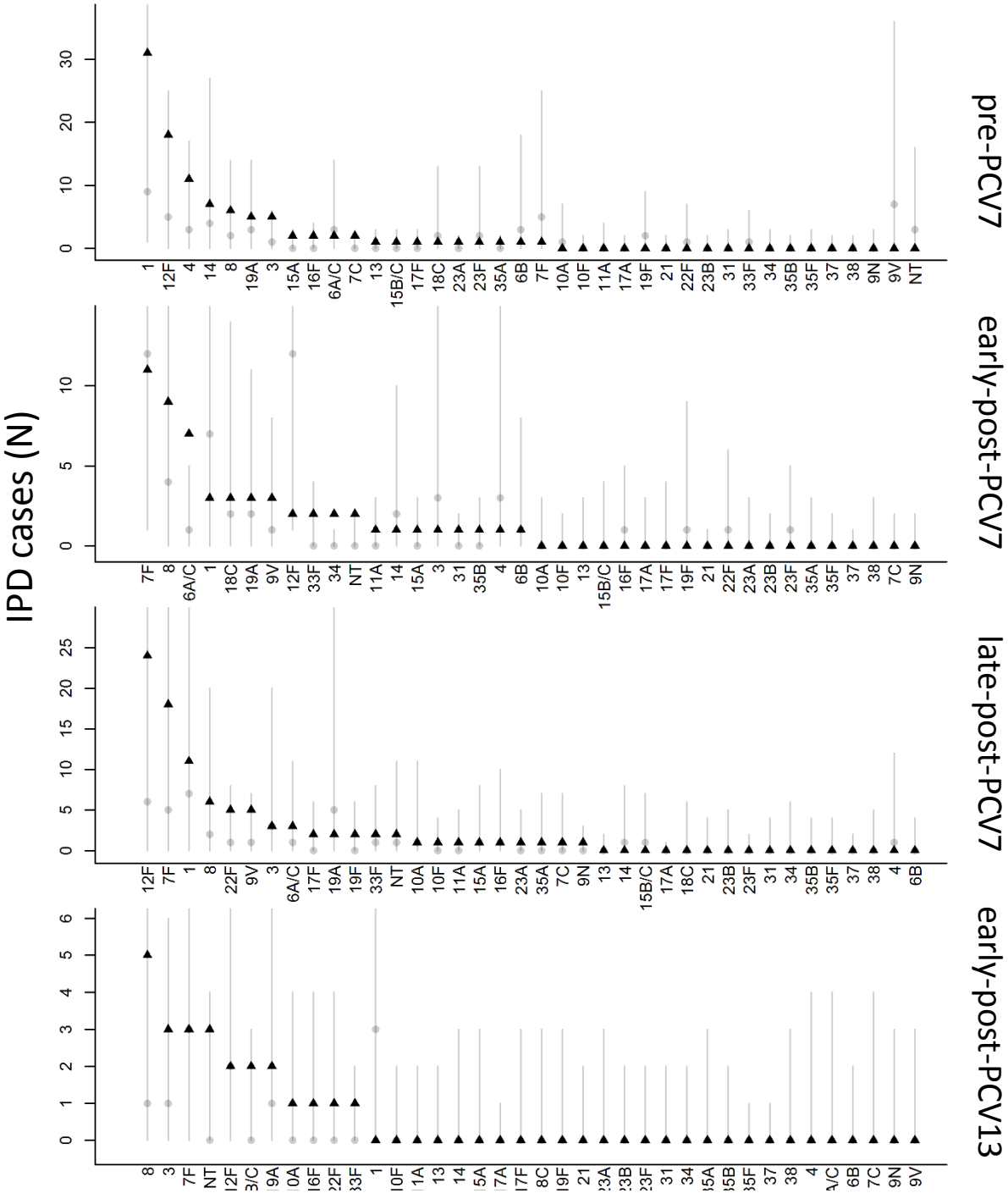
Web Figure 2. Comparison of invasiveness patterns (incidence of disease divided by frequency of carriage) between children <7 years and (A) adults 18-39 years or (B) adults 40+ years. The size of the bubble is proportional to the precision of the estimates (more confidence in larger bubbles). The diagonal line denotes X=Y.

Web Figure 3
<7 years old



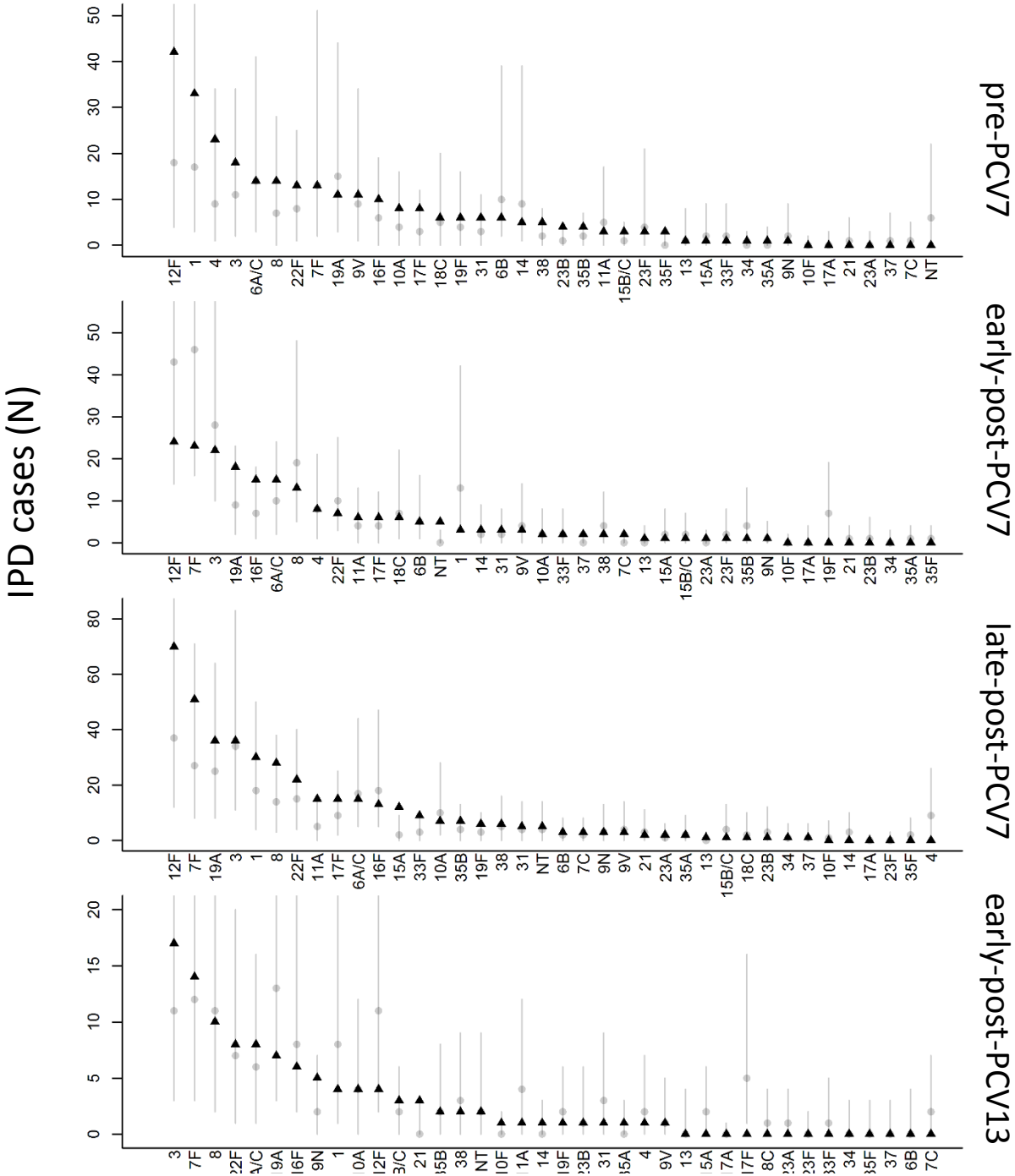
The observed number of cases of invasive pneumococcal disease caused by each serotype in in each time period (black triangle) compared with the number of cases predicted from a regression model (gray circle with gray vertical lines representing the 95% credible interval). The regression included carriage frequency in children, invasiveness in children, and a random effect for serotype. The estimates are based on a cross-validation approach, where the time period shown was held out during model fitting.

Web Figure 4
18-39 year olds

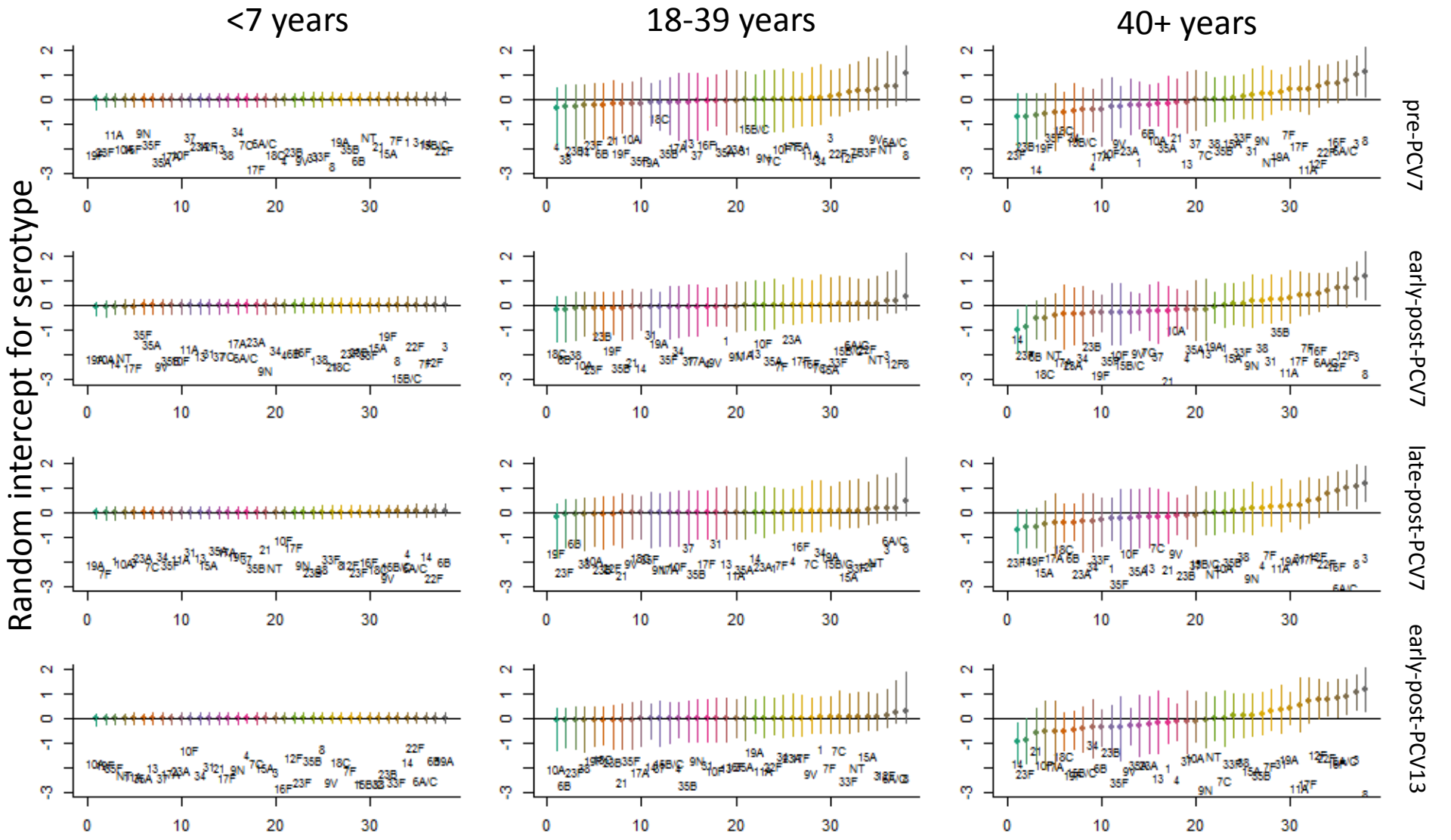


The observed number of cases of invasive pneumococcal disease caused by each serotype in in each time period (black triangle) compared with the number of cases predicted from a regression model (gray circle with gray vertical lines representing the 95% credible interval). The regression included carriage frequency in children, invasiveness in children, and a random effect for serotype. The estimates are based on a cross-validation approach, where the time period shown was held out during model fitting

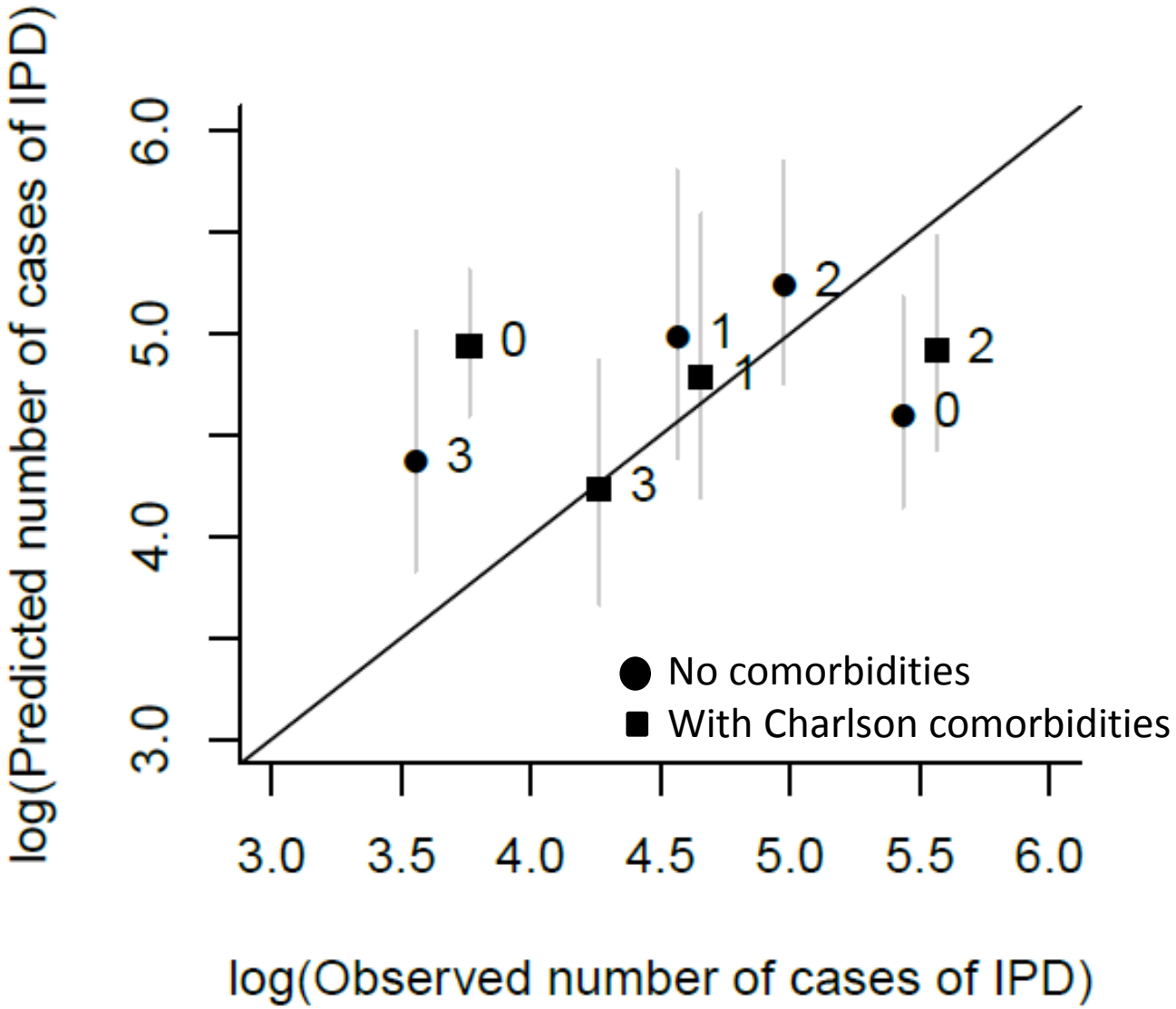
Web Figure 5
40+ year olds



The observed number of cases of invasive pneumococcal disease caused by each serotype in in each time period (black triangle) compared with the number of cases predicted from a regression model (gray circle with gray vertical lines representing the 95% credible interval). The regression included carriage frequency in children, invasiveness in children, and a random effect for serotype. The estimates are based on a cross-validation approach, where the time period shown was held out during model fitting

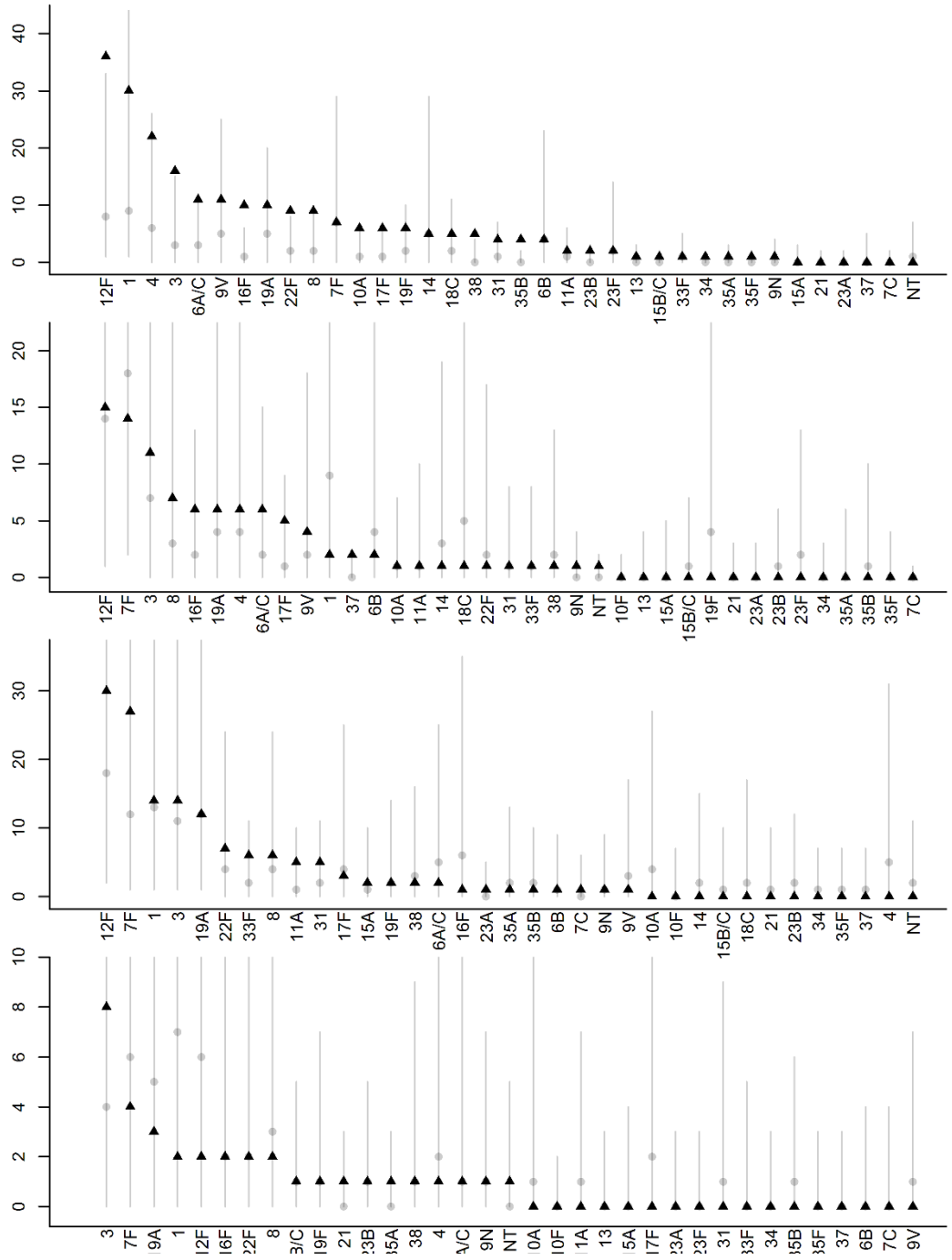


Web Figure 6. Serotype-specific random intercepts from cross-validation models. Higher values indicate that the serotype tends to cause more disease than would be expected based only on invasiveness and carriage prevalence in children.



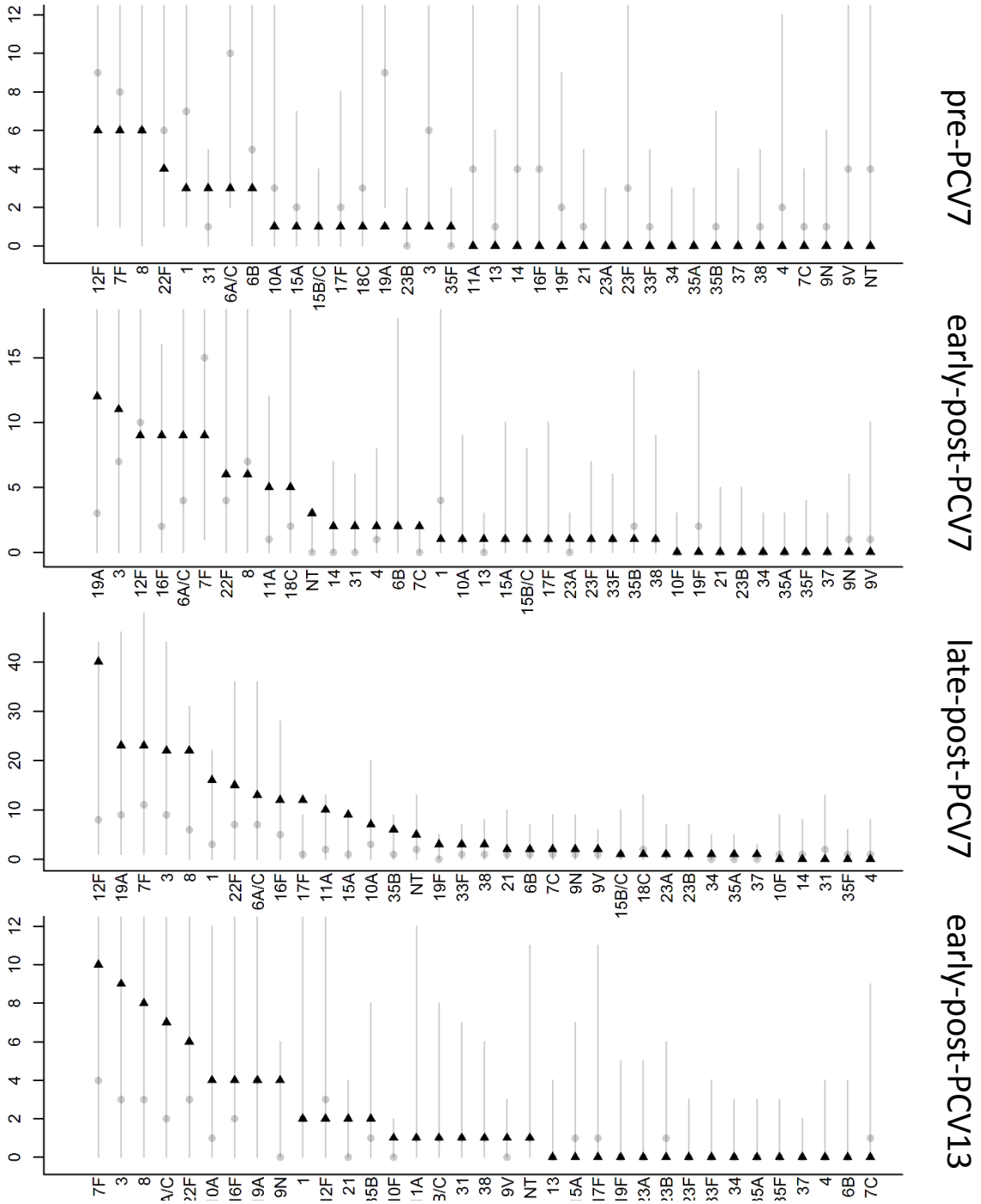
Web Figure 7. The observed number of cases of invasive pneumococcal disease in each comorbidity group and time period among adults 40+ years of age compared with the number of cases predicted from a regression model. The regression included serotype-specific carriage frequency in children, invasiveness in children, and a random effect for serotype. The model was fit to serotype level data and then summed. The estimates are based on a cross-validation approach, where the time period shown was held out during model fitting, and predictions are based on observed carriage prevalence during the time period. The numbers indicate the time period (0=pre-pneumococcal conjugate vaccine (PCV), 1=early-post-PCV7, 2=late-post-PCV7, 3=early-post-PCV13), and the colors distinguish the comorbidity groups.

Web Figure 8. Adults 40+ without comorbidities

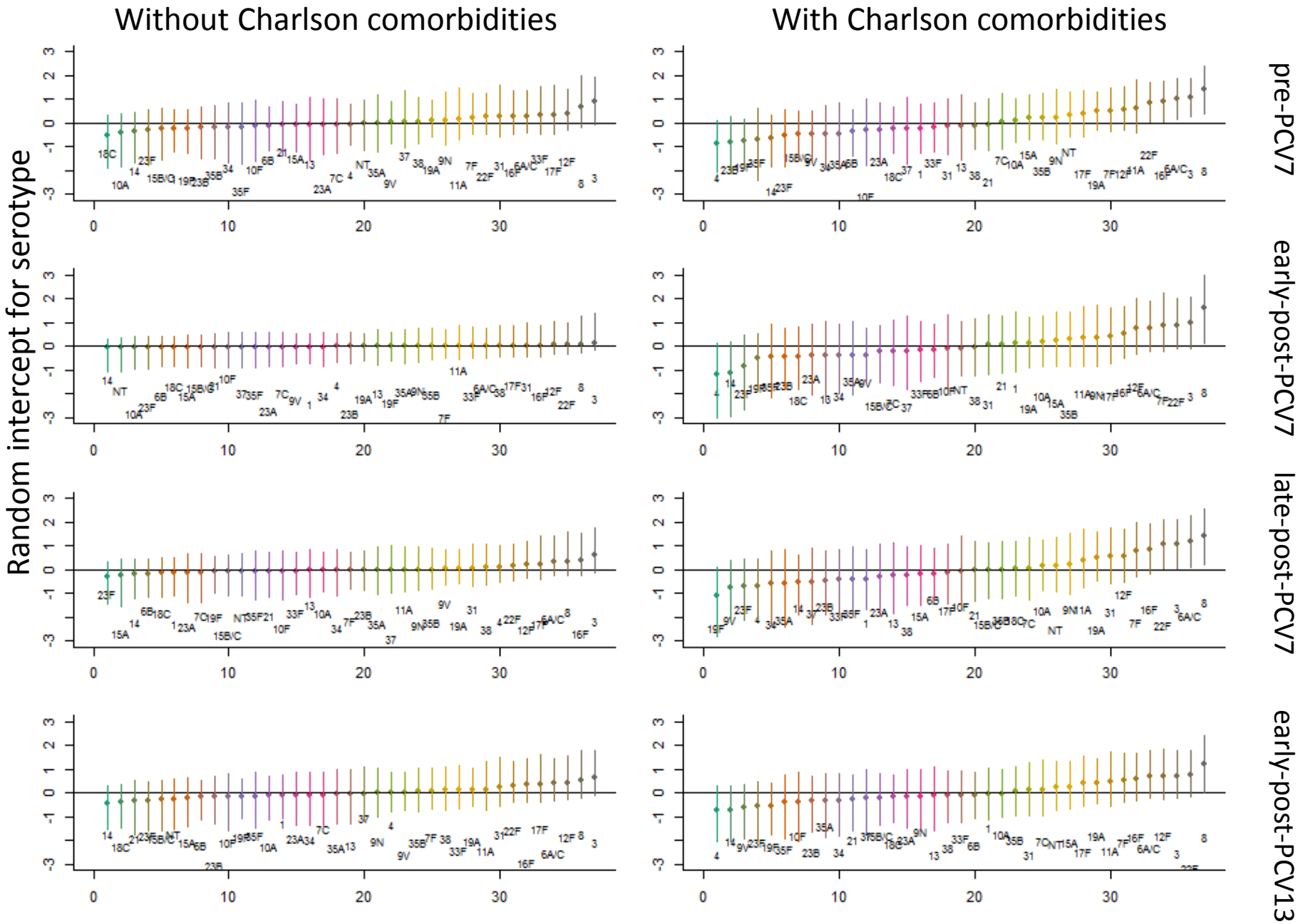


The observed rate of invasive pneumococcal disease among 40+ year olds without known comorbidities (Charlson index) compared with the number of cases predicted from regression models where the indicated time period was held out during model fitting. The regression included serotype-specific carriage frequency in children, invasiveness in children, and a random effect for serotype. The estimates are based on a cross-validation approach, where the time period shown was held out during model fitting, and predictions are based on observed carriage prevalence during the time period, invasiveness, and the serotype-specific intercept.

Web Figure 9. Adults 40+ with comorbidities



The observed rate of invasive pneumococcal disease among 40+ year olds with known comorbidities (Charlson index) compared with the number of cases predicted from regression models where the indicated time period was held out during model fitting. The regression included serotype-specific carriage frequency in children, invasiveness in children, and a random effect for serotype. The estimates are based on a cross-validation approach, where the time period shown was held out during model fitting, and predictions are based on observed carriage prevalence during the time period, invasiveness, and the serotype-specific intercept.



Web Figure 10. Random intercepts for each serotype from models fit to data from adults 40+ years of age for those with or without Charlson comorbidities. Higher values indicate that the expected incidence of the serotypes tends to be higher than what would be expected based only on invasiveness and carriage prevalence in children.