Analysing pneumococcal invasiveness using Bayesian models of pathogen progression rates

Alessandra Løchen[1,2], James E. Truscott[1] and Nicholas J. Croucher[1,2,*]

[1] Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, W2 1PG, UK

[2] MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, W2 1PG, UK

* n.croucher@imperial.ac.uk

Short title: Bayesian analyses of microbial progression rates

Abstract

The disease burden attributable to opportunistic pathogens depends on their prevalence in asymptomatic colonisation and the rate at which they progress to cause symptomatic disease. Increases in infections caused by commensals can result from the emergence of “hyperinvasive” strains. Such pathogens can be identified through quantifying progression rates using matched samples of typed microbes from disease cases and healthy carriers. This study describes Bayesian models for analysing such datasets, implemented in an RStan package (https://github.com/nickjcroucher/progressionEstimation). The models converged on stable fits that accurately reproduced observations from meta-analyses of *Streptococcus pneumoniae* datasets. The estimates of invasiveness, the progression rate from carriage to invasive disease, in cases per carrier per year correlated strongly with the dimensionless values from meta-analysis of odds ratios when sample sizes were large. At smaller sample sizes, the Bayesian models produced more informative estimates. This identified historically rare but high-risk *S. pneumoniae* serotypes that could be problematic following vaccine-associated disruption of the bacterial population. The package allows for hypothesis testing through model comparisons with Bayes factors. Application to datasets in which strain and serotype information were available for *S. pneumoniae* found significant evidence for within-strain and within-serotype variation in invasiveness. The heterogeneous geographical distribution of these genotypes is therefore likely to contribute to differences in the impact of vaccination in between locations. Hence genomic surveillance of opportunistic pathogens is crucial for quantifying the effectiveness of public health interventions, and enabling ongoing meta-analyses that can identify new, highly invasive variants.

Author summary

Opportunistic pathogens are microbes that are commonly carried by healthy hosts, but can occasionally cause severe disease. The progression rate quantifies the risk of such a pathogen transitioning from a harmless commensal to causing a symptomatic infection. The incidence of infections caused by opportunistic pathogens can rise with the emergence of “hyperinvasive” strains, which have high progression rates. Therefore methods for calculating progression rates of different pathogen strains using surveillance data are crucial for rapidly identifying emerging infectious disease threats. Existing methods typically measure progression rates relative to the overall mix of microbes in the population, but these populations can vary substantially between locations and times, making the outputs challenging to combine across studies. This work presents a new method for estimating progression rates from surveillance data that generates values useful for modelling pathogen populations, even from relatively small sample sizes.

54 **Introduction**

55 Opportunistic pathogens are commonly found in the environment, or microbiota of
56 healthy hosts, but have the capacity to cause disease in some host organisms [1].
57 The frequency with which they transition from asymptomatic colonisation to
58 symptomatic infection can be quantified as a progression rate [2]. Many opportunistic
59 pathogens are diverse, and can be typed by a variety of phenotypic methods [3], or
60 divided into strains using genotyping or whole genome sequencing [4,5].
61 Understanding the variation in progression rates across such species is crucial for
62 understanding changes in the incidence of diseases caused by opportunistic
63 pathogens. Recently, strains with high progression rates, sometimes referred to as
64 "hypervirulent" [6] or "hyperinvasive" [7], have been ascribed as the primary cause of
65 rising case numbers of disease caused by *Neisseria meningitidis* [8], *Streptococcus*
66 *pyogenes* [9] and *Klebsiella pneumoniae* [10], among others. However,
67 understanding whether such strains drive elevated disease burdens through more
68 rapid transmission, or heightened progression rates, typically requires additional
69 information on the carried population of the opportunistic pathogen.
70

71 Progression rates have been extensively studied in the nasopharyngeal commensal
72 and respiratory pathogen *Streptococcus pneumoniae* (the pneumococcus). Over one
73 hundred serotypes have been identified in this species, each corresponding to a
74 structurally-distinct polysaccharide capsule [11,12]. These are clustered into 48
75 serogroups, based on serological cross-reactivity [13]. This capsule inhibits
76 pneumococci being cleared by the immune system, and is crucial to their ability to
77 cause invasive pneumococcal disease (IPD) [14,15], infections of normally sterile
78 anatomical sites. Hence the rate of progression from carriage to IPD is referred to as

79 invasiveness [16]. It has long been assumed the capsule is an important determinant
80 of pneumococcal invasiveness, based on differences in serotypes' ability to cause
81 disease in animal experiments [17]. Epidemiological differences are also apparent in
82 human disease, with paediatric serotypes (including serotypes within serogroups 6,
83 9, 19, 23, and serotype 14) common in sporadic infant disease, and 'epidemic'
84 serotypes (such as 1, 2, 5 and 12F) capable of causing IPD outbreaks in adults
85 [18,19].
86
87 Understanding these differences between serotypes has become crucial for *S.*
88 *pneumoniae* epidemiology, owing to the introduction of polysaccharide conjugate
89 vaccines (PCVs) [20]. A seven valent PCV (PCV7) was introduced in the USA in
90 2000, which has been supplanted by ten- and thirteen-valent formulations (PCV10
91 and PCV13, respectively). These induce immune responses against a specific
92 subset of serotypes, resulting in their elimination from carriage [21]. However, the
93 overall frequency of *S. pneumoniae* colonisation typically remains stable after any
94 PCV's introduction, owing to the replacement of vaccine serotypes by the plethora of
95 non-vaccine serotypes [22,23]. Nevertheless, PCVs have usually proved highly
96 effective at reducing IPD through facilitating the replacement of the pre-PCV carried
97 population with pneumococcal serotypes associated with lower invasiveness [24,25].
98 Hence optimal PCVs are those which minimise the overall invasiveness of the
99 carried pneumococcal population [26,27]. Therefore estimation of invasiveness
100 across vaccine-targeted serotypes, and those which may replace them post-PCV, is
101 vital for reducing the incidence of IPD. Such quantification of invasiveness is typically
102 achieved through paired studies of serotypes' prevalence in geographically and
103 temporally matched surveys of pneumococcal carriage, and surveillance of IPD.

104

105 Multiple methods have been used to estimate invasiveness using such paired case
106 and carrier data. The most common approaches use odds ratios: the ratio of isolates
107 of a given serotype recovered from disease against those recovered from carriage,
108 divided by the same ratio calculated across all other serotypes in the population
109 [28,29]. However, this statistic is intrinsically imperfect, because even if a serotype
110 causes disease at a consistent rate across populations, the mix of serotypes against
111 which it is compared will differ between locations [30,31]. This geographical
112 heterogeneity has been exacerbated by post-PCV serotype replacement, which has
113 driven increasing divergence between countries' serotype compositions [32]. The
114 interpretability of the odds ratios may be improved by standardising them relative to
115 the geometric mean across all odds ratios [33], but this does not resolve the
116 underlying problem, which likely contributes to the substantial heterogeneity
117 observed for some serotypes when odds ratios are combined in meta-analyses [34].
118 This means carefully-conducted odds ratio analyses can be hampered by having to
119 subsample data, and employ both fixed and random effects analyses, within a single
120 analysis [35]. One solution has been to estimate invasiveness relative to a standard
121 serotype that is common in both carriage and IPD across studied locations [34,36].
122 However, these properties mean such serotypes are likely to be targeted by PCVs.
123 Correspondingly, the original standard, serotype 14 [29], has been eliminated by
124 PCV7 in many settings. Similarly, PCV13 has removed the replacement standard
125 serotype used in post-PCV7 studies, 19A [34].
126

127 Invasiveness has been more directly characterised as the ratio of disease cases
128 relative to the estimated prevalence in carriage [25]. However, this approach did not

account for the stochasticity of disease cases being observed, and assumed the uncertainty associated with the estimated case-to-carrier ratio derived only from the colonisation survey data. More comprehensive quantification of the uncertainty in invasiveness can be achieved with Bayesian frameworks, which have been applied to individual pneumococcal populations [2,37]. Invasiveness has been modelled by jointly fitting distributions to carriage and disease data, then using random effects models to estimate invasiveness in different age groups [37]. An alternative approach, using Poisson regression, was used to estimate serotypes' progression rates for causing otitis media [2].

In this work, we apply Bayesian modelling to estimate progression rates from meta-analyses of multiple studies. The ability to synthesise data across populations is important for increasing the sample size of each of the many serotypes observed in species such as *S. pneumoniae*, particularly as many genotypes that have emerged post-PCV were previously rare, and vary geographically [32,38]. The model enables the estimation of progression rates as an opportunistic pathogen's hazard of progressing from carriage to disease in a specified population. Such absolute values avoid a measure that is relative either to the rest of the local pneumococcal population, or to one standard serotype. These properties are crucial for quantifying and modelling changes in disease incidence, such as following vaccine introduction, or the emergence of a hyperinvasive strain. They can be particularly valuable in settings where disease surveillance is not comprehensive, and only carriage data are available [37,39].

153 Using a Bayesian approach also enables hypothesis testing through statistical
154 comparison of alternative model structures [2,37]. This is crucial for understanding
155 the host and pathogen factors that affect progression rates. For instance, the
156 correlation between pneumococcal serotypes and invasiveness has not been
157 indisputably established as a causal link. In *Haemophilus influenzae*, changes to an
158 isolate's serotype altered its virulence in an animal model of disease in such a
159 manner that reflected the epidemiology of human disease [40]. While some
160 equivalent experiments in *S. pneumoniae* have replicated observations from human
161 IPD [41,42], others have found changes in an isolate's serotype did not change its
162 invasiveness in an animal model [43,44], suggesting serotype-independent factors
163 may contribute to this phenotype. Studies of isolates from disease and carriage
164 using comparative genomic hybridisation [45,46] and whole genome sequencing
165 [47,48] have both supported links between non-capsular loci and invasiveness. An
166 unambiguous association cannot even be made between the distinctive phenotypes
167 of the epidemic serotypes and their capsules, as these pneumococci tend to have
168 low genetic diversity [35,49], and therefore their capsule polysaccharide synthesis
169 locus is in linkage disequilibrium with many other variable genomic loci [50,51].
170
171 Epidemiological studies using genotyped and serotyped isolates have also found
172 examples of within-serotype differences in invasiveness [33,52,53], although other
173 studies found evidence of serotype being the primary determinant of invasiveness
174 [29]. Genomic analysis by the Global Pneumococcal Sequencing project also
175 identified differences in the invasiveness of strains of the same serotype [31].
176 However, given the multiple pairwise comparisons conducted in such analyses, it is
177 not clear whether such observations might be expected, even if the capsule is the

primary determinant of invasiveness. Such questions can be addressed through comparison of different model fits to the data, thereby improving our understanding of the determinants of progression rates. This is essential for deciding on the most effective methods for ongoing pathogen surveillance, and to enable the rapid identification of emerging highly invasive types. Therefore the methods, models and data described in this study are made available as an R package (https://github.com/nickjcroucher/progressionEstimation), to enable the ongoing aggregation of case and carrier studies for any opportunistic pathogens,.

Methods

Model definitions and assumptions

All models were designed to be applied to a meta-population of a multi-strain opportunistic pathogen. Each population *i* corresponded to a matched sample of microbes from asymptomatic carriers, or the environment, and a sample from disease. It was assumed all cases of disease emerged independently from the carried or environmental population. Across the meta-population, the microbes were classified according to their serotype *j*, strain *k*, or the combination of both, *j*,*k*. If *x* denotes any one of these classifications, then the carriage rate in population *i*, $\rho_{i,x}$, was estimated based on the total number of samples (including negative results), η_i, and the number of samples positive for isolates of type *x*, $c_{i,x}$:

$$c_{i,x} \sim Binom(\rho_{i,x}, \eta_i)$$

Modelling the number of positive samples using a Binomial distribution follows the precedent of Weinberger *et al* [37], and assumes all carriage samples are independently drawn from the circulating population.

A set of related statistical models were used to estimate the progression rate [2], v, the hazard of developing a disease given carriage of a microbe per unit time. As v was assumed to be constant for each type x in each study i, the number of observed cases of disease caused by x in i, $d_{i,x}$, was modelled as the product of the number of potential hosts in i, N_i; the proportion carrying x in i, $\rho_{i,x}$; the per unit time probability of x causing disease, v_x; and the duration of the study, t_i. . Hence, the expected number of isolates from disease, $\mathbb{E}[d_{i,x}]$, in all models was proportional to all these quantities:

$$\mathbb{E}[d_{i,x}] \propto \rho_{i,x} N_i \mathrm{t}_i \upsilon_x$$

Some models also adjusted the expected $d_{i,x}$ between i through a scaling parameter, γ_i, that was relative to a reference population. This corrects for differences between studies both in the actual hazard of developing a disease, due to host population variation, and differences in the comprehensiveness of surveillance across the study populations. Hence, the expected number of isolates from disease, $\mathbb{E}[d_{i,x}]$, in such models was proportional the quantities:

$$\mathbb{E}[d_{i,x}] \propto \rho_{i,x} N_i \mathrm{t}_i \upsilon_x \gamma_i$$

All these parameters were assumed to be constant within a study over t_i. All disease cases were assumed to develop independently, and therefore were modelled as a Poisson process. However, relevant unmodelled variation, or oversampling of local transmission chains in either carriage or disease sampling [54], could cause heterogeneity in the observed isolate counts in either component of the paired samples. Hence $d_{i,x}$ was additionally modelled as following a negative binomial distribution, in which overdispersion was quantified by the precision parameter, ϕ, relative to the mean, μ. Lower ϕ values indicate greater overdispersion, as the variance of the distribution was calculated as:

$$Var[d_{i,x}] = \mu + \frac{\mu^2}{\phi}$$

Models of disease prevalence

Four different structures were used to model $\mathbb{E}[d_{i,x}]$. For each, two different versions were fitted to the data. In the simpler version, the observed disease case counts were assumed to follow a Poisson distribution:

$$d_{i,x} \sim Poisson(\mathbb{E}[d_{i,x}])$$

The more complex version assumed a negative binomial distribution of the observed disease case counts:

$$d_{i,x} \sim NegBin(\mathbb{E}[d_{i,x}], \phi)$$

Null model: In this simplest model, the progression rate was independent of the pathogen's type (v_x = v for all x) and host population (γ_i = 1 for all i). Therefore $d_{i,x}$ is expected to be correlated with $c_{i,x}$:

$$\mathbb{E}[d_{i,x}] = \rho_{i,x} N_i \mathrm{t}_i \upsilon$$

Type-specific model: This model, similar to that Weinberger *et al* applied to a single population [37], allowed for variation in progression rate between types, but still assumed host populations to be homogeneous (γ_i = 1 for all i). Therefore $d_{i,x}$ depends on both the carriage prevalence and progression rate of x:

$$\mathbb{E}[d_{i,x}] = \rho_{i,x} N_i \mathrm{t}_i \upsilon_x$$

Study-adjusted model: In this model, variation in disease prevalence between studies reflected differences in host population and surveillance, rather than differences between types. Hence the progression rate was independent of the pathogen's type (v_x = v for all x), but varied between i due to γ_i:

$$\mathbb{E}[d_{i,x}] = \rho_{i,x} N_i \mathrm{t}_i \upsilon \gamma_i$$

Study-adjusted type-specific model: This model allowed for both variation in the progression rate between types x and study populations i:

$$\mathbb{E}[d_{i,x}] = \rho_{i,x} N_i \mathrm{t}_i \upsilon_x \gamma_i$$

275 **Joint modelling of strain and type progression rates**

276 For datasets in which information was available on both type (for *S. pneumoniae*,
277 serotype) *j* and strain *k*, v_x was modelled in five different ways:

278

279 **Serotype-determined progression rate**: The progression rate was entirely
280 determined by an isolate's type, and independent of its strain background ($v_x = v_j$).

281

282 **Strain-determined progression rate**: The progression rate was entirely determined
283 by the strain to which an isolate belonged, and independent of its type ($v_x = v_k$).

284

285 **Strain-modified type-determined progression rate**: Two approaches were taken
286 to modelling progression rates as being independently affected by type and genetic
287 background. The first modelled the progression rate as being primarily determined
288 by an isolate's serotype, which was modified by its strain background. This was
289 calculated as $v_x = v_j v_k$, where the model priors meant v_j was less constrained than v_k.
290 Hence most variation in v_x should be attributed to the type *j*.

291

292 **Type-modified strain-determined progression rate**: The second approach to
293 analysing the independent effects of type and genetic background modelled
294 progression rate as being primarily determined by the strain to which an isolate
295 belonged, which was modified by the serotype it expressed. This was calculated as
296 $v_x = v_j v_k$, where the model priors meant v_k was less constrained than v_j. most
297 variation in v_x should therefore be attributed to the strain *k*.

298

Type and strain-determined progression rate: Each combination of serotype *j* and strain *k* was modelled as having a unique progression rate ($v_x = v_{j,k}$), suggesting non-multiplicative interactions between strain background and type on an isolate's progression rate.

Model priors and implementation

The characteristics of each study population (η_i, N_i, t_i) and observed counts of type *x* in carriage ($c_{i,x}$) and disease ($d_{i,x}$) were assumed to be accurate and without error. As each recovery of a type *x* isolate from carriage contributes to the estimation of $\rho_{i,x}$, even if an individual carries multiple types, the prior distribution was:

$$\rho_{i,x} \sim Beta(1,1)$$

This means each type's carriage was independently estimated, and the sum of carriage prevalences ($\sum_x \rho_{i,x}$) may be greater than one, if there is high level of multiple carriage in a population. The beta distribution was used, as it is the conjugate prior of the binomial distribution, although setting the scale and shape parameters to one make it equivalent to a uniform distribution bounded by zero and one.

Invasiveness has been estimated to vary over orders of magnitude in *S. pneumoniae* [27], and therefore a uniform prior was placed on the logarithm of v_x in all models. The lower bound was set at one IPD case per million carriers per year, as the largest studies used surveillance of host populations consisting of tens of millions of individuals (Table S1). The upper bound was ten cases per carrier per year, which

would effectively correspond to an obligate pathogen for a bacterium with the carriage duration of *S. pneumoniae* [55]:

$$log_{10}(\nu_x) \sim U(-6,1)$$

As the scaling factors were relative measures of disease prevalence between populations, γ_i for the population with the largest sample size was fixed at one. The scaling factor for each other population *i* was allowed to vary higher or lower. The Cauchy distribution was assumed as the prior for the logarithm of γ_i, as this describes the ratio of two normally-distributed random variables. However, the Hamiltonian Monte Carlo algorithm used for model fitting does not efficiently sample heavy-tailed distributions, and therefore the prior was reparameterised to use a random variable Γ_i with a unit-sized uniform prior [56]:

$$\Gamma_i \sim U\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$$

The Γ_i values were then transformed to generate γ_i values with a Cauchy distribution using the location ($\mu_{study} = 0$) and scale ($\tau_{study} = 2$) parameters:

$$log_{10}(\gamma_i) = \mu_{study} + \tau_{study} tan(\Gamma_i))$$

The τ_{study} value allowed for variation in observed incidence over multiple orders of magnitude between datasets, ensuring the model makes allowance for substantial heterogeneity in observed disease incidence [57].

A similar reparameterization of the Cauchy distribution was used for the models in which type and strain background independently contributed to a genotype's progression rate (the strain-modified type-determined, and type-modified strain-determined progression rate models). The categorisation modelled as determining the progression rate (v_x) had the same prior as when a single factor determined the progression rate. This was multiplied by a second value, v_y, determined by the classification modifying the progression rate. The prior distribution for the logarithm of this value was a truncated Cauchy, symmetrical around zero. This represented the expectation that the modification of the progression rate would be small (approximately one) in most cases, but may be substantial in rare instances. The truncation was found to be necessary for efficient sampling with the Hamiltonian Monte Carlo algorithm. This was parameterised with a random variable Υ_y, sampled from a uniform distribution with narrower boundaries than those for the analogous Γ_i variable:

$$\Upsilon_y \sim U(-1.25, 1.25)$$

The Υ_y values were then transformed to generate v_y values using the location ($\mu_{mod} = 0$) and scale ($\tau_{mod} = 0.5$) parameters:

$$log_{10}(v_y) = \mu_{mod} + \tau_{mod} tan(\Upsilon_y))$$

This meant v_y values could vary over three orders of magnitude (between 0.0313 and 32.0), while enabling models to converge with reasonable MCMC lengths.

The precision parameter of the negative binomial distribution, ϕ, was sampled from a uniform distribution. Where there is little evidence of overdispersion, ϕ will be large relative to $\mathbb{E}[d_{i,x}]$, whereas highly overdispersed data can be modelled with a smaller, albeit non-zero, ϕ value. Therefore, the prior distribution was chosen to be broad, but always positive:

$$log_{10}(\phi) \sim U(-3,3)$$

Model parameters are summarised in Table 1. All models were implemented using R [58] and Rstan [59]. Each model was fitted using two Hamiltonian Markov chain Monte Carlo (MCMC) evaluations for 25,000 iterations. Convergence was assessed through analysing the MCMC traces, using the criterion of all $\hat{R}$ values below 1.05, and no reports of divergent transitions in the sampling [56]. This was achieved using default MCMC sampling parameters for the serotype analyses, but the strain analyses required a smaller step size (0.01), higher acceptance probability (0.99) and larger tree depth (20). The data and code used for the analyses are available as an R package from https://github.com/nickjcroucher/progressionEstimation/.

Model evaluation and comparison

Models were compared using likelihoods, independent of priors, through approximate leave-one-out cross-validation with the loo package [60]. The marginal likelihoods of different models, given the data, were compared with Bayes factors calculated using the bridgesampling package [61]. The same number of iterations were used to calculate the logarithmic marginal likelihoods as were used to fit the

398 models. Model outputs were analysed and plotted with tidyverse and ggpubr
399 packages [62,63].
400
401 ***S. pneumoniae* serotype data**
402 The matched carriage and invasive pneumococcal disease serotype datasets were
403 combined from two meta-analyses [27,34]. Twelve of the reported studies were
404 omitted due to lack of publicly-available data [29,64–69]; difficultly in defining
405 independent cross-sectional carriage samples [70,71]; small sample sizes once
406 stratified by age [72], or study design biased towards particular serotypes [73]. If a
407 PCV were introduced during a study, then samples were divided into pre-PCV and
408 post-PCV datasets. This generated 21 paired case and carrier datasets, each with
409 substantial matched systematically-sampled and comprehensively serotyped
410 asymptomatic carriage and disease samples (Table S1). Of these studies, 20
411 included data on child carriage and child IPD, and five included data on child
412 carriage and adult IPD.
413
414 Data that were only resolved to serogroup level, based on the currently-known
415 serotypes, were omitted from the analysis, although the overall number of samples
416 taken (η_i) was not adjusted. However, data on serotypes in historical studies that are
417 now known to correspond to multiple serotypes (e.g. serotype 6A now being
418 resolved into 6A and 6C [74]) were included unaltered. Additionally, isolates of
419 serotypes 15B and 15C were combined into the single 15B/C category.
420
421 ***S. pneumoniae* strain data**

422 Analysis of variation of invasiveness between *S. pneumoniae* strains in children used
423 population genomic data from South Africa [35], a mixture of population genomic and
424 genotyped data from the USA [35], and genotyped data from Finland [71], Oxford
425 [29] and Stockholm [75]. The equivalent analysis of strain invasiveness in adults
426 used data from Portugal [52]. Some of the genotyped datasets were not as
427 thoroughly documented as the serotyped datasets, and therefore it was not
428 appropriate to fit models that lacked a study-specific scale factor that reduced their
429 reliance upon accurate carriage sample information (Text S1).
430
431 **Results**
432
433 **Bayesian meta-analysis of *S. pneumoniae* serotype invasiveness**
434 Twenty matched disease case and nasopharyngeal carrier datasets were used to
435 quantify *S. pneumoniae* serotype invasiveness in children (Table S1). From these,
436 7,340 carriage isolates and 2,851 disease isolates were extracted, across 72
437 serotypes (Fig. S1). Multiple models of invasiveness were proposed to analyse these
438 data (see Methods). The null model assumed there was no systematic difference
439 between serotypes' invasiveness across datasets. The type-specific model assumed
440 each serotype had a characteristic invasiveness that was consistent across
441 locations. The study-adjusted null model assumed invasiveness did not vary across
442 serotypes, but could differ between studies. Finally, the study-adjusted type-specific
443 model allowed invasiveness to vary between both types and studies. All four
444 approaches to estimating the expected numbers of disease cases and carriage
445 isolates were fitted assuming the number of disease cases would follow a Poisson

446 distribution, and allowing for overdispersion by fitting a negative binomial distribution
447 to these values, resulting in eight total models.
448
449 Datapoints at which both the count of both carriage and disease isolates were low for
450 a given type in a study offered limited power for distinguishing the different model
451 structures. Therefore only paired observations from an individual study in which
452 either the number of carried or disease isolates was at least five were used for the
453 model comparison. This reduced the dataset to 7,048 carriage isolates and 2,617
454 disease isolates across 45 serotypes (Fig. S2). The eight models were each fitted to
455 the data using two Hamiltonian Markov chain Monte Carlo samplings, both run for
456 25,000 iterations. All chains appeared to have converged, based on the traces of the
457 posterior probabilities (Fig. S3) and all $\hat{R}$ values being below 1.001 (Fig. S4).
458 Comparisons of the observed and predicted isolate counts showed all models could
459 accurately reproduce the observations from carriage studies, with the exception of
460 the simplest null Poisson model (Fig. 1). However, only the most complex models,
461 adjusting for both *S. pneumoniae* serotype and study, were able to accurately
462 reproduce the disease case counts.
463
464 These deviations can be quantified as root mean squared error (RMSE) values (Fig.
465 S5). These show the negative binomial models replicate the carriage values closely,
466 relative to the equivalent Poisson models, with correspondingly greater deviation
467 from the disease counts enabled by the overdispersion permitted by the negative
468 binomial distribution's precision parameter, ϕ. The low ϕ values for the three simplest
469 negative binomial models suggest there is substantial unmodelled variation (Fig. S6).
470 However, for the study-adjusted type-specific models, the high ϕ value suggests little

471 evidence of overdispersion, and the similar RMSE values for the Poisson and
472 negative binomial versions of this model suggested adjusting for study and serotype
473 enabled disease cases to be modelled using a Poisson distribution (Fig. S5).
474
475 The success of the study-adjusted type-specific models in reproducing the
476 observations may represent overfitting by the most complex models. Therefore
477 formal model comparisons were undertaken using leave-one-out cross validation
478 [60]. Although this suggested the study-adjusted type-specific models were the most
479 appropriate, the large number of model parameters meant the Pareto k statistic
480 diagnostic values were too high for these comparisons to be reliable (Table S2) [76].
481 This was still true if the models were compared using only the likelihoods calculated
482 from the disease counts, which were more constrained than those calculated from
483 the carriage data (Table S3). Hence models were instead compared using Bayes
484 factors calculated from bridge sampling [61]. This found the null models to be the
485 poorest fitting, and identified the study-adjusted type-specific Poisson model to be
486 the most strongly favoured by the data (Table S4).
487
488 **Identification of highly invasive non-vaccine serotypes**
489 The study-adjusted type-specific model was therefore applied to the full dataset.
490 Both the MCMC traces and $\hat{R}$ statistics indicated the model fit converged after
491 25,000 iterations (Fig. S7). This enabled the estimation of invasiveness for 72
492 serotypes (Fig. 2; Dataset S1), including all 24 included in formulations that are
493 currently licensed or under development [20,77,78]. The analysis shows the
494 serotypes associated with high invasiveness and narrow credibility intervals are the
495 three serotypes added to PCV7 to generate PCV10 (1, 5 and 7F), which are also

496 present in higher-valency PCVs, and serotype 12F, included in PCV20. Other
497 serotypes not included in current vaccine designs, but likely to be highly invasive,
498 include 9L, 19C, 28A and 46. However, the credibility intervals of these estimates
499 were substantial, resulting from small overall sample sizes.
500
501 The serotypes associated with low invasiveness and narrow credible intervals were
502 11A (included in PCV20), 6C, 21, 23A, 23B, 34, 35B and 35F. Many other serotypes
503 (e.g. 10F, 11F, 19B, 33A, 35A, 42 and 45) are likely to have low invasiveness, but
504 these estimates have wide credibility intervals. The higher uncertainty generally
505 corresponded with lower sample sizes, although some of these serotypes had a total
506 sample size above 50 (e.g. 10F, 35A, 42), but were nevertheless rare in IPD. The
507 reference population for the model fit was the post-PCV7 sample from the Navajo
508 nation [37], as this study contributed the largest number of isolates (Fig. S1). The
509 model fits found the maximal progression rates were above 10^{-2} IPD cases per
510 carrier per year, with the minimum point estimates approximately 10^{-4} IPD cases per
511 carrier per year. However, the incidence of IPD in the Navajo population is high, with
512 excellent disease surveillance [79], and therefore other studies were expected to
513 report lower disease cases per carriage episode. Correspondingly, the study-
514 adjustment scale factors were below one for most other datasets (Fig. S8),
515 suggesting most locations would expect to detect around half the number of IPD
516 cases per carriage episode for the same carriage serotype composition.
517
518 The fitted study-adjustment scale factors were all associated with similarly narrow
519 credibility intervals, demonstrating there were enough shared types between
520 samples to robustly infer these parameters. For four locations, both pre-PCV and

521 post-PCV samples were included in the meta-analysis, which might be expected to
522 have similar scale factors, if the host population and disease surveillance were
523 consistent across these eras. However, this was not the case for three locations
524 (Atlanta, Bogota and the Netherlands). The model's assumption that serotype
525 invasiveness is constant across studies may be violated by a reduction in vaccine-
526 type invasiveness following PCV introduction, if vaccine-induced immunity inhibited
527 progression from carriage to IPD. Therefore the study-adjusted type-specific Poisson
528 model was refitted, treating vaccine-type serotypes as different types pre- and post-
529 PCV (including 6A as a PCV7 type, and 6C as a PCV13 type [80]; Fig. S9). The
530 point estimates of invasiveness for most vaccine serotypes dropped following PCV
531 introduction, with the evidence strongest for serotypes 14 and 7F falling after the
532 introduction of PCV7 and PCV10, respectively (Fig. 2). This fit meant the scale factor
533 credibility intervals for the Netherlands' pre-PCV7, post-PCV7 and post-PCV10
534 studies overlapped (Fig. S10). These data suggest PCVs can reduce the
535 invasiveness of vaccine types before they are eliminated by herd immunity, and
536 therefore distinguishing between pre- and post-PCV invasiveness may help
537 standardisation across meta-analyses.
538
539 **Comparison of Bayesian modelling with odds ratios**
540 The outputs of the adjusted type-specific Poisson model were compared to those
541 from estimating pneumococcal invasiveness from the same dataset using odds
542 ratios (Fig. 3). Odds ratios were combined across studies using either random (Fig.
543 3A) or fixed effects (Fig. 3B) models, and the results disaggregated based on the
544 total number of isolates, summed across carriage and disease, for each serotype.
545 Across all sample sizes, there was a significant positive correlation between the two

546 measures of invasiveness. The range of invasiveness values extended over two
547 orders of magnitude for both methods. Hence the study-adjusted type-specific
548 Poisson model produced similar relative estimates of invasiveness to odds ratio
549 analyses, while having the advantage of providing both overall, and location-specific,
550 absolute estimates of these progression rates that can be used in quantitative
551 analyses.
552
553 Fixed effects analyses assume there is a single invasiveness value across all
554 locations, whereas random effects analyses allow for variation in the underlying
555 invasiveness estimate between studies. The identification of between-location
556 heterogeneity typically suggests random effects models are more appropriate for
557 analyses of pneumococcal invasiveness [27,34,35]. This variation may either reflect
558 genuine differences in serotype behaviour, or arise from standardisation relative to a
559 variable mix of serotypes between locations. By accounting for this variation, random
560 effects models often calculate wider confidence intervals than fixed effect models
561 (Fig. 3). Additionally, random effects models are typically not recommended where
562 there are fewer than five studies in which a serotype features [81], and therefore
563 meta-analyses of odds ratios may be a mixture of fixed- and random-effects models
564 [35]. By contrast, the Bayesian framework rejected the over-dispersed negative
565 binomial model in favour of the Poisson model (Fig. 1). This may reflect reduced
566 heterogeneity in invasiveness estimates, because standardisation does not use
567 differing mixtures of serotypes, as for odds ratios. This single, consistent model can
568 be applied across serotypes, regardless of their frequency.
569

570 Deviations between the methods in the rank order of invasiveness point estimates
571 were most evident at small sample sizes. This was most notable for serotypes 33A
572 and 36, both of which were associated with relatively high log odds ratios (above
573 zero) and relatively low Bayesian progression rate estimates (below 10^{-3} cases per
574 carrier per year). Both were observed only in carriage: five isolates across four
575 studies for serotype 33A, and one isolate in each of three studies for serotype 36.
576 Whereas logarithmic odds ratio analyses necessarily generate symmetrical
577 confidence intervals, those from the Bayesian analysis were asymmetric, with long
578 tails spanning lower values representing the possibility that the invasiveness of these
579 serotypes may be negligible, as they were not observed to cause IPD. Hence the
580 Bayesian model estimates provide a more informative representation of the limited
581 available data.
582
583 At the upper end of the scale, the Bayesian model found the most highly invasive
584 serotype to be 46, with credibility intervals indicating this serotype is unlikely to be of
585 intermediate invasiveness [28]. By contrast, the confidence intervals for the odds
586 ratios estimated using random effects spanned the full range of invasiveness point
587 estimates across the species. This serotype was observed in two studies, but never
588 isolated from carriage, and there is further circumstantial evidence that serotype 46
589 is likely to be highly invasive (see Discussion). Hence this Bayesian framework can
590 use case-carrier studies to provide informative estimates of invasiveness even from
591 small sample sizes.
592
593 Odds ratios were previously found to correlate with another measure of
594 invasiveness, the “attack rate”, which in turn negatively correlated with carriage

duration [18]. This could result from IPD being most common shortly after the acquisition of a novel serotype in the nasopharynx, which occurs more frequently for serotypes carried for only a short period in each host. The invasiveness values from child IPD data were compared with recent estimates of carriage duration from multi-state modelling of longitudinally-sampled carriage studies in Maela, Thailand [55,82]. Although the highest invasiveness values were associated with short carriage durations, likely as longer carriage duration is expected to result in higher cross-sectional carriage prevalences, there was no strong overall relationship (Fig. S11).

Differences in invasiveness between host age demographics

Although *S. pneumoniae* primarily circulate between children, they frequently cause disease in adults. Five datasets were identified in which adult disease cases could be matched with child nasopharyngeal carrier datasets, to estimate serotype invasiveness in adults. Overall, 3,756 carriage isolates and 3,041 disease isolates were extracted across 53 serotypes (Fig. S12). As with the child IPD samples, this dataset was filtered for observations where either the number of carriage or disease isolates was at least five for the model comparison, to improve the power for identifying the best fitting model. This filtering left in 3,704 carriage isolates and 2,969 disease isolates across 40 serotypes (Fig. S13). As with the child dataset, two Hamiltonian MCMC samplings for 25,000 iterations were used to fit the eight models, which converged based on the traces of the posterior probability (Fig. S14) and $\hat{R}$ values being below 1.001 (Fig. S15). Similar to the analysis of invasiveness in children, comparisons of the observed and predicted isolate counts again showed that only the models adjusting for both *S. pneumoniae* serotype and study were able to accurately reproduce the adult disease case counts (Fig. S16). Bayes factors

concurred that the study-adjusted type-specific Poisson model was the most likely, given the data (Table S5).

Thus, this model was applied to the complete adult disease dataset. MCMC traces and $\hat{R}$ statistics again indicated the model fit converged after 25,000 iterations (Fig. S17). Like the child analysis, the reference population for the model fit was the post-PCV7 sample from the Navajo nation [37]. The high incidence of IPD in this population is again evident in the adult population, and therefore the other studies report lower study-adjustment scale factors (Fig. S18), with most locations expecting to detect around a third or fewer the number of IPD cases per carriage episode for the same carriage serotype composition.

This analysis therefore enabled the estimation of invasiveness of 53 serotypes in adults (Fig. 4A; Dataset S2), including the 24 included in PCV formulations currently licensed or under development. The analysis shows the non-vaccine serotypes associated with high invasiveness and narrow credibility intervals were serotypes 8 and 12F, included in the PCV20 formulation, and the non-vaccine type serotype 13. Other serotypes not included in current vaccine designs, but likely to be highly invasive, include serotypes 9L, 12B and 29. However, the credibility intervals of these estimates were substantial, resulting from small overall sample sizes. The non-PCV serotypes associated with low invasiveness (mean estimates $<10^{-3}$ disease cases per carrier per year) and narrow credible intervals were serotypes 7C, 15A, 16F, 23A, 31, 34, 35A, 35B and 35F.

644 Comparisons with the corresponding child invasiveness estimates (Fig. 4B) showed
645 many PCV serotypes had similar invasiveness estimates in both children and adults
646 [37], with PCV7 and PCV10 types having mid-range and high invasiveness
647 estimates for both children and adults, respectively. Serotype 6C, which is affected
648 by PCV13-induced immunity against the 6A component, was low in both children
649 and adults. The non-PCV serotype 9L appears to be highly invasive in both
650 populations, although this value was associated with considerable uncertainty.
651 However, other non-PCV types showed a discrepancy between the two age groups.
652 For example, serotypes 12B and 29 were highly invasive in adults but not children,
653 though the credible interval ranges were large for both.
654
655 Serotypes included in the PCV15 and PCV20 formulations appeared to have mid-
656 range invasiveness in both age groups, except for the consistently highly invasive
657 serotypes 8 and 12F. This suggests the removal of these two serotypes through herd
658 immunity from higher-valency PCV infant immunisation programmes would likely be
659 beneficial to the adult population [83,84]. However, this would concomitantly reduce
660 the effectiveness of the current PPV23 adult vaccine, due to its overlap in serotype
661 coverage with PCVs meaning it would protect against a lower proportion of the post-
662 vaccine *S. pneumoniae* population [27,85]. Its residual effect would be determined
663 by the prevalence and invasiveness of the non-PCV PPV23 serotypes. Both
664 serotypes 2 and 20 were estimated to be highly invasive for both age groups, albeit
665 with broad credible intervals for adults for the former, due to it only being included in
666 one adult dataset. Other PPV23 serotypes had elevated invasiveness compared to
667 many non-vaccine types, suggesting that PPV23 may offer protection against a

668 limited number of higher-risk serotypes following the implementation of PCV15 or
669 PCV20 infant vaccination programmes.
670
671 **Invasiveness varies between pneumococcal strains and serotypes**
672 *S. pneumoniae* populations are genetically diverse, and can be divided into discrete
673 strains [50,86]. Isolates of a particular serotype are often found across multiple strain
674 backgrounds, and a single strain may be associated with multiple serotypes through
675 switching [51]. To test how each of these characteristics affected pneumococcal
676 progression rates, five different models of isolates' invasiveness were fitted using the
677 study-adjusted type-specific Poisson framework (see Methods). These corresponded
678 to the hypotheses that an isolate's invasiveness was determined by its serotype; by
679 its strain background; primarily by its serotype, but modified by strain background;
680 primarily by its strain background, but modified by its serotype; and determined by
681 the combination of its serotype and strain background. All five models were fitted
682 using Poisson and negative binomial distributions.
683
684 These ten models were used to conduct a meta-analysis of six studies in which both
685 serotype and strain background could be determined, for at least some isolates (Fig.
686 S19). Three of the studies (post-PCV7 and post-PCV13 South Africa; post-PCV7
687 USA) were modified versions of a comparison primarily conducted using genomic
688 data [35]. These were combined with pre-PCV genotyped studies from Oxford, UK
689 [29]; Stockholm, Sweden [75], and Finland [71] (see Text S1 and Table S6). For
690 each dataset *i*, isolates were grouped by both their serotype *j* and strain *k*. Models
691 were fitted to a subset of these samples that excluded rare strain-serotypes (with a
692 maximum $c_{i,j,k}$ or $d_{i,j,k}$ below five; Fig. S20), leaving 38 serotypes, 74 strains, and 99

693 serotype-strain combinations. All models converged on stable set of parameter
694 estimates, as inferred from the traces of logarithmic posterior probability values (Fig.
695 S21) and distribution of $\hat{R}$ values (Fig. S22). All models were able to generate
696 reasonably accurate predictions (Fig. S23). There was little evidence of negative
697 binomial distributions improving the fit of these models, and the precision parameters
698 were correspondingly high, albeit with evidence of greater overdispersion when
699 serotype information was omitted from the model fit (Fig. S24). The lowest RMSEs
700 were associated with the models that accounted for both strain background and
701 serotype, either through modification, or when considered in combination (Fig. S25).
702
703 Model comparison with LOO-CV found the best-performing models were those that
704 combined information on isolates' serotype and strain background (Tables S7 and
705 S8), although the Pareto k statistic diagnostic values were again too high for this
706 comparison to be reliable. Comparisons using Bayes factors agreed that the most
707 likely model was that in which invasiveness was determined individually for each
708 combination of strain and serotype (Table S9). Some carriage sample sizes had to
709 be approximated in this analysis (for the Oxford and Finland studies in particular; see
710 Text S1), but the study adjustment factors were robustly estimated and significantly
711 differed across studies, suggesting the model was able to compensate for this
712 aspect of the data (Fig. S26). To test whether these problems could have affected
713 relative model likelihoods, the model comparison was repeated with a 100-fold
714 change in the sample size values for those studies in which the precise number was
715 not reported. The study-adjustment factor meant the comparisons were relatively
716 insensitive to these changes, and the same model was identified as the most likely,

717 given the data (Table S10). Hence model comparisons demonstrated that neither
718 serotype nor strain background alone determines a genotype's invasiveness.
719
720 **Identification of invasive pneumococcal genotypes**
721 The best-fitting model converged on a stable solution when applied to the full dataset
722 (Fig. S27). The invasiveness estimates can be plotted to analyse variation by
723 serotype within strains (Fig. 5; Dataset S3). Some strains were uniformly low
724 invasiveness (serotypes 11A, 15A and 20A within GPSC22), whereas others had
725 consistently higher invasiveness estimates (serotypes 14, 9V, 19F and 19A within
726 GPSC6). By contrast, 19A had an elevated invasiveness relative to 19F within
727 GPSC1, and relative to serotype 15B/C in GPSC4. Similarly, serotypes 8 and 33F
728 were substantially more invasive than 11A within GPSC3. Hence there was
729 considerable variation in estimated invasiveness both between, and within, strains.
730
731 The same invasiveness estimates were plotted grouped by serotype (Fig. 6). The
732 meta-analysis identified significant evidence to support many of the observations of
733 within-serotype invasiveness variation noted in the individual analyses (Table S11).
734 Heterogeneity was most strongly evident within some of the traditional paediatric
735 types [19] (including 6A, 6B, 14, 19F), as well as 16F and 19A. This may represent
736 these common serotypes being found across a wider diversity of strain backgrounds.
737 However, the estimates for other low invasiveness serotypes disseminated across
738 multiple backgrounds, such as 15B/C, 23F and 35B, were more consistent.
739 Furthermore, some serotypes (such as 1 and 12F) were similarly highly invasive in
740 multiple strains. This implies some serotypes may have a strong effect on an
741 isolate's invasiveness.

742

743 The adjusted type-specific model was applied to the full dataset, including low
744 frequency types, to identify high-risk genotypes that might emerge as causes of IPD
745 post-PCV (Fig. S28). The most invasive genotypes not targeted by current PCVs
746 were mainly of serotypes typically considered invasive, such as 8 and 12F. Yet there
747 were also examples of genotypes expressing serotypes regarded as less invasive
748 (e.g. 10A, 16F, 23B and 35B) with high invasiveness point estimates, albeit with wide
749 credibility intervals reflecting small sample sizes. Most of these are likely to remain
750 rare, or regress to the population-wide mean as more data emerge, but ongoing
751 surveillance to rapidly identify whether any represent a potentially problematic type
752 post-PCV will be valuable.

753

754 An additional genotyped dataset from Portugal (Fig. S29), in which the majority of
755 disease cases were isolated from adults [52], was analysed with the type-specific
756 Poisson model in which invasiveness was specific to each combination of serotype
757 and strain background (Fig. S30). There were few instances of multiple serotypes
758 being isolated from the same strain, but several serotypes were distributed across
759 multiple genetic backgrounds (Fig. S31). This found evidence of heterogeneity in
760 invasiveness within serotypes 3, 14 and 23F. The elevated invasiveness of serotype
761 3 was associated with the GPSC6 strain, found to have high invasiveness when
762 expressing multiple serotypes in children (Fig. 5). For serotypes 14 and 23F, this
763 involved strains that did not substantially differ in the child dataset (Fig. 6),
764 suggesting further within-serotype variation in invasiveness may emerge as studies
765 include a greater diversity of host age groups.

766

767 **Discussion**

768 This study describes novel models for using data from cases and carriers to estimate
769 the hazard of an asymptomatically carried pathogen causing a defined disease over
770 a time period, previously defined as a progression rate [2]. This quantity is crucial for
771 determining the efficacy of partial coverage vaccines, such as the PCVs [26,27], and
772 the flexible framework developed here should be applicable to any diverse microbial
773 population. These models calculate progression rates as absolute quantities, rather
774 than dimensionless ratios as with previous meta-analyses [27,34,36]. These
775 estimates can enable the translation of alterations in microbial population structures
776 to changes in disease incidence, which is crucial for understanding the
777 consequences of strain- or type-specific interventions [27]. Furthermore, given
778 suitable datasets for multiple pathogens, these models could also be used to
779 estimate changes in disease burdens for alterations across microbiota.
780
781 When applied to *S. pneumoniae* serotype data, the best-performing model structures
782 were those in which type-specific progression rates were adjusted by dataset-
783 specific scale factors. In principle, these scale factors should correlate with the
784 burden of disease in a given host population, representing a combination of socio-
785 economic factors and co-morbidities that affect IPD incidence [87,88]. Yet in
786 practice, they will also represent the probability of an isolate causing IPD being
787 included in a study dataset. This is affected by multiple factors, including the
788 probability of attending a hospital participating in the study, the efficiency of retrieving
789 bacteria from blood cultures, and the consistency with which cultured isolates are
790 collected by research centres [87].
791

792 These model structures facilitate comparisons across populations by standardising
793 progression rate estimates between datasets using all common serotypes or strains
794 shared between them. Existing methods for meta-analysis typically standardise data
795 relative to a single type, required to be present in all datasets [34,36], or combine
796 estimates derived from ratios calculated relative to variable mixes of types in
797 different locations. The dataset-adjusted type-specific progression rate models are
798 better suited to the extensive international variation in *S. pneumoniae* population
799 structures [35,38], particularly given their ongoing diversification following vaccine
800 introduction [32]. Comparison with odds ratios showed that estimates were
801 correlated at large sample sizes, demonstrating both methods converged on similar
802 values when data are available on many isolates. However, this comes with the
803 caveat that 16 of the 21 studies in the meta-analysis originated from North America
804 or Europe, with most collected prior to PCV introduction. Hence there is a pressing
805 need for paired case and carrier studies from post-PCV settings, particularly from
806 Asia and Africa. As more diverse populations are combined, the improved
807 standardisation across studies from these Bayesian analyses will become more
808 important.
809
810 At smaller sample sizes, the Bayesian models were able to produce more
811 informative estimates of invasiveness than odds ratios. Information on types that are
812 rare in individual datasets is particularly important for an opportunistic pathogen as
813 diverse as *S. pneumoniae*, in which serotypes can emerge and expand rapidly post-
814 PCV [23,32,83]. A recent meta-analysis using odds ratios highlighted the conclusion
815 that 12F was the only non-PCV13 serotype to have an invasiveness higher than 19A
816 [34], the serotype responsible for much post-PCV7 IPD [32,89]. However, this

817 previous study limited its results to 25 serotypes, omitting less common examples
818 that could rise in frequency when serotype replacement is driven by higher-valency
819 vaccines [32,89]. The Bayesian analysis described in this work identified rare
820 serotypes that may be highly invasive in adults (e.g. 9L, 12B and 29) and children
821 (e.g. 19C, 28A and 46), based on small sample sizes. Serotype 46 has already been
822 observed causing child IPD post-PCV13 within the strain GPSC26, which is more
823 commonly associated with the highly-invasive serotype 12F [31]. These two capsule
824 types are structurally similar [90], which is common for serotypes expressed by
825 isolates of the same strain [51]. Future post-PCV13 surveillance will likely show
826 whether 12F and 46 are truly similarly invasive. As higher-valency PCVs become
827 available, with the potential to remove commonly-carried serotypes such as 11A and
828 15B/C [20,91], reducing the uncertainty in invasiveness estimates for remaining non-
829 vaccine types will be of great importance in understanding the vaccine-associated
830 changes in the incidence of IPD.
831
832 The need for meta-analysis of multiple datasets is exacerbated by the model
833 comparison in this study that concluded that invasiveness is most accurately
834 quantified at the level of strain-serotype combination. This conclusion is consistent
835 with the results of some individual studies [35,52,53], although not all [29]. Such a
836 result raises the spectre of separately estimating invasiveness for multiple serotypes
837 within the hundreds of known *S. pneumoniae* strains, many of which are rare in
838 individual populations [35,86]. However, that such models fit better than those using
839 only strain or serotype as a predictor only implies that neither factor alone
840 determines an isolate's invasiveness. The improved fit might simply represent the
841 benefit of more precise classification of bacterial genotypes, and therefore remains

842 possible that some other component of the *S. pneumoniae* genome that varies over
843 short, within-strain timescales may additionally affect, or entirely determine, isolates'
844 invasiveness [50,92,93]. This could explain the heterogeneity of invasiveness within
845 strain GPSC3, where the serotypes are separated into distinct clades
846 (https://microreact.org/project/gpsGPSC3), in contrast to the comparative
847 homogeneity of invasiveness within GPSC6 (containing the PMEN3, or Spain^{9V}-3,
848 lineage), where a single lineages dominates the strain's structure [94].
849
850 However, it seems unlikely that serotypes and invasiveness are not causally linked.
851 Some serotypes (e.g. 1, 12F) exhibited high invasiveness across multiple strain
852 backgrounds, whereas others were consistently low (e.g. 15B/C, 35F). This is
853 consistent with the model fitting to serotype data, in which there was not a
854 substantial improvement of model fitting when using negative binomial versions of
855 the models. Additionally, the associated precision parameters were large when
856 dataset-adjusted models were fitted (Fig. S6). If other loci were strongly affecting
857 invasiveness, then a greater degree of overdispersion in the observed incidence of
858 IPD might be expected, unless they were in linkage disequilibrium with the locus
859 determining serotype.
860
861 By contrast, some paediatric serotypes exhibited a more heterogeneous distribution
862 of invasiveness values. This may reflect these capsules having little effect on
863 invasiveness, or their greater prevalence in the pre-PCV pneumococcal population
864 providing greater power to detect variation between strains [30]. Heterogeneity has
865 previously been noted in serotype 14, with both GPSC6 [52,71] and GPSC18 [29,35]
866 identified as highly invasive clones (Table S11). In this analysis, both were found to

867 be more invasive than GPSC9 within the same serotype. This analysis estimated
868 high invasiveness values for all serotypes within GPSC6, including an elevated
869 invasiveness within serotype 3 in the Portuguese dataset. Similarly, invasiveness
870 was consistently low for serotypes 14 and 15A for GPSC9. This could result in
871 heterogeneous vaccine impacts between regions, as elimination of serotype 14 by
872 PCVs will likely have a greater benefit across Europe and South America, where the
873 highly-invasive GPSC6 is mainly distributed
874 (https://microreact.org/project/gpsGPSC6), rather than Africa and India, where the
875 less invasive GPSC9 is common (https://microreact.org/project/gpsGPSC9).
876
877 Whether these models are useful long-term, or will ultimately be superseded by
878 alternatives based on genome-wide association studies, will depend on the extent to
879 which progression rates are affected by interactions between genetic loci [95]. It may
880 prove possible to explain a high proportion of the variation in invasiveness between
881 microbes through a tractable number of polymorphic loci that each independently
882 contribute to a microbe's invasiveness using genome-wide association studies [45–
883 48]. If so, then the individual estimates for each of the multitude of subtype
884 categorisations can be replaced with a simpler genome-based approach, which
885 would be capable of predicting the invasiveness of previously unseen genotypes.
886 However, if invasiveness depends on combinations of loci [95], it may only be
887 possible to estimate progression rates for extant common types from epidemiological
888 studies. Other alternative model structures include random effects models and
889 multivariate regression analyses, which can incorporate information on the host
890 population characteristics [2,37,96]. However, these models typically require
891 comprehensive surveillance of disease and detailed information on host populations,

892 which limits the range of studies that can be combined in meta-analyses of diverse
893 bacterial populations.
894
895 Even using more conventional approaches to estimating invasiveness, whole
896 genome sequencing clearly represents an important tool in future case-carrier
897 studies of opportunistic pathogen invasiveness. In addition to assigning isolates to
898 strains, genomic data is a reliable and cost-effective means of classifying microbes
899 according to their capsular structures [97,98]. Furthermore, it also enables the
900 extraction of additional clinically-relevant information, such as antibiotic resistance
901 loci [94,99], that may be incorporated into future extensions of these models.
902 However, the use of genomic data exacerbates an underlying problem with these
903 models, as high-quality genome assemblies are not necessarily generated from all
904 isolates recovered from carriage. The model assumes there are no false negative
905 swabs in carriage studies, meaning there is an expectation that all carriage samples
906 will be present in the genomic data. Hence the omission of genomic data adds to a
907 false negative rate that, even with serotype data, already reflects the imperfect
908 detection of colonisation by nasopharyngeal swabbing. This can be the consequence
909 of all resident bacteria being missed, or the difficulty of detecting lower prevalence
910 types in instances of multiple carriage [100]. Unless there is variation in the ability to
911 detect carriage of different types, the underestimation of carriage means all
912 progression rates will be uniformly overestimated. However, if only a subset of
913 datasets in a meta-analysis has a lower sensitivity for detecting carriage, the model
914 will adjust for this by increasing the associated scale factors. Hence the model
915 structure is sufficiently flexible to account for such differences. This is necessary, as
916 even with standard protocols for detecting colonisation, the population-wide

917 estimates for carriage rates for a location will depend on the exact demographics
918 sampled.
919
920 Some of these problems with detecting colonisation can be addressed using new
921 techniques with improved sensitivity for detecting multiple serotypes in carriage
922 [101]. However, to simultaneously obtain the genotyping information needed to most
923 precisely characterise an isolate's invasiveness, deep genome sequencing, or
924 similar high-sensitivity molecular genotyping techniques, will be necessary. Such
925 data will not only improve the input to epidemiological models, such as those
926 outlined here, but also potential future models based on combinations of genetic loci.
927 Yet meta-analyses of case-carrier studies can also be improved by simple changes
928 to reporting, without any alterations to current procedures. The data and models
929 described in this work are made available for re-use, modification, and application to
930 other multi-strain microbes (see Methods). The fitting of the model to new datasets
931 will require studies to report their raw data in a standard format, and include the total
932 number of swabs included in the carriage study, as well as an informed estimate of
933 the size of the population under surveillance for IPD. To improve comparability
934 between future studies, it would also be ideal for studies to stratify these data by age
935 category, to enable specific subsets to be employed in different meta-analyses
936 studying specific demographics. Therefore, as the repertoire of vaccines against
937 diverse pathogens expands, we can hope to monitor and understand their impact on
938 population-wide microbial invasiveness using improved models, advances in
939 sequencing technology, and greater transparency of reporting.
940
941 **Acknowledgements**

We thank Raquel Sá-Leão, Mário Ramirez, Bill Hanage and Birgitta Henriques-Normark for providing further information on case-carrier datasets, and Daniel Weinberger for helpful discussions on model structure and fitting. AL was funded by an investigator-initiated grant from GSK to NJC. NJC was supported by the UK Medical Research Council and Department for International Development (grants MR/R015600/1 and MR/T016434/1) and a Sir Henry Dale Fellowship, jointly funded by Wellcome and the Royal Society (grant no. 104169/Z/14/A). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

We have read the journal's policy and the authors of this manuscript have the following competing interests: this work was partially funded by GSK, who manufacture the PCV10 vaccine, although they had no role in study design, data collection or analysis.

Data Availability

All data and code used for the described analyses are available in a GitHub repository at https://github.com/nickjcroucher/progressionEstimation/. This repository has been assigned the DOI 10.5281/zenodo.5154066.

962 **References**

963

964 1. Brown SP, Cornforth DM, Mideo N. Evolution of virulence in opportunistic
965 pathogens: generalism, plasticity, and control. Trends Microbiol. 2012;20: 336–
966 342. doi:10.1016/j.tim.2012.04.005
967 2. Lewnard JA, Givon-Lavi N, Weinberger DM, Lipsitch M, Dagan R. Pan-
968 serotype reduction in progression of *Streptococcus pneumoniae* to otitis media
969 after rollout of pneumococcal conjugate vaccines. Clin Infect Dis. 2017;65:
970 1853–1861. doi:10.1093/cid/cix673
971 3. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical
972 microbiology with bacterial genome sequencing. Nat Rev Genet. 2012;13:
973 601–612. doi:10.1038/nrg3226
974 4. Feil EJ, Enright MC. Analyses of clonality and the evolution of bacterial
975 pathogens. Curr Opin Microbiol. 2004;7: 308–313.
976 doi:10.1016/j.mib.2004.04.002
977 5. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo S, Weiser JN, et al. Fast
978 and flexible bacterial genomic epidemiology with PopPUNK. Genome Res.
979 2019;29: 304–316. doi:10.1101/360917
980 6. Cartman ST, Heap JT, Kuehne SA, Cockayne A, Minton NP. The emergence
981 of “hypervirulence” in *Clostridium difficile*. Int J Med Microbiol. 2010;300: 387–
982 395. doi:10.1016/j.ijmm.2010.04.008
983 7. Fearnley C, Manning G, Bagnall M, Javed MA, Wassenaar TM, Newell DG.
984 Identification of hyperinvasive *Campylobacter jejuni* strains isolated from
985 poultry and human clinical sources. J Med Microbiol. 2008;57: 570–580.
986 doi:10.1099/jmm.0.47803-0

987 8. Knol MJ, Hahné SJM, Lucidarme J, Campbell H, de Melker HE, Gray SJ, et al.
988 Temporal associations between national outbreaks of meningococcal
989 serogroup W and C disease in the Netherlands and England: an observational
990 cohort study. Lancet Public Heal. 2017;2: e473–e482. doi:10.1016/S2468-
991 2667(17)30157-3
992 9. Lynskey NN, Jauneikaite E, Li HK, Zhi X, Turner CE, Mosavie M, et al.
993 Emergence of dominant toxigenic M1T1 *Streptococcus pyogenes* clone during
994 increased scarlet fever activity in England: a population-based molecular
995 epidemiological study. Lancet Infect Dis. 2019;19: 1209–1218.
996 doi:10.1016/S1473-3099(19)30446-3
997 10. Du P, Zhang Y, Chen C. Emergence of carbapenem-resistant hypervirulent
998 *Klebsiella pneumoniae*. Lancet Infect Dis. 2018;18: 23–24. doi:10.1016/S1473-
999 3099(17)30625-4
1000 11. Ganaie F, Saad JS, McGee L, van Tonder AJ, Bentley SD, Lo SW, et al. A
1001 new pneumococcal capsule type, 10D, is the 100th serotype and has a large
1002 *cps* fragment from an oral streptococcus. MBio. 2020;11: e00937-20.
1003 doi:10.1128/mBio.00937-20
1004 12. Ganaie F, Maruhn K, Li C, Porambo RJ, Elverdal PL, Abeygunwardana C, et
1005 al. Structural, genetic, and serological elucidation of *Streptococcus*
1006 *pneumoniae* serogroup 24 serotypes: Discovery of a new serotype, 24C, with
1007 a variable capsule structure. J Clin Microbiol. 2021; JCM.00540-21.
1008 doi:10.1128/JCM.00540-21
1009 13. Geno KA, Gilbert GL, Song JY, Skovsted IC, Klugman KP, Jones C, et al.
1010 Pneumococcal capsules and their types: Past, present, and future. Clin
1011 Microbiol Rev. 2015;28: 871–899. doi:10.1128/CMR.00024-15

1012 14. Nelson AL, Roche AM, Gould JM, Chim K, Ratner AJ, Weiser JN. Capsule
1013 enhances pneumococcal colonization by limiting mucus-mediated clearance.
1014 Infect Immun. 2007;75: 83–90. doi:10.1128/IAI.01475-06
1015 15. Hyams C, Camberlein E, Cohen JM, Bax K, Brown JS. The *Streptococcus*
1016 *pneumoniae* capsule inhibits complement activity and neutrophil phagocytosis
1017 by multiple mechanisms. Infect Immun. 2009/12/02. 2010;78: 704–715.
1018 doi:IAI.00881-09 [pii]10.1128/IAI.00881-09 [doi]
1019 16. Barker J, Gratten M, Riley M, Lehmann D, Montgomery J, Kajoi M, et al.
1020 Pneumonia in children in the eastern highlands of Papua New Guinea: A
1021 bacteriologic study of patients selected by standard clinical criteria. J Infect
1022 Dis. 1989;159: 348–352. doi:10.1093/infdis/159.2.348
1023 17. Eyre JW, Washbourn JW. Varities and virulence of the pneumococcus. Lancet.
1024 1899;153: 19–22. doi:10.1016/S0140-6736(01)78950-7
1025 18. Sleeman KL, Griffiths D, Shackley F, Diggle L, Gupta S, Maiden MC, et al.
1026 Capsular serotype-specific attack rates and duration of carriage of
1027 *Streptococcus pneumoniae* in a population of children. J Infect Dis. 2006;194:
1028 682–688. doi:10.1086/505710
1029 19. Hausdorff WP, Feikin DR, Klugman KP. Epidemiological differences among
1030 pneumococcal serotypes. Lancet Infect Dis. 2005;5: 83–93.
1031 doi:10.1016/S1473-3099(05)01280-6
1032 20. Croucher NJ, Løchen A, Bentley SD. Pneumococcal Vaccines: Host
1033 Interactions, Population Dynamics, and Design Principles. Annu Rev Microbiol.
1034 2018;72: 521–549. doi:10.1146/annurev-micro-090817-062338
1035 21. Miller E, Andrews NJ, Waight PA, Slack MPE, George RC. Herd immunity and
1036 serotype replacement 4 years after seven-valent pneumococcal conjugate

1037 vaccination in England and Wales: An observational cohort study. Lancet
1038 Infect Dis. 2011;11: 760–68. doi:10.1016/S1473-3099(11)70090-1
1039 22. Weinberger DM, Malley R, Lipsitch M. Serotype replacement in disease after
1040 pneumococcal vaccination. The Lancet. 2011. pp. 1962–1973.
1041 doi:10.1016/S0140-6736(10)62225-8
1042 23. Lewnard JA, Hanage WP. Making sense of differences in pneumococcal
1043 serotype replacement. Lancet Infect Dis. 2019. doi:10.1016/s1473-
1044 3099(18)30660-1
1045 24. Devine VT, Cleary DW, Jefferies JMC, Anderson R, Morris DE, Tuck AC, et al.
1046 The rise and fall of pneumococcal serotypes carried in the PCV era. Vaccine.
1047 2017;35: 1293–1298. doi:10.1016/j.vaccine.2017.01.035
1048 25. Southern J, Andrews N, Sandu P, Sheppard CL, Waight PA, Fry NK, et al.
1049 Pneumococcal carriage in children and their household contacts six years after
1050 introduction of the 13-valent pneumococcal conjugate vaccine in England.
1051 PLoS One. 2018;13: e0195799. doi:10.1371/journal.pone.0195799
1052 26. Nurhonen M, Auranen K. Optimal Serotype Compositions for Pneumococcal
1053 Conjugate Vaccination under Serotype Replacement. PLoS Comput Biol.
1054 2014;10: e1003477. doi:10.1371/journal.pcbi.1003477
1055 27. Colijn C, Corander J, Croucher NJ. Designing ecologically optimized
1056 pneumococcal vaccines using population genomics. Nat Microbiol. 2020;5:
1057 473–485. doi:10.1038/s41564-019-0651-y
1058 28. Smith T, Lehmann D, Montgomery J, Gratten M, Riley ID, Alpers MP.
1059 Acquisition and invasiveness of different serotypes of *Streptococcus*
1060 *pneumoniae* in young children. Epidemiol Infect. 1993;111: 27–39.
1061 doi:10.1017/S0950268800056648

1062 29. Brueggemann AB, Griffiths DT, Peto T, Meats E, Crook DW, Spratt BG. Clonal
1063 Relationships between Invasive and Carriage *Streptococcus pneumoniae* and
1064 Serotype- and Clone-Specific Differences in Invasive Disease Potential. J
1065 Infect Dis. 2003;187: 1424–32. doi:10.1086/374624
1066 30. Johnson HL, Deloria-Knoll M, Levine OS, Stoszek SK, Hance LF, Reithinger
1067 R, et al. Systematic evaluation of serotypes causing invasive pneumococcal
1068 disease among children under five: The pneumococcal global serotype project.
1069 PLoS Med. 2010;7: e1000348. doi:10.1371/journal.pmed.1000348
1070 31. Lo SW, Gladstone RA, van Tonder AJ, Lees JA, du Plessis M, Benisty R, et al.
1071 Pneumococcal lineages associated with serotype replacement and antibiotic
1072 resistance in childhood invasive pneumococcal disease in the post-PCV13 era:
1073 an international whole-genome sequencing study. Lancet Infect Dis. 2019;19:
1074 759–769. doi:10.1016/S1473-3099(19)30297-X
1075 32. Løchen A, Croucher NJ, Anderson RM. Divergent serotype replacement trends
1076 and increasing diversity in pneumococcal disease in high income settings
1077 reduce the benefit of expanding vaccine valency. Sci Rep. 2020;10: 18977.
1078 doi:10.1038/s41598-020-75691-5
1079 33. Hanage WP, Auranen K, Syrjanen R, Herva E, Makela PH, Kilpi T, et al. Ability
1080 of pneumococcal serotypes and clones to cause acute otitis media:
1081 implications for the prevention of otitis media by conjugate vaccines. Infect
1082 Immun. 2004;72: 76–81.
1083 34. Balsells E, Dagan R, Yildirim I, Gounder PP, Steens A, Muñoz-Almagro C, et
1084 al. The relative invasive disease potential of *Streptococcus pneumoniae*
1085 among children after PCV introduction: A systematic review and meta-
1086 analysis. J Infect. 2018;77: 368–378. doi:10.1016/j.jinf.2018.06.004

1087 35. Gladstone RA, Lo SW, Lees JA, Croucher NJ, van Tonder AJ, Corander J, et
1088 al. International genomic definition of pneumococcal lineages, to contextualise
1089 disease, antibiotic resistance and vaccine impact. EBioMedicine. 2019;43:
1090 338–346. doi:10.1016/j.ebiom.2019.04.021
1091 36. Brueggemann AB, Peto TEA, Crook DW, Butler JC, Kristinsson KG, Spratt
1092 BG. Temporal and Geographic Stability of the Serogroup-Specific Invasive
1093 Disease Potential of *Streptococcus pneumoniae* in Children. J Infect Dis.
1094 2004/09/04. 2004;190: 1203–1211. doi:10.1086/423820
1095 37. Weinberger DM, Grant LR, Weatherholtz RC, Warren JL, O'Brien KL, Hammitt
1096 LL. Relating Pneumococcal Carriage among Children to Disease Rates among
1097 Adults before and after the Introduction of Conjugate Vaccines. Am J
1098 Epidemiol. 2016;183: 1055–62. doi:10.1093/aje/kwv283
1099 38. Corander J, Fraser C, Gutmann MU, Arnold B, Hanage WP, Bentley SD, et al.
1100 Frequency-dependent selection in vaccine-associated pneumococcal
1101 population dynamics. Nat Ecol Evol. 2017;1: 1950–1960. doi:10.1038/s41559-
1102 017-0337-x
1103 39. Weinberger DM, Bruden DT, Grant LR, Lipsitch M, O'Brien KL, Pelton SI, et al.
1104 Using pneumococcal carriage data to monitor postvaccination changes in
1105 invasive disease. Am J Epidemiol. 2013;178: 1488–1495.
1106 doi:10.1093/aje/kwt156
1107 40. Zwahlen A, Winkelstein JA, Moxon ER. Surface Determinants of *Haemophilus*
1108 *influenzae* Pathogenicity: Comparative Virulence of Capsular Transformants in
1109 Normal and Complement-Depleted Rats. J Infect Dis. 1983;148: 385–394.
1110 doi:10.1093/infdis/148.3.385
1111 41. Kelly T, Dillard JP, Yother J. Effect of genetic switching of capsular type on

1112 virulence of *Streptococcus pneumoniae*. Infect Immun. 1994;62: 1813–1819.
1113 Available: https://iai.asm.org/content/iai/62/5/1813.full.pdf
1114 42. Hathaway LJ, Grandgirard D, Valente LG, Täuber MG, Leib SL. *Streptococcus*
1115 *pneumoniae* capsule determines disease severity in experimental
1116 pneumococcal meningitis. Open Biol. 2016;6: 150269.
1117 doi:10.1098/rsob.150269
1118 43. Hu FZ, Eutsey R, Ahmed A, Frazao N, Powell E, Hiller NL, et al. *In vivo*
1119 capsular switch in *Streptococcus pneumoniae* - Analysis by whole genome
1120 sequencing. PLoS One. 2012;7: e47983–e47983.
1121 doi:10.1371/journal.pone.0047983
1122 44. Mizrachi Nebenzahl Y, Porat N, Lifshitz S, Novick S, Levi A, Ling E, et al.
1123 Virulence of *Streptococcus pneumoniae* may be determined independently of
1124 capsular polysaccharide. FEMS Microbiol Lett. 2004;233: 147–152.
1125 doi:10.1016/j.femsle.2004.02.003
1126 45. Obert C, Sublett J, Kaushal D, Hinojosa E, Barton T, Tuomanen EI, et al.
1127 Identification of a candidate *Streptococcus pneumoniae* core genome and
1128 regions of diversity correlated with invasive pneumococcal disease. Infect
1129 Immun. 2006;74: 4766–4777. doi:10.1128/IAI.00316-06
1130 46. Blomberg C, Dagerhamn J, Dahlberg S, Browall S, Fernebro J, Albiger B, et al.
1131 Pattern of Accessory Regions and Invasive Disease Potential in *Streptococcus*
1132 *pneumoniae*. J Infect Dis. 2009;199: 1032–1042. doi:10.1086/597205
1133 47. Lees JA, Ferwerda B, Kremer PHC, Wheeler NE, Serón MV, Croucher NJ, et
1134 al. Joint sequencing of human and pathogen genomes reveals the genetics of
1135 pneumococcal meningitis. Nat Commun. 2019;10: 2176. doi:10.1038/s41467-
1136 019-09976-3

1137 48. Cremers AJ, Mobegi F, van der Gaast –de Jongh C, van Weert M, van
1138 Opzeeland F, Vehkala M, et al. The contribution of genetic variation of
1139 *Streptococcus pneumoniae* to the clinical manifestation of invasive
1140 pneumococcal disease. Clin Infect Dis. 2019;68: 61–69. doi:10.1101/169722
1141 49. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli S V, et al.
1142 Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and
1143 closely related species. Genome Biol. 2010;11: R107–R107. doi:10.1186/gb-
1144 2010-11-10-r107
1145 50. Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD,
1146 Hanage WP. Diversification of bacterial genome content through distinct
1147 mechanisms over different timescales. Nat Commun. 2014;5: 5471.
1148 51. Croucher NJ, Kagedan L, Thompson CM, Parkhill J, Bentley SD, Finkelstein
1149 JA, et al. Selective and Genetic Constraints on Pneumococcal Serotype
1150 Switching. PLoS Genet. 2015;11: e1005095.
1151 doi:10.1371/journal.pgen.1005095
1152 52. Sá-Leao R, Pinto F, Aguiar S, Nunes S, Carriço JAJA, Frazao N, et al.
1153 Analysis of invasiveness of pneumococcal serotypes and clones circulating in
1154 Portugal before widespread use of conjugate vaccines reveals heterogeneous
1155 behavior of clones expressing the same serotype. J Clin Microbiol. 2011;49:
1156 1369–75. doi:10.1128/jcm.01763-10
1157 53. Browall S, Norman M, Tångrot J, Galanis I, Sjöström K, Dagerhamn J, et al.
1158 Intraclonal Variations Among *Streptococcus pneumoniae* Isolates Influence the
1159 Likelihood of Invasive Disease in Children. J Infect Dis. 2014;209: 377–388.
1160 doi:10.1093/infdis/jit481
1161 54. Fraser C, Hanage WP, Spratt BG. Neutral microepidemic evolution of bacterial

1162 pathogens. Proc Natl Acad Sci U S A. 2005;102: 1968–1973.
1163 doi:10.1073/pnas.0406993102
1164 55. Lees JA, Croucher NJ, Goldblatt D, Nosten F, Parkhill J, Turner C, et al.
1165 Genome-wide identification of lineage and locus specific variation associated
1166 with pneumococcal carriage duration. Elife. 2017;6: e26255.
1167 doi:10.7554/eLife.26255
1168 56. Stan Development Team. Stan User's Guide, Version 2.27. In: Interaction Flow
1169 Modeling Language [Internet]. 2020 p. 23.7. Available: https://mc-
1170 stan.org/docs/2_27/stan-users-guide/reparameterization-section.html
1171 57. Wahl B, O'Brien KL, Greenbaum A, Majumder A, Liu L, Chu Y, et al. Burden of
1172 *Streptococcus pneumoniae* and *Haemophilus influenzae* type b disease in
1173 children in the era of conjugate vaccines: global, regional, and national
1174 estimates for 2000–15. Lancet Glob Heal. 2018;6: E744–E757.
1175 doi:10.1016/S2214-109X(18)30247-X
1176 58. R Core Team. R: A Language and Environment for Statistical Computing. R
1177 Found Stat Comput. 2019. doi:10.1007/978-3-540-74686-7
1178 59. Stan Development Team. RStan: the R interface to Stan. 2020.
1179 60. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using
1180 leave-one-out cross-validation and WAIC. Stat Comput. 2017;27: 1413–1432.
1181 doi:10.1007/s11222-016-9696-4
1182 61. Gronau QF, Singmann H, Wagenmakers EJ. Bridgesampling: An R package
1183 for estimating normalizing constants. J Stat Softw. 2020;92: 1–29.
1184 doi:10.18637/jss.v092.i10
1185 62. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al.
1186 Welcome to the Tidyverse. J Open Source Softw. 2019;4: 1686.

1187 doi:10.21105/joss.01686
1188 63. Kassambara A. ggpubr: “ggplot2” Based Publication Ready Plots. 2020.
1189 Available: https://cran.r-project.org/package=ggpubr
1190 64. Valles X, Roca A, Lozano F, Morais L, Suarez B, Casals F, et al. Serotype-
1191 specific pneumococcal disease may be influenced by mannose-binding lectin
1192 deficiency. Eur Respir J. 2010;36: 856–863. doi:10.1183/09031936.00171409
1193 65. Principi N, Marchisio P. Epidemiology of *Streptococcus pneumoniae* in Italian
1194 children. Acta Paediatr Suppl. 2000;89: 40–43.
1195 66. Smart LE, Platt DJ, Timbury MC. A comparison of the distribution of
1196 pneumococcal types in systemic disease and the upper respiratory tract in
1197 adults and children. Epidemiol Infect. 1987;98: 203–209.
1198 67. Mogdasy MC, Camou T, Fajardo C, Hortal M. Colonizing and invasive strains
1199 of *Streptococcus pneumoniae* in Uruguayan children: type distribution and
1200 patterns of antibiotic resistance. Pediatr Infect Dis J. 1992;11: 648–652.
1201 68. Scott J, Hall A, Hannington A, Edwards R, Mwarumba S, Lowe B, et al.
1202 Serotype distribution and prevalence of resistance to benzylpenicillin in three
1203 representative populations of *Streptococcus pneumoniae* isolates from the
1204 coast of Kenya. Clin Infect Dis. 1998;27: 1442–50.
1205 69. Levidiotou S, Vrioni G, Tzanakaki G, Pappa C, Gesouli H, Gartzonika C, et al.
1206 Serotype distribution of *Streptococcus pneumoniae* in north-western Greece
1207 and implications for a vaccination programme. FEMS Immunol Med Microbiol.
1208 2006;48: 179–82. doi:10.1111/j.1574-695X.2006.00126.x
1209 70. Azzari C, Cortimiglia M, Nieddu F, Moriondo M, Indolfi G, Mattei R, et al.
1210 Pneumococcal serotype distribution in adults with invasive disease and in
1211 carrier children in Italy: Should we expect herd protection of adults through

1212 infants' vaccination? Hum Vaccin Immunother. 2016;12: 344–350.
1213 doi:10.1080/21645515.2015.1102811
1214 71. Hanage WP, Kaijalainen TH, Syrjänen RK, Auranen K, Leinonen M, Mäkelä
1215 PH, et al. Invasiveness of serotypes and clones of *Streptococcus pneumoniae*
1216 among children in Finland. Infect Immun. 2005;73: 431–5.
1217 doi:10.1128/IAI.73.1.431-435.2005
1218 72. Cullotta AR, Kalter HD, Delgado J, Gilman RH, Facklam RR, Velapatino B, et
1219 al. Antimicrobial susceptibilities and serotype distribution of *Streptococcus*
1220 *pneumoniae* isolates from a Low socioeconomic area in Lima, Peru. Clin Diagn
1221 Lab Immunol. 2002;9: 1328–1331.
1222 73. Takala AK, Vuopio-Varkila J, Tarkka E, Leinonen M, Musser JM. Subtyping of
1223 common pediatric pneumococcal serotypes from invasive disease and
1224 pharyngeal carriage in Finland. J Infect Dis. 1996;173: 128–135.
1225 74. In HP, Pritchard DG, Cartee R, Brandao A, Brandileone MCC, Nahm MH.
1226 Discovery of a new capsular serotype (6C) within serogroup 6 of
1227 *Streptococcus pneumoniae*. J Clin Microbiol. 2007;45: 1225–1233.
1228 doi:10.1128/JCM.02199-06
1229 75. Browall S, Backhaus E, Naucler P, Galanis I, Sjöström K, Karlsson D, et al.
1230 Clinical manifestations of invasive pneumococcal disease by vaccine and non-
1231 vaccine types. Eur Respir J. 2014;44: 1646–57.
1232 doi:10.1183/09031936.00080814
1233 76. Vehtari A, Simpson D, Gelman A, Yao Y, Gabry J. Pareto Smoothed
1234 Importance Sampling. arXiv. 2021; 1507.02646v7.
1235 77. Stacey HL, Rosen J, Peterson JT, Williams-Diaz A, Gakhar V, Sterling TM, et
1236 al. Safety and immunogenicity of 15-valent pneumococcal conjugate vaccine

1237 (PCV-15) compared to PCV-13 in healthy older adults. Hum Vaccines
1238 Immunother. 2019;15: 530–539. doi:10.1080/21645515.2018.1532249
1239 78. Hurley D, Griffin C, Young M, Scott DA, Pride MW, Scully IL, et al. Safety,
1240 Tolerability, and Immunogenicity of a 20-Valent Pneumococcal Conjugate
1241 Vaccine (PCV20) in Adults 60 to 64 Years of Age. Clin Infect Dis. 2020;
1242 ciaa1045. doi:10.1093/cid/ciaa1045
1243 79. O'Brien KL, Shaw J, Weatherholtz R, Reid R, Watt J, Croll J, et al.
1244 Epidemiology of invasive *Streptococcus pneumoniae* among Navajo children
1245 in the era before use of conjugate pneumococcal vaccines, 1989-1996. Am J
1246 Epidemiol. 2004;160: 270–8. doi:10.1093/aje/kwh191
1247 80. Ladhani SN, Collins S, Djennad A, Sheppard CL, Borrow R, Fry NK, et al.
1248 Rapid increase in non-vaccine serotypes causing invasive pneumococcal
1249 disease in England and Wales, 2000–17: a prospective national observational
1250 cohort study. Lancet Infect Dis. 2018;18: 441–451. doi:10.1016/S1473-
1251 3099(18)30052-5
1252 81. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, et
1253 al. Generalized linear mixed models: a practical guide for ecology and
1254 evolution. Trends Ecol Evol. 2009;24: 127–135. doi:10.1016/j.tree.2008.10.008
1255 82. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et
1256 al. Dense genomic sampling identifies highways of pneumococcal
1257 recombination. Nat Genet. 2014;46: 305–309. doi:10.1038/ng.2895
1258 83. Feikin DR, Kagucia EW, Loo JD, Link-Gelles R, Puhan MA, Cherian T, et al.
1259 Serotype-Specific Changes in Invasive Pneumococcal Disease after
1260 Pneumococcal Conjugate Vaccine Introduction: A Pooled Analysis of Multiple
1261 Surveillance Sites. PLoS Med. 2013;10: e1001517.

1262 doi:10.1371/journal.pmed.1001517
1263 84. Shiri T, Datta S, Madan J, Tsertsvadze A, Royle P, Keeling MJ, et al. Indirect
1264 effects of childhood pneumococcal conjugate vaccination on invasive
1265 pneumococcal disease: a systematic review and meta-analysis. Lancet Glob
1266 Heal. 2017;5: e51–e59. doi:10.1016/S2214-109X(16)30306-0
1267 85. José RJ, Brown JS. Adult pneumococcal vaccination: advances, impact, and
1268 unmet needs. Curr Opin Pulm Med. 2017;23.
1269 86. Harrow GL, Lees JA, Hanage WP, Lipsitch M, Corander J, Colijn C, et al.
1270 Negative frequency-dependent selection and asymmetrical transformation
1271 stabilise multi-strain bacterial population structures. ISME J. 2021;15: 1523–
1272 1538. doi:10.1038/s41396-020-00867-w
1273 87. Fedson DS, Anthony J, Scott G. The burden of pneumococcal disease among
1274 adults in developed and developing countries: what is and is not known.
1275 Vaccine. 1999;17: S11–S18. doi:https://doi.org/10.1016/S0264-
1276 410X(99)00122-X
1277 88. Torres A, Blasi F, Dartois N, Akova M. Which individuals are at increased risk
1278 of pneumococcal disease and why? Impact of COPD, asthma, smoking,
1279 diabetes, and/or chronic heart disease on community-acquired pneumonia and
1280 invasive pneumococcal disease. Thorax. 2015;70: 984 LP – 989.
1281 doi:10.1136/thoraxjnl-2015-206780
1282 89. Kyaw MH, Lynfield R, Schaffner W, Craig AS, Hadler J, Reingold A, et al.
1283 Effect of Introduction of the Pneumococcal Conjugate Vaccine on Drug-
1284 Resistant *Streptococcus pneumoniae*. N Engl J Med. 2006;354: 1455–1463.
1285 doi:10.1056/NEJMoa051642
1286 90. Mavroidi A, Aanensen DM, Godoy D, Skovsted IC, Kaltoft MS, Reeves PR, et

1287 al. Genetic relatedness of the *Streptococcus pneumoniae* capsular
1288 biosynthetic loci. J Bacteriol. 2007;189: 7841–7855. doi:10.1128/JB.00836-07
1289 91. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, et al.
1290 Population genomics of post-vaccine changes in pneumococcal epidemiology.
1291 Nat Genet. 2013;45: 656–663. doi:10.1038/ng.2625
1292 92. Manso AS, Chai MH, Atack JM, Furi L, De Ste Croix M, Haigh R, et al. A
1293 random six-phase switch regulates pneumococcal virulence via global
1294 epigenetic changes. Nat Commun. 2014;5: 5055. doi:10.1038/ncomms6055
1295 93. Kwun MJ, Oggioni MR, De Ste Croix M, Bentley SD, Croucher NJ. Excision-
1296 reintegration at a pneumococcal phase-variable restriction-modification locus
1297 drives within- and between-strain epigenetic differentiation and inhibits gene
1298 acquisition. Nucleic Acids Res. 2018;46: 11438–11453.
1299 doi:10.1093/nar/gky906
1300 94. D'Aeth JC, van der Linden MPG, McGee L, de Lencastre H, Turner P, Song J-
1301 H, et al. The role of interspecies recombination in the evolution of antibiotic-
1302 resistant pneumococci. Elife. 2021;10: e67113. Available:
1303 https://doi.org/10.7554/eLife.67113
1304 95. Skwark MJ, Croucher NJ, Puranen S, Chewapreecha C, Pesonen M, Xu YY,
1305 et al. Interacting networks of resistance, virulence and core machinery genes
1306 identified by genome-wide epistasis analysis. PLoS Genet. 2017;13:
1307 e1006508. doi:10.1371/journal.pgen.1006508
1308 96. Lewnard JA, Givon-Lavi N, Tähtinen PA, Dagan R. Pneumococcal Phenotype
1309 and Interaction with Nontypeable *Haemophilus influenzae* as Determinants of
1310 Otitis Media Progression. Infect Immun. 2018;86: e00727-17.
1311 doi:10.1128/IAI.00727-17

1312 97. Epping L, van Tonder AJ, Gladstone RA, Bentley SD, Page AJ, Keane JA.
1313 SeroBA: rapid high-throughput serotyping of *Streptococcus pneumoniae* from
1314 whole genome sequence data. Microb Genomics. 2018;4:
1315 10.1099/mgen.0.000186. doi:10.1099/mgen.0.000186
1316 98. Kapatai G, Sheppard CL, Al-Shahib A, Litt DJ, Underwood AP, Harrison TG, et
1317 al. Whole genome sequencing of *Streptococcus pneumoniae*: development,
1318 evaluation and verification of targets for serogroup and serotype prediction
1319 using an automated pipeline. PeerJ. 2016;4: e2477. doi:10.7717/peerj.2477
1320 99. Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE, Walker H, et al. Penicillin-Binding
1321 Protein Transpeptidase Signatures for Tracking and Predicting β-Lactam
1322 Resistance Levels in *Streptococcus pneumoniae*. MBio. 2016;7: e00756-16.
1323 doi:10.1128/mbio.00756-16
1324 100. Huebner RE, Dagan R, Porath N, Wasas AD, T M, Klugman KP. Lack of utility
1325 of serotyping multiple colonies for detection of simultaneous nasopharyngeal
1326 carriage of different pneumococcal serotypes. Pediatr Infect Dis J. 2000;19:
1327 1017–1020. doi:10.1097/00006454-200010000-00019
1328 101. Turner P, Hinds J, Turner C, Jankhot A, Gould K, Bentley SD, et al. Improved
1329 detection of nasopharyngeal cocolonization by multiple pneumococcal
1330 serotypes by use of latex agglutination or molecular serotyping by microarray.
1331 J Clin Microbiol. 2011;49: 1784–1789. doi:10.1128/JCM.00157-11
1332 102. Nzenze SA, Von Gottberg A, Shiri T, Van Niekerk N, De Gouveia L, Violari A,
1333 et al. Temporal Changes in Pneumococcal Colonization in HIV-infected and
1334 HIV-uninfected Mother-Child Pairs Following Transitioning from 7-valent to 13-
1335 valent Pneumococcal Conjugate Vaccine, Soweto, South Africa. J Infect Dis.
1336 2015. doi:10.1093/infdis/jiv167

1337 103. Madhi SA, Nzenze SA, Nunes MC, Chinyanganya L, Van Niekerk N, Kahn K,
1338 et al. Residual colonization by vaccine serotypes in rural South Africa four
1339 years following initiation of pneumococcal conjugate vaccine immunization.
1340 Expert Rev Vaccines. 2020;19: 383–393.
1341 doi:10.1080/14760584.2020.1750377
1342 104. Nzenze SA, Shiri T, Nunes MC, Klugman KP, Kahn K, Twine R, et al.
1343 Temporal changes in pneumococcal colonization in a rural African community
1344 with high HIV prevalence following routine infant pneumococcal immunization.
1345 Pediatr Infect Dis J. 2013;32: 1270–1278.
1346 doi:10.1097/01.inf.0000435805.25366.64
1347 105. Statistics South Africa. Mid-year population estimates: 2009-2013. 2013.
1348 106. Chang Q, Stevenson AE, Croucher NJ, Lee GM, Pelton SI, Lipsitch M, et al.
1349 Stability of the pneumococcal population structure in Massachusetts as PCV13
1350 was introduced. BMC Infect Dis. 2015;15: 68. doi:10.1186/s12879-015-0797-z
1351 107. Centres for Disease Prevention and Control. Active Bacterial Core surveillance
1352 matrix. 2018. Available: https://www.cdc.gov/abcs/downloads/abcs-
1353 surveillance-matrix.pdf
1354 108. Enright MC, Spratt BG. A multilocus sequence typing scheme for
1355 *Streptococcus pneumoniae*: Identification of clones associated with serious
1356 invasive disease. Microbiology. 1998;144: 3049–3060. doi:10.1099/00221287-
1357 144-11-3049
1358 109. Syrjänen RK, Kilpi TM, Kaijalainen TH, Herva EE, Takala AK. Nasopharyngeal
1359 carriage of *Streptococcus pneumoniae* in Finnish children younger than 2
1360 years old. J Infect Dis. 2001;184: 451–459. doi:10.1086/322048
1361 110. Sleeman KL, Daniels L, Gardiner M, Griffiths D, Deeks JJ, Dagan R, et al.

1362 Acquisition of *Streptococcus pneumoniae* and nonspecific morbidity in infants
1363 and their families: A cohort study. Pediatr Infect Dis J. 2005;24: 121–127.
1364 doi:10.1097/01.inf.0000151030.10159.b1
1365 111. Normark BH, Christensson B, Sandgren A, Noreen B, Sylvan S, Burman LG,
1366 et al. Clonal Analysis of *Streptococcus pneumoniae* Nonsusceptible to
1367 Penicillin at Day-Care Centers with Index Cases, in a Region with Low
1368 Incidence of Resistance: Emergence of an Invasive Type 35B Clone among
1369 Carriers. Microb Drug Resist. 2003;9: 337–344.
1370 doi:10.1089/107662903322762761
1371 112. Galanis I, Lindstrand A, Darenberg J, Browall S, Nannapaneni P, Sjöström K,
1372 et al. Effects of PCV7 and PCV13 on invasive pneumococcal disease and
1373 carriage in Stockholm, Sweden. Eur Respir J. 2016;47: 1208–1218.
1374 doi:10.1183/13993003.01451-2015
1375 113. Mato R, Santos Sanches I, Simas C, Nunes S, Carriço JA, Sousa NG, et al.
1376 Natural history of drug-resistant clones of *Streptococcus pneumoniae*
1377 colonizing healthy children in Portugal. Microb Drug Resist. 2005;11: 309–322.
1378 doi:10.1089/mdr.2005.11.309
1379 114. Gibson C. Population of the 100 Largest Cities and Other Urban Places In The
1380 United States: 1790 to 1990. 1998. Available:
1381 https://www.census.gov/library/working-papers/1998/demo/POP-
1382 twps0027.html
1383 115. Gray BM, Converse GM, Dillon HC. Serotypes of *Streptococcus pneumoniae*
1384 causing disease. J Infect Dis. 1979;140: 979–83. doi:10.1093/infdis/140.6.979
1385 116. Sharma D, Baughman W, Holst A, Thomas S, Jackson D, Da Gloria Carvalho
1386 M, et al. Pneumococcal carriage and invasive disease in children before

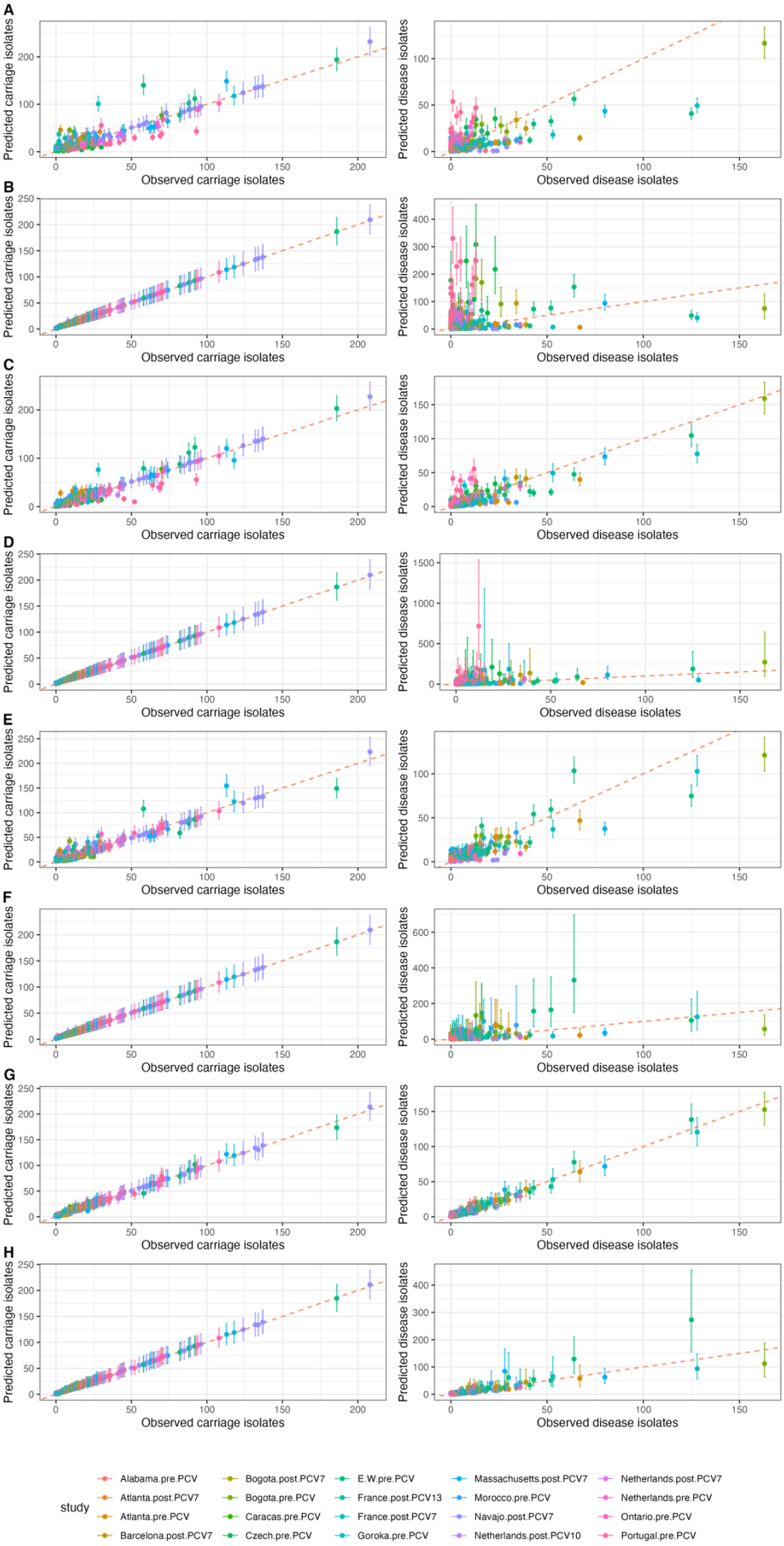
1387 introduction of the 13-valent conjugate vaccine: Comparison with the era
1388 before 7-valent conjugate vaccine. Pediatr Infect Dis J. 2013;32: e45-53.
1389 doi:10.1097/INF.0b013e3182788fdd
1390 117. Parra EL, De La Hoz F, Díaz PL, Sanabria O, Realpe ME, Moreno J. Changes
1391 in *Streptococcus pneumoniae* serotype distribution in invasive disease and
1392 nasopharyngeal carriage after the heptavalent pneumococcal conjugate
1393 vaccine introduction in Bogotá, Colombia. Vaccine. 2013;31: 4033–8.
1394 doi:10.1016/j.vaccine.2013.04.074
1395 118. Instituto Nacional de Estadistica Venezuela. Censo Nacional de Poblacion y
1396 Vivienda. 2011. Available:
1397 http://www.redatam.ine.gob.ve/Censo2011/index.html
1398 119. Rivera-Olivero IA, Del Nogal B, Sisco MC, Bogaert D, Hermans PWM, De
1399 Waard JH. Carriage and invasive isolates of *Streptococcus pneumoniae* in
1400 Caracas, Venezuela: The relative invasiveness of serotypes and vaccine
1401 coverage. Eur J Clin Microbiol Infect Dis. 2011;30: 1489–95.
1402 doi:10.1007/s10096-011-1247-5
1403 120. Czech Statistical Office. Age Distribution of the Population of the Czech
1404 Republic - 2005. 2005. Available: https://www.czso.cz/csu/czso/4003-06-2005-
1405 01n
1406 121. Zemlickova H, Jakubu V, Urbaskova P, Motlova J, Musilek M, Adamkova V.
1407 Serotype-specific invasive disease potential of *Streptococcus pneumoniae* in
1408 Czech children. J Med Microbiol. 2010;59: 1079–83.
1409 doi:10.1099/jmm.0.018390-0
1410 122. Trotter CL, Waight P, Andrews NJ, Slack M, Efstratiou A, George R, et al.
1411 Epidemiology of invasive pneumococcal disease in the pre-conjugate vaccine

1412 era: England and Wales, 1996-2006. J Infect. 2010;60: 200–8.
1413 doi:10.1016/j.jinf.2009.12.008
1414 123. Papua New Guinea National Statistical Office. National Population and
1415 Housing Census Ward Population Profile: Highlands Region. 2011. Available:
1416 https://www.nso.gov.pg/index.php/population-and-social/other-
1417 indicators#highlands-region
1418 124. Plan H-C au. Annuaire statistique de la région Rabat-Salé-Zemmour-Zaër,
1419 2010. 2012. Available: https://www.hcp.ma/downloads/Annuaires-statistiques-
1420 regionaux_t11956.html
1421 125. Jroundi I, Mahraoui C, Benmessaoud R, Moraleda C, Munoz Almagro C,
1422 Seffar M, et al. *Streptococcus pneumoniae* carriage among healthy and sick
1423 pediatric patients before the generalized implementation of the 13-valent
1424 pneumococcal vaccine in Morocco from 2010 to 2011. J Infect Public Health.
1425 2017;10: 165–170. doi:10.1016/j.jiph.2016.02.012
1426 126. United Nations Department of Economic and Social Affairs. Population by age
1427 groups - both sexes. 2020. Available:
1428 https://population.un.org/wpp/Download/Standard/Population/
1429 127. Vissers M, Wijmenga-Monsuur AJ, Knol MJ, Badoux P, van Houten MA, van
1430 der Ende A, et al. Increased carriage of non-vaccine serotypes with low
1431 invasive disease potential four years after switching to the 10-valent
1432 pneumococcal conjugate vaccine in The Netherlands. PLoS One. 2018;13: 1–
1433 15. doi:10.1371/journal.pone.0194823
1434 128. Statistics Canada. Data tables, 1996. 1996. Available:
1435 https://www.statcan.gc.ca/
1436 129. Instituto Nacionalde Estadistica Portugal. Censos 2001. 2002.

1437 130. Statistics Sweden. Population by region, marital status, age and sex. Year
1438 1968 - 2018. Available:
1439 http://www.statistikdatabasen.scb.se/pxweb/en/ssd/START__BE__BE0101__
1440 BE0101A/BefolkningNy/?rxid=e550e75b-4cef-4e77-b5e5-83adcc09c97e
1441 131. Sandgren A, Christensson B, Samuelsson A, Olsson-Liljequist B, Sjöström K,
1442 Kronvall G, et al. Effect of Clonal and Serotype-Specific Properties on the
1443 Invasive Capacity of *Streptococcus pneumoniae*. J Infect Dis. 2004;189: 785–
1444 96. doi:10.1086/381686
1445 132. del Amo E, Brotons P, Monsonis M, Triviño M, Iñigo M, Selva L, et al. High
1446 invasiveness of pneumococcal serotypes included in the new generation of
1447 conjugate vaccines. Clin Microbiol Infect. 2014;20: 684–9. doi:10.1111/1469-
1448 0691.12422
1449 133. Institut National de la Statistique et des Etudes Economiques. Age pyramid
1450 2020 – France and metropolitan France. 2008.
1451 134. Varon E, Cohen R, Béchet S, Doit C, Levy C. Invasive disease potential of
1452 pneumococci before and after the 13-valent pneumococcal conjugate vaccine
1453 implementation in children. Vaccine. 2015;33: 6178–85.
1454 doi:10.1016/j.vaccine.2015.10.015
1455 135. Yildirim I, Hanage WP, Lipsitch M, Shea KM, Stevenson A, Finkelstein J, et al.
1456 Serotype specific invasive capacity and persistent reduction in invasive
1457 pneumococcal disease. Vaccine. 2010/10/30. 2010;29: 283–288. doi:S0264-
1458 410X(10)01502-1 [pii]10.1016/j.vaccine.2010.10.032 [doi]
1459 136. Scott JR, Millar E V., Lipsitch M, Moulton LH, Weatherholtz R, Perilla MJ, et al.
1460 Impact of more than a decade of pneumococcal conjugate vaccine use on
1461 carriage and invasive potential in native American communities. J Infect Dis.

1462 2012. doi:10.1093/infdis/jir730
1463 137. Statistics Finland. Tilastokeskus. 2021.
1464 138. Office for National Statistics. Mid-1998 Population Estimates: Single year of
1465 age and sex for local authorities in England and Wales; estimated resident
1466 population. 2004.
1467 139. Huang SS, Hinrichsen VL, Stevenson AE, Rifas-Shiman SL, Kleinman K,
1468 Pelton SI, et al. Continued Impact of Pneumococcal Conjugate Vaccine on
1469 Carriage in Young Children. Pediatrics. 2009;124: e1-11.
1470 doi:10.1542/peds.2008-3099
1471 140. Wroe PC, Lee GM, Finkelstein JA, Pelton SI, Hanage WP, Lipsitch M, et al.
1472 Pneumococcal carriage and antibiotic resistance in young children before 13-
1473 valent conjugate vaccine. Pediatr Infect Dis J. 2012;31: 249–254.
1474 doi:10.1097/INF.0b013e31824214ac
1475

Figures


A
B
C
D
E
F
G
H
Predicted carriage isolates
Observed carriage isolates
Predicted disease isolates
Observed disease isolates
study
Alabama.pre.PCV
Atlanta.post.PCV7
Atlanta.pre.PCV
Barcelona.post.PCV7
Bogota.post.PCV7
Bogota.pre.PCV
Caracas.pre.PCV
Czech.pre.PCV
E.W.pre.PCV
France.post.PCV13
France.post.PCV7
Goroka.pre.PCV
Massachusetts.post.PCV7
Morocco.pre.PCV
Navajo.post.PCV7
Netherlands.post.PCV10
Netherlands.post.PCV7
Netherlands.pre.PCV
Ontario.pre.PCV
Portugal.pre.PCV

1477

1478 **Figure 1** Comparison of observed and predicted counts of each serotype within each
1479 study of *S. pneumoniae* invasiveness in children. The points are coloured by the
1480 study to which they correspond, and show the observed value on the horizontal axis,
1481 and the mean predicted value on the vertical axis. The error bars show the 95%
1482 credibility intervals. The red dashed line shows the line of identity, corresponding to a
1483 perfect match between prediction and observation. The left column shows the
1484 correspondence for carriage data (values of $c_{i,j}$), and the right column shows the
1485 correspondence for disease isolates (values of $d_{i,j}$). Each row corresponds to a
1486 different model: (A) null Poisson model; (B) null negative binomial model; (C) type-
1487 specific Poisson model; (D) type-specific negative binomial model; (E) study-
1488 adjusted Poisson model; (F) study-adjusted negative binomial model; (G) study-
1489 adjusted type-specific Poisson model; (H) study-adjusted type-specific negative
1490 binomial model.
1491
1492

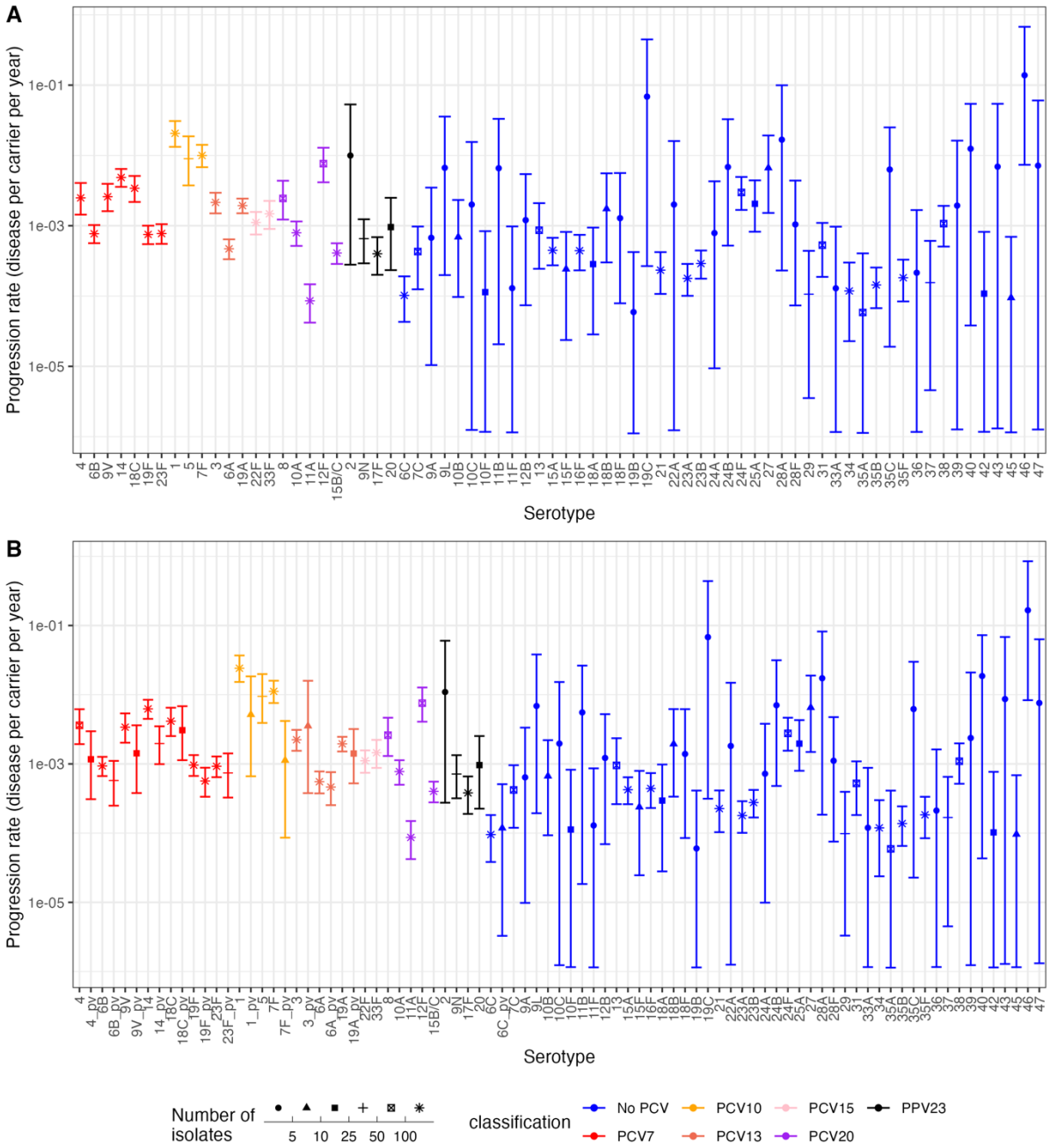
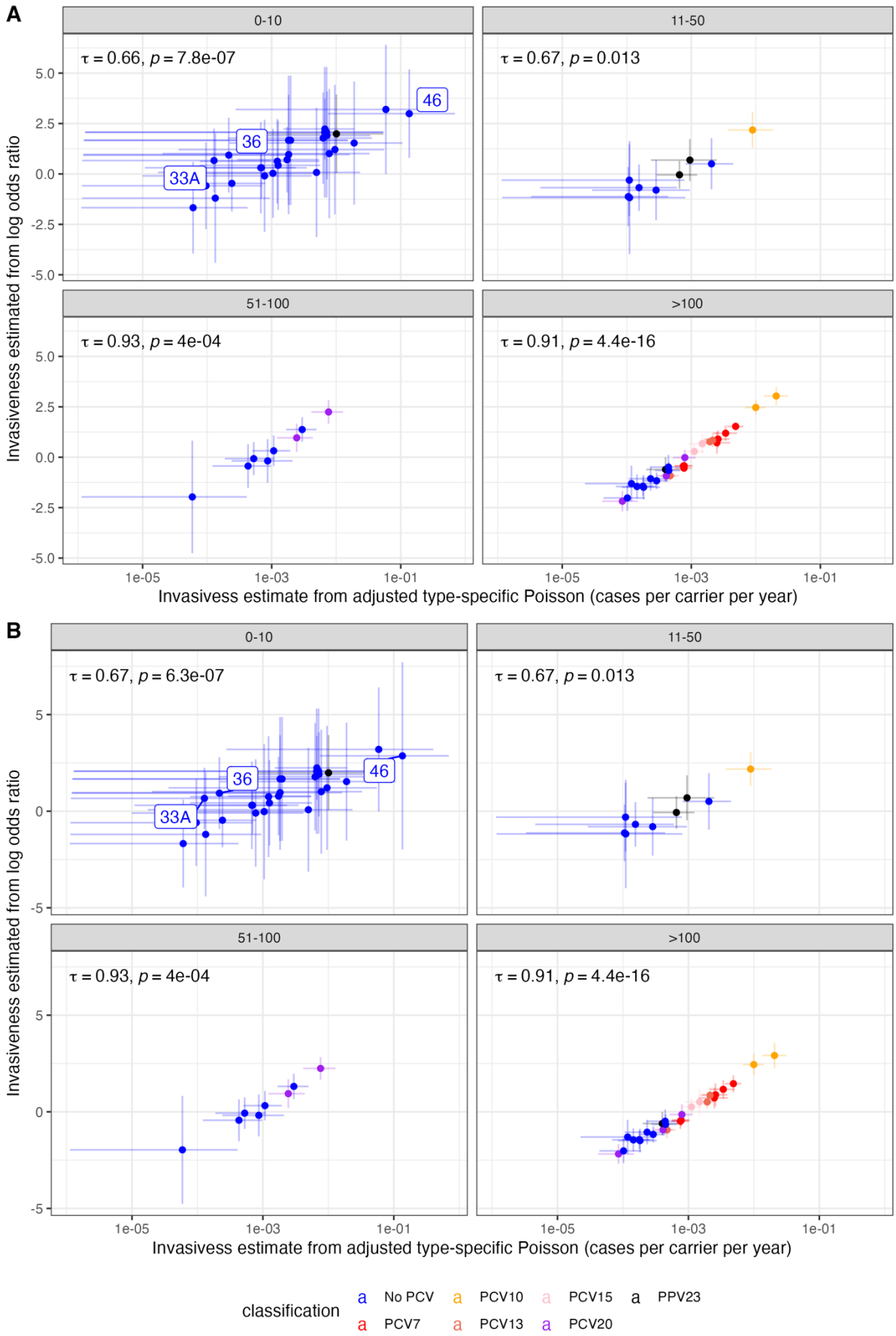


Figure 2 Invasiveness estimates for *S. pneumoniae* serotypes from the study-adjusted type-specific Poisson model applied to the full dataset of serotypes from child carriage and disease. Points represent the mean estimate, and are coloured according to the vaccine formulation in which they are found; all higher valency formulations encompass the serotypes found in lower valency formulations, with the exception of 6A not being present in PPV23. The point shape represents the sample size, summed across disease and carriage isolates from all studies, on which the estimates are based. The error bars show the 95% credibility intervals. (A) Estimates for all 72 serotypes in the full dataset. (B) Estimates when distinguishing between vaccine-type serotypes in unvaccinated and vaccinated populations. Vaccine serotypes in vaccinated populations are denoted with the suffix "_pv".


A
0-10
τ = 0.66, p = 7.8e-07
46
36
33A
11-50
τ = 0.67, p = 0.013
51-100
τ = 0.93, p = 4e-04
>100
τ = 0.91, p = 4.4e-16
Invasiveness estimated from log odds ratio
Invasivess estimate from adjusted type-specific Poisson (cases per carrier per year)
B
0-10
τ = 0.67, p = 6.3e-07
46
36
33A
11-50
τ = 0.67, p = 0.013
51-100
τ = 0.93, p = 4e-04
>100
τ = 0.91, p = 4.4e-16
Invasiveness estimated from log odds ratio
Invasivess estimate from adjusted type-specific Poisson (cases per carrier per year)
classification
No PCV
PCV10
PCV15
PPV23
PCV7
PCV13
PCV20

1505

1506 **Figure 3** Comparison of the serotype invasiveness estimates from the study-
1507 adjusted type-specific Poisson model with those from meta-analyses of the same
1508 dataset calculated as logarithmic odds ratios combined using (A) fixed effects
1509 models and (B) random effects models. Points are coloured based on the vaccine
1510 formulations in which serotypes are found. The horizontal error bars are the 95%
1511 credibility intervals from the Bayesian model, and the vertical error bars show the
1512 95% confidence intervals from the logarithmic odds ratio analyses. Each plot is
1513 divided based on the sample size associated with each serotype, calculated as the
1514 sum of isolates from carriage and disease across all studies. The correlation
1515 between the two estimates is summarised as Kendall's τ on each panel, with an
1516 associated *p* value.

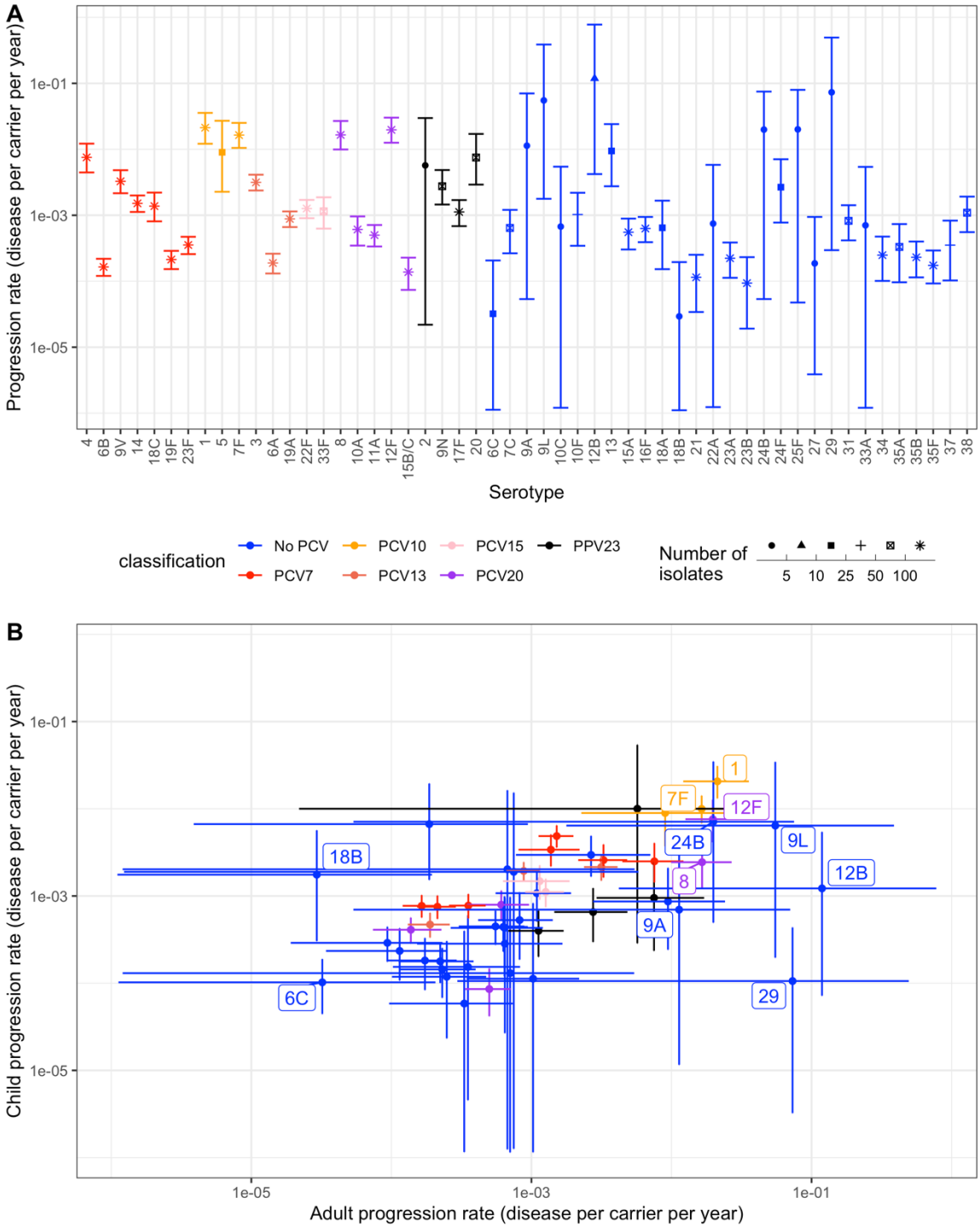
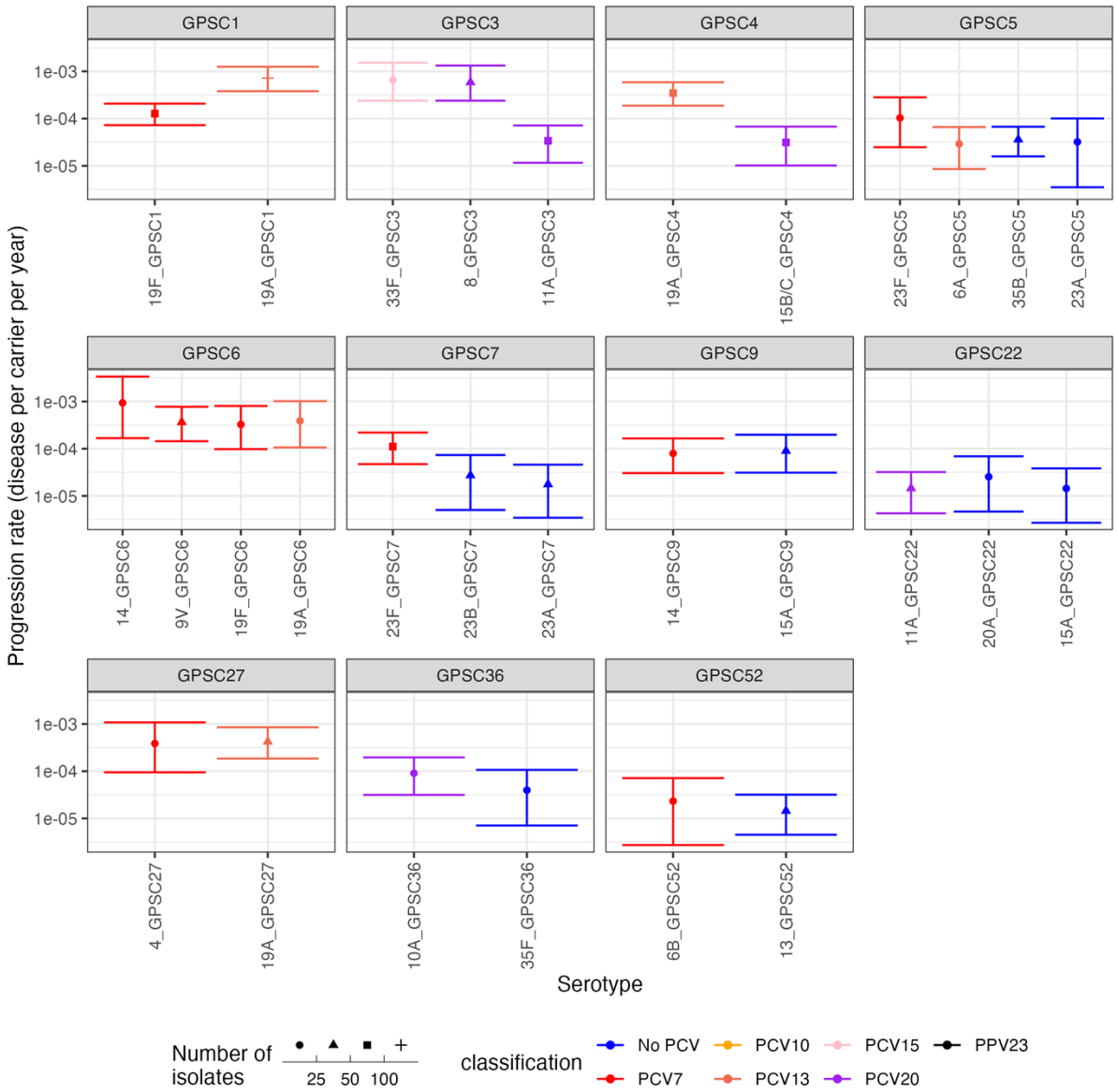


Figure 4 Invasiveness estimates for *S. pneumoniae* serotypes from the study-adjusted type-specific Poisson model applied to the full dataset of serotypes from child carriage and adult disease. Points represent the mean estimate, and are coloured according to the vaccine formulation in which they are found; all higher valency formulations encompass the serotypes found in lower valency formulations, with the exception of 6A not being present in PPV23. The error bars show the 95% credibility intervals. (A) Estimates for all 53 serotypes in the full dataset. The point shape represents the sample size, summed across disease and carriage isolates from all studies, on which the estimates are based. (B) Scatterplot comparing the invasiveness estimates for adults and children.

1529



1530
1531 **Figure 5** Estimates of within- and between-strain variation in invasiveness for strain
1532 and serotype combinations represented by at least ten isolates across the studies
1533 with genotype information. Points represent the mean estimate, and are coloured by
1534 the vaccine formulations in which the corresponding serotype is present. The shape
1535 of the point represents the sample size on which the estimate is based. Estimates
1536 are grouped by strain. Only strains for which estimates of multiple strain and
1537 serotype combinations were possible are displayed.

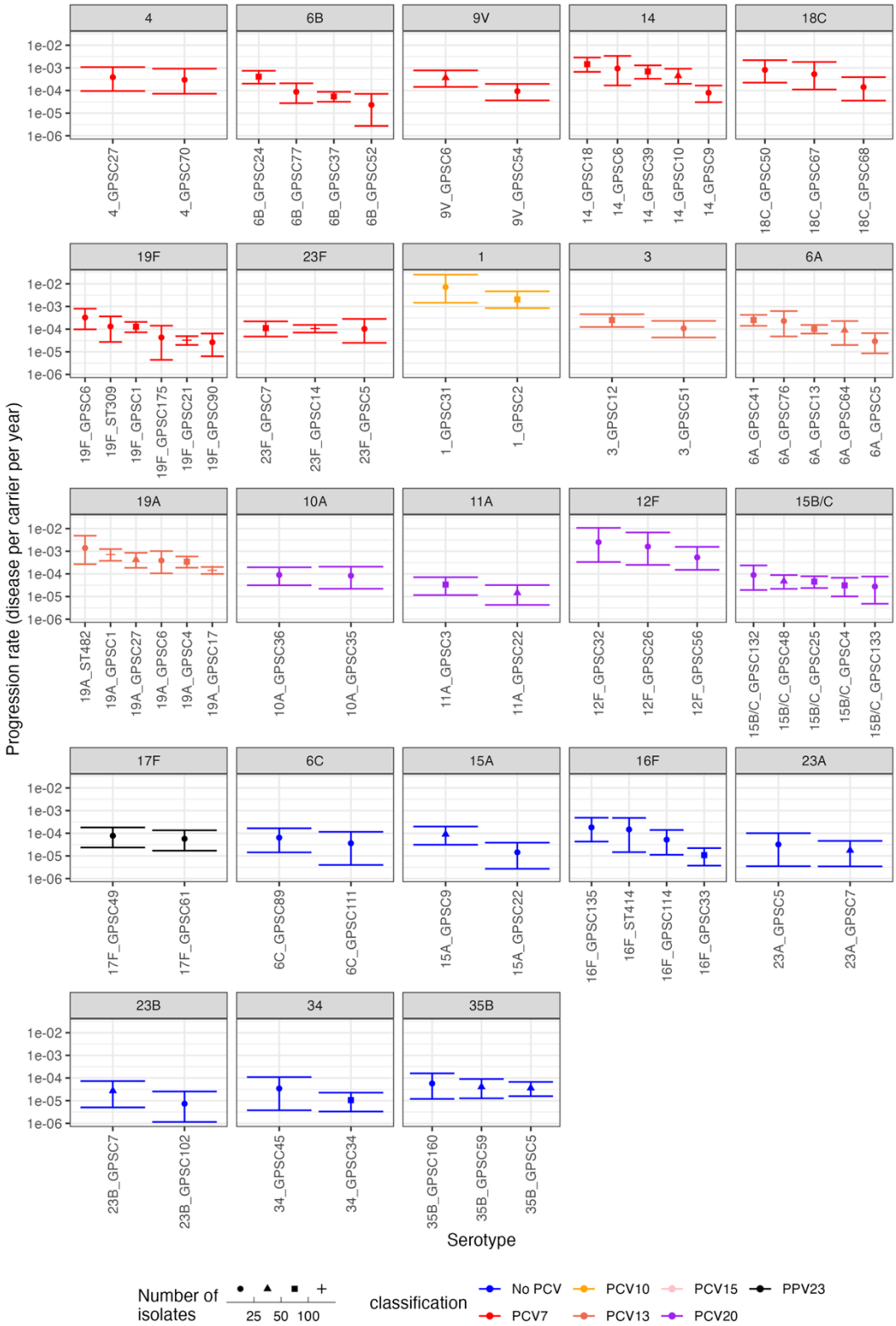


Figure 6 Estimates of within- and between-serotype variation in invasiveness for strain and serotype combinations represented by at least ten isolates across the included studies. Points represent the mean estimate, and are coloured by the vaccine formulations in which the corresponding serotype is present. The shape of the point represents the sample size on which the estimate is based. Estimates are grouped by serotype. Only serotypes for which estimates of multiple strain and serotype combinations were possible are displayed.

1546 **Tables**

Parameter	Description	Value or prior distribution
i	Index denoting study	1..(number of studies)
j	Index denoting type	1..(number of types)
k	Index denoting strain	1..(number of strains)
x	Generic index for type or strain	-
y	Generic index of type or strain that modifies the invasiveness of a genotype	-
c	Number of isolates detected in carriage study	-
d	Number of isolates detected in disease surveillance	-
ρ	Carriage prevalence	ρ ~Beta(1,1)
η	Total number of positive and negative samples in carriage study	From previous publications
N	Number of individuals of the specified demographic under disease surveillance	From previous publications
t	Duration of disease surveillance	From previous publications
v	Progression rate	$\log_{10}(v)$ ~U(-6,1) or truncated Cauchy(μ_{mod}, τ_{mod}); see below
γ	Study-specific scale parameters	1 or $\log_{10}(\gamma)$~ Cauchy(μ_{study},τ_{study})
ϕ	Precision parameter of negative binomial distribution	$\log_{10}(\phi)$~U(-3,3)
Γ_i	Random variable for sampling from the Cauchy distribution for the study-specific scale parameters	Γ_i ~U(-0.5π,0.5π)
μ_{study}	Location parameter of the Cauchy distribution for the study-specific scale parameters	0
τ_{study}	Scale parameter of the Cauchy distribution for the study-specific scale parameters	2
Y_y	Random variable for sampling from the truncated Cauchy distribution for the invasiveness modification terms	Y_y~U(-1.25,1.25)
μ_{mod}	Location parameter of the truncated Cauchy distribution for the invasiveness modification terms	0
τ_{mod}	Scale parameter of the truncated Cauchy distribution for the invasiveness modification terms	0.5

1547

1548 **Table 1** Summary of parameters used in the models.

1549 **Supporting Information**
1550
1551

1552 **Analysing pneumococcal invasiveness using Bayesian models of pathogen**

1553 **progression rates**

1554

1555 Alessandra Løchen, James E. Truscott and Nicholas J. Croucher

1556

1557 **Text S1**

1558 **Figures S1-S31**

1559 **Tables S1-11**

1560 **Datasets S1-S3** (legends only)

1561 **Text S1: Sample descriptions for genomic and genotyped datasets**

1562 The *S. pneumoniae* population was previously divided into strains by the Global
1563 Pneumococcal Sequencing project [31] using PopPUNK [5]. Previous estimation of
1564 the effect of strain background on invasiveness by Gladstone *et al* [35] used two
1565 study populations. The first combined two carriage studies from Agincourt and
1566 Soweto, in South Africa, with nationwide surveillance of IPD by the National Institute
1567 for Communicable Diseases. After excluding carriage and disease isolates from
1568 individuals reported as being HIV positive, genomes were available for 1,089
1569 carriage isolates and 855 IPD isolates from children under seven years old. Model
1570 fitting required calculating the number of swabs taken from HIV negative children
1571 under seven in the carriage studies. In Soweto, the reported age stratification meant
1572 this number could only be calculated for children under five [102], and therefore four
1573 carriage isolates from children aged five or six were excluded from the study. In
1574 Agincourt, the reported age stratification meant 33 isolates from children aged six
1575 were excluded [103]. As the prevalence of HIV was low in infants in the Agincourt
1576 study [104], and the exact number of swabs from HIV positive children was not
1577 reported, no adjustment was made to the sampling based on HIV status. The
1578 population of children under surveillance for IPD was estimated from nationwide
1579 census data between 2009-2013 [105]. All samples were stratified into the post-
1580 PCV7 (2009-2011) and post-PCV13 (2012-2013) periods [31].

1581

1582 The second study population combined a carriage survey from Massachusetts, USA
1583 and isolates from the Active Bacterial Core Surveillance (ABCS) of IPD across
1584 several states. Isolates were collected over the sampling periods of 2000-2001,
1585 2003-2004, 2006-2007 and 2008-2009. As only 51 IPD samples were available for

1586 the earliest two periods, and the population was changing rapidly over these years
1587 [91], only samples from the more stable post-PCV7 2006-2009 period were studied
1588 [106]. This included 405 IPD isolates, modelled as coming from a time interval of two
1589 years (as samples only came from 2007 and 2009) from a population size calculated
1590 from the ABCS population matrix [107]. Of the carriage isolates, 280 were assigned
1591 to strains from genomic data, while 291 were only associated with multi-locus
1592 sequence typing (MLST) genotypes [108]. Gladstone *et al* assigned these isolates to
1593 strains using the close relationship between MLST and PopPUNK population
1594 classifications [5,35].
1595
1596 As these studies were estimating odds ratios across samples stratified by year,
1597 Gladstone *et al* randomly subsampled isolates to ensure an even ratio of disease
1598 and carriage at each sampling timepoint [35]. This correction is unnecessary with the
1599 models described in this study, as the γ_i scaling parameters can adjust for differing
1600 levels of surveillance between samples (incorporating variation in both the samples
1601 collected, and the subset sequenced). Furthermore, such subsampling would distort
1602 the inferred $\rho_{i,x}$.
1603
1604 MLST genotypes were also available for isolates from the studies of serotype
1605 invasiveness from Stockholm [75], Finland [71,109] and Oxford, UK [29]. Where
1606 possible, these were assigned to GPS strains using the isolate's MLST data, or the
1607 clonal complex to which they belonged at the time of publication [4]. If an isolate
1608 could not be assigned to a GPS strain, then its MLST sequence type was used to
1609 designate its genotype instead. Isolates that could not be unequivocally assigned a
1610 strain, serotype and source (carriage or disease) were omitted.

1611

1612 For the Oxford study, these criteria meant 84 of the 150 disease isolates, and 113 of
1613 the 351 carriage isolates, could be analysed. The carriage isolates were assembled
1614 from three different surveys, only one of which could be linked to a more detailed
1615 published description of the total carriage sample size [110]. As the isolates in this
1616 study were recovered from 213 individuals, and the Oxford study attempted to
1617 minimise multiple sampling from any one host [29], the overall carriage sample size
1618 for the three studies was approximated as three times the number of individuals in
1619 this published study (639).

1620

1621 For the Finland study, the criteria described above meant 143 of the 224 disease
1622 isolates, and 102 of the 217 carriage isolates, could be analysed. The carriage
1623 isolates were recovered from ten longitudinal samples from each of 329 individuals
1624 [33]. As the Finland study sought to avoid multiple sampling of the same carriage
1625 episode [71], the host population size was used as the carriage sample size.

1626

1627 For the Stockholm study, the criteria described above meant 65 of the 165 disease
1628 isolates, and 178 of the 550 carriage isolates, could be analysed. The carriage
1629 isolates were recovered from studies conducted between 1997 and 2004. Two
1630 published descriptions of the original carriage studies reported the recovery of 506
1631 isolates from 1,330 individuals [111,112]. As this accounted for over 90% of the
1632 nasopharyngeal isolates, this number was used as the carriage sample size.

1633

1634 An additional study from Portugal used a collection of disease isolates primarily
1635 recovered from adults [52]. Applying the same criteria as to the child disease studies

allowed unambiguous assignation of genotype and serotype to 463 of the 769 carriage isolates, and 152 of the 475 of the disease isolates. The carriage isolates were a subset of a previous study [113], and originated from 1,170 nasopharyngeal swabs (R. Sá-Leão, personal communication).

Supplementary Figures

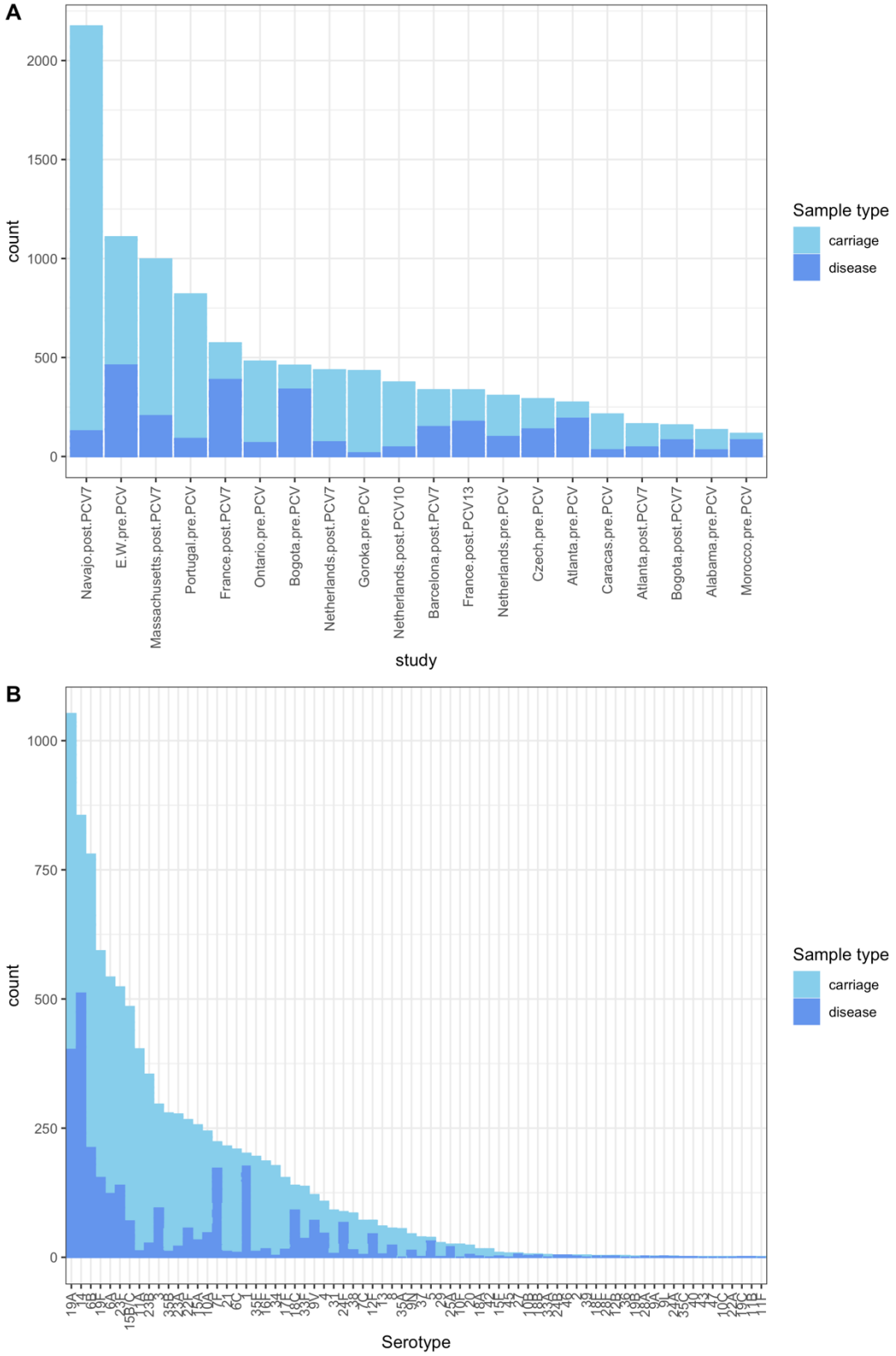


Figure S1 Properties of the child serotype dataset. (A) Stacked bar plot showing the distribution of carriage and disease isolates between studies. (B) Stacked bar plot showing the distribution of carriage and disease isolates between serotypes.

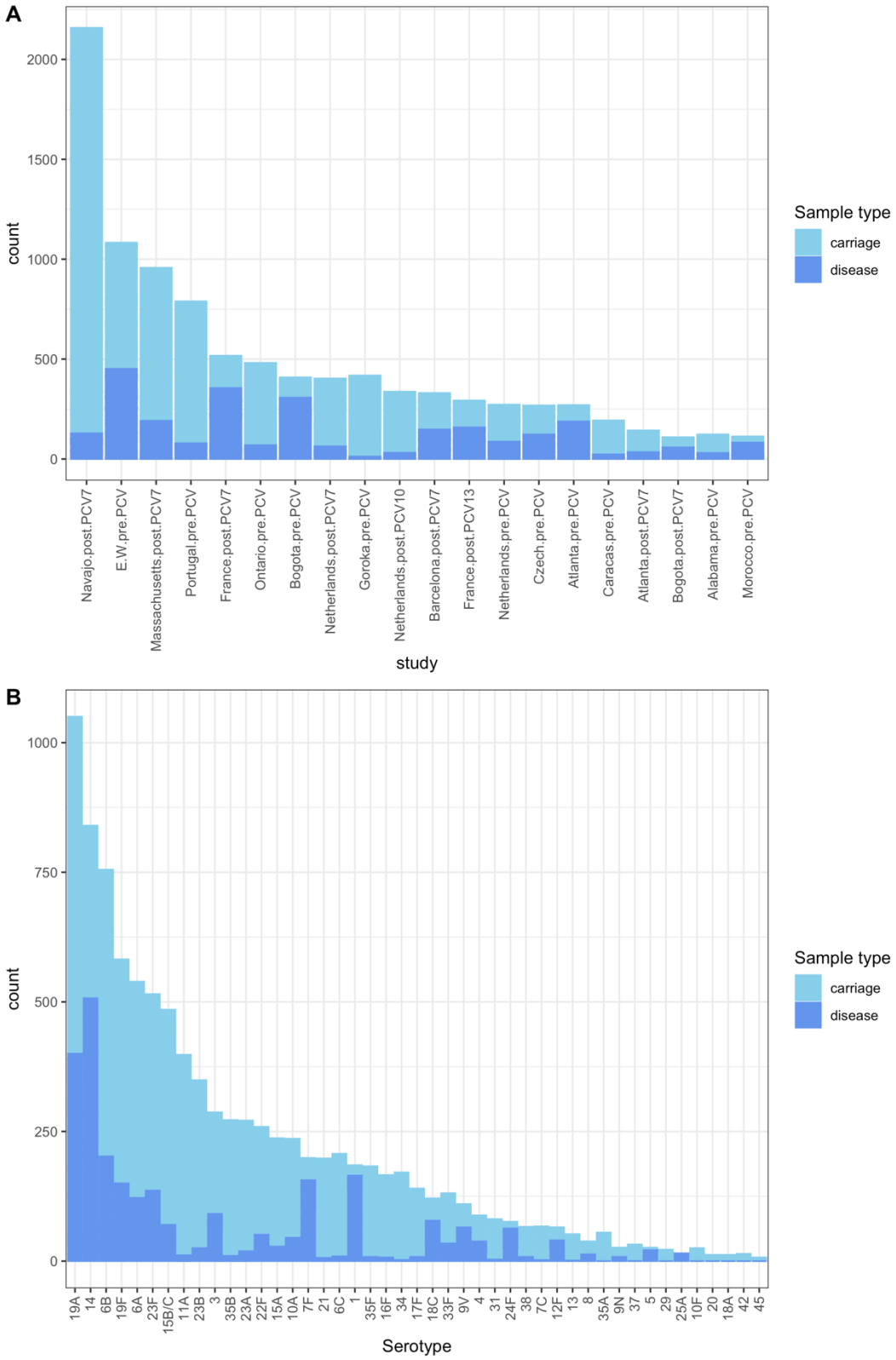


Figure S2 Properties of the child serotype dataset after removing datapoints in which fewer than five isolates of a serotype were recorded from both carriage and disease in a given study. (A) Stacked bar plot showing the distribution of carriage and disease isolates between studies. (B) Stacked bar plot showing the distribution of carriage and disease isolates between serotypes.

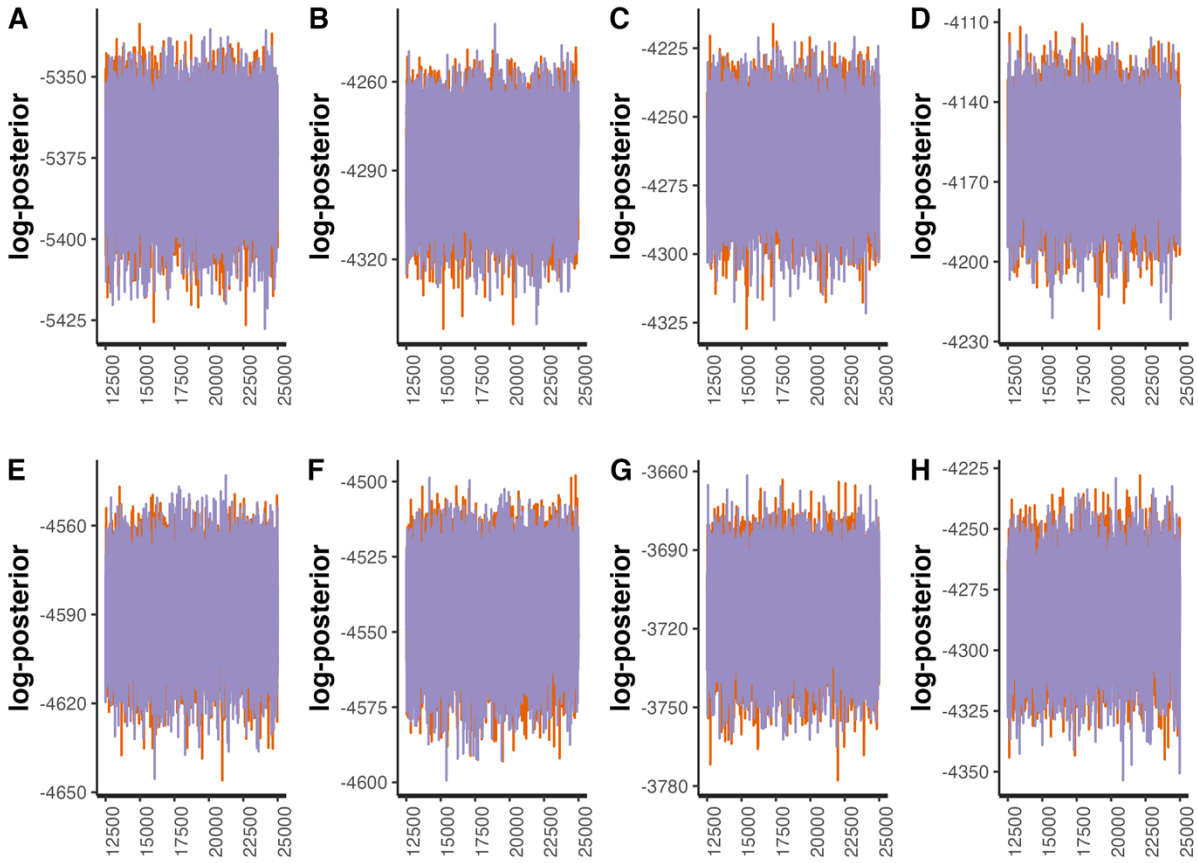


Figure S3 Line plots showing the post-warmup MCMC traces for the logarithmic posterior probabilities across two independent chains for models fitted to the filtered serotype data from child carriage and disease. The horizontal axis shows the generation of the MCMC, with values for the two chains shown by orange and purple lines. Each panel corresponds to a different model: (A) null Poisson model; (B) null negative binomial model; (C) type-specific Poisson model; (D) type-specific negative binomial model; (E) study-adjusted Poisson model; (F) study-adjusted negative binomial model; (G) study-adjusted type-specific Poisson model; (H) study-adjusted type-specific negative binomial model.

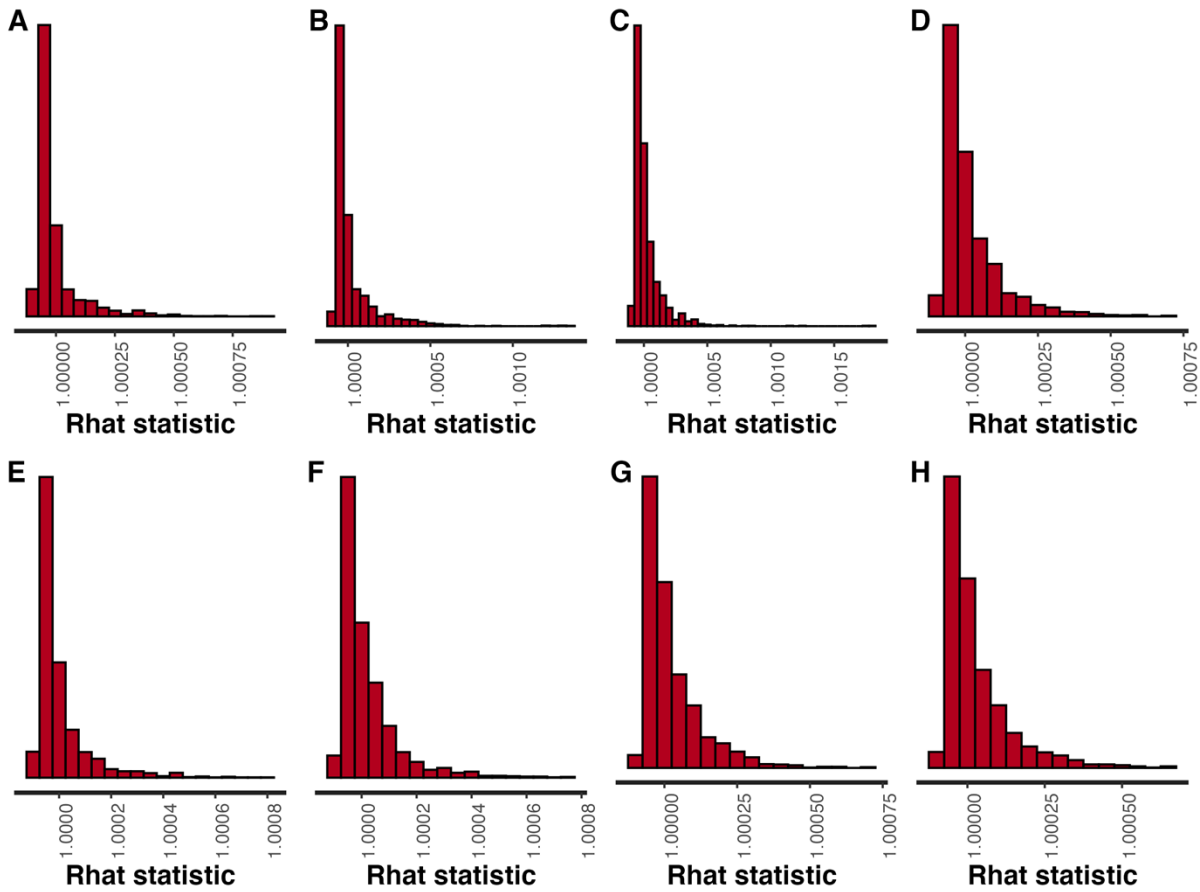


Figure S4 Histograms of Rhat values calculated from the paired MCMCs for models fitted to the filtered serotype data from child carriage and disease. Each panel corresponds to a different model: (A) null Poisson model; (B) null negative binomial model; (C) type-specific Poisson model; (D) type-specific negative binomial model; (E) study-adjusted Poisson model; (F) study-adjusted negative binomial model; (G) study-adjusted type-specific Poisson model; (H) study-adjusted type-specific negative binomial model.

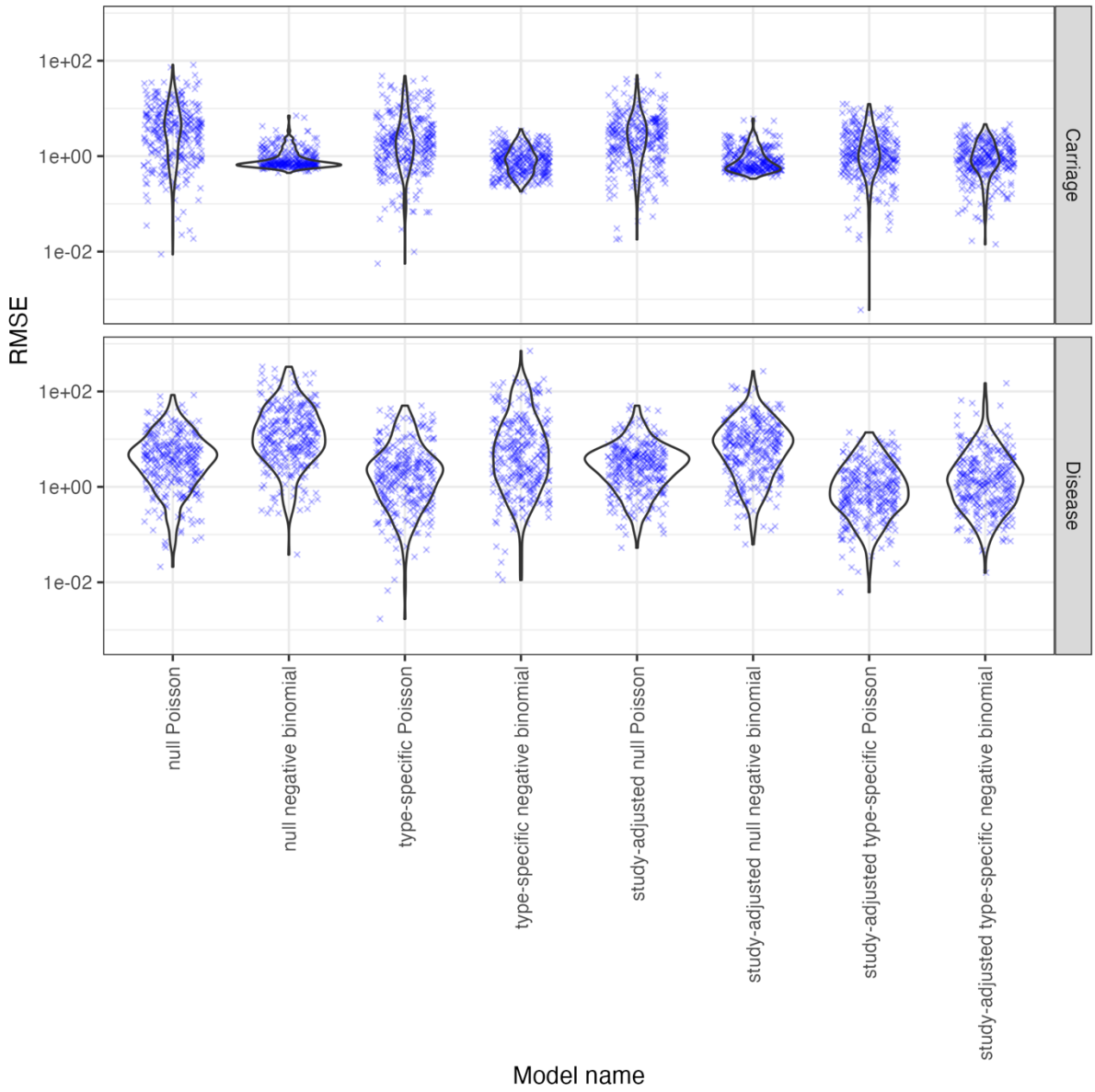


Figure S5 Violin plots showing the root mean square error between observed and predicted values across model fits to the filtered serotype data from child carriage and disease. Blue crosses represent the individual observations.

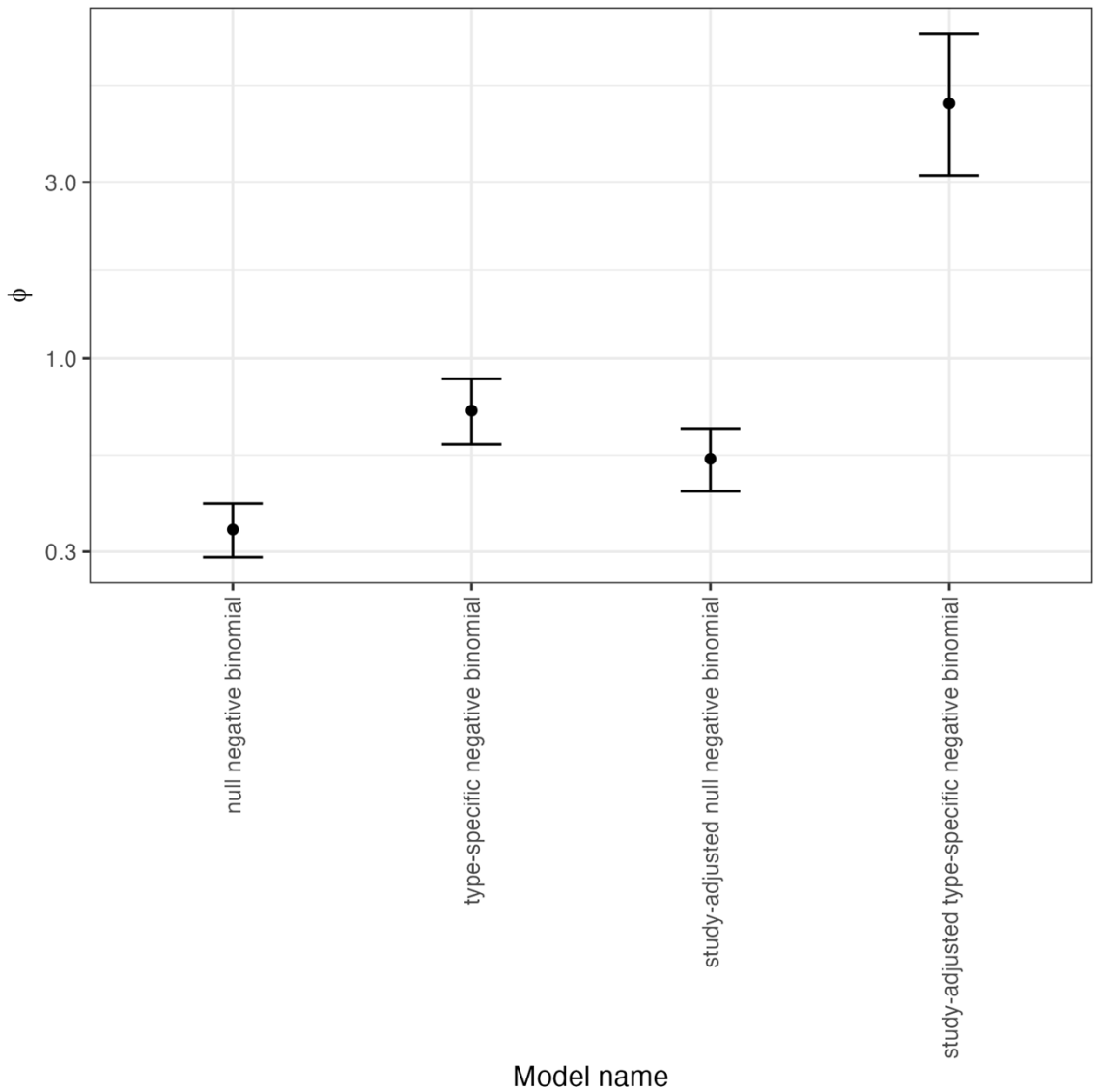


Figure S6 Graph showing the values of the negative binomial distribution's precision parameter, ϕ, from model fits to the filtered serotype data from child carriage and disease. The points represent the mean estimate from the MCMCs, and the error bars show the 95% credibility interval.

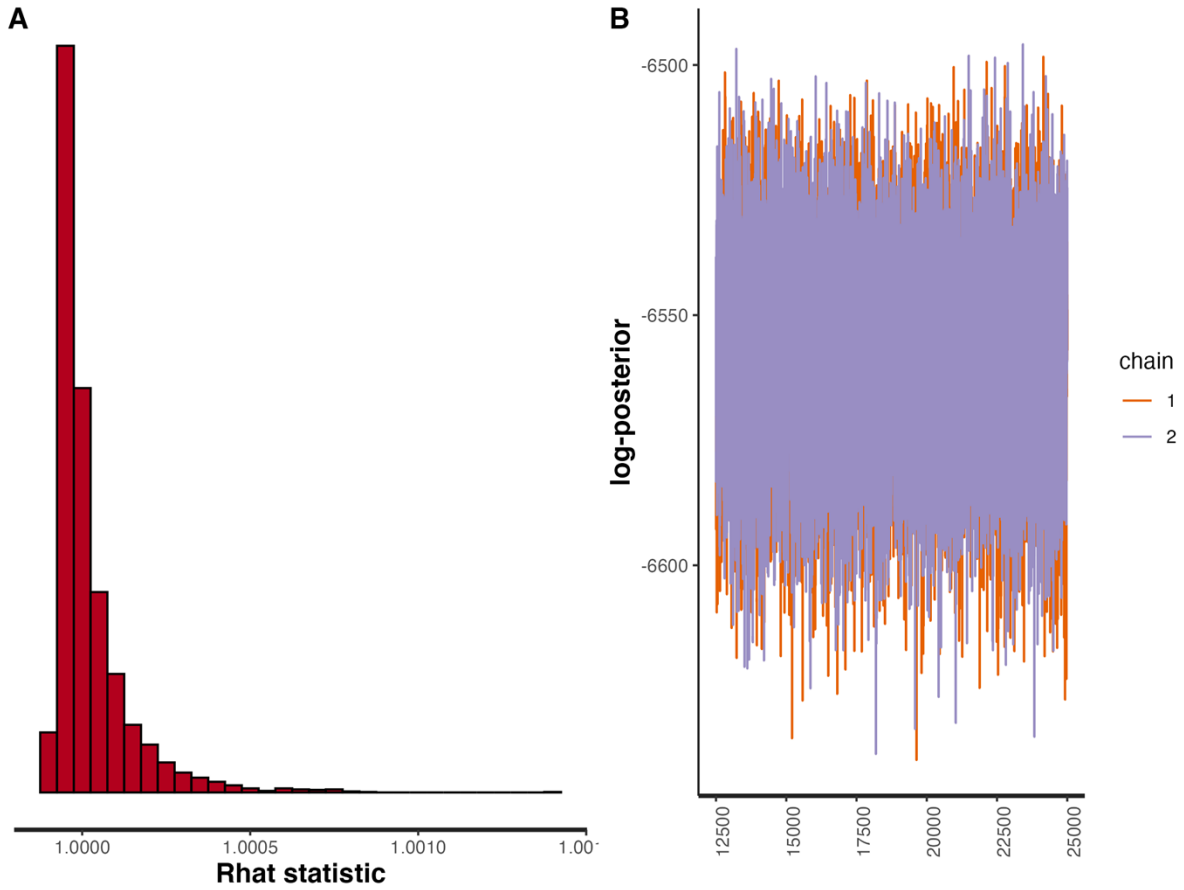


Figure S7 Plots validating the fit of the study-adjusted type-specific Poisson model to the full serotype data from child carriage and disease. (A) Histogram showing the distribution of Rhat values. (B) Post-warmup MCMC traces of the log posterior probability.

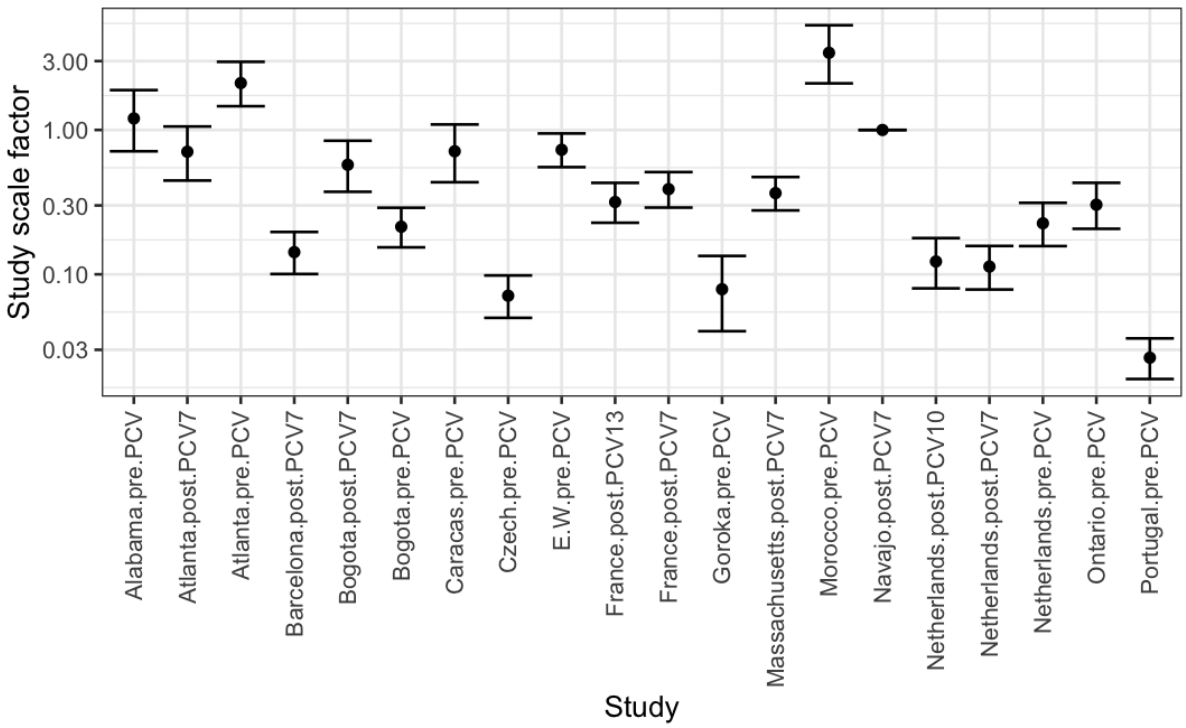


Figure S8 Study-adjustment scale factors from the study-adjusted type-specific Poisson model fitted to the full serotype data from child carriage and disease. The reference study, for which the value was fixed at one, was the Navajo post-PCV7 dataset, which had the greatest sample size in this meta-analysis (Fig. S1).

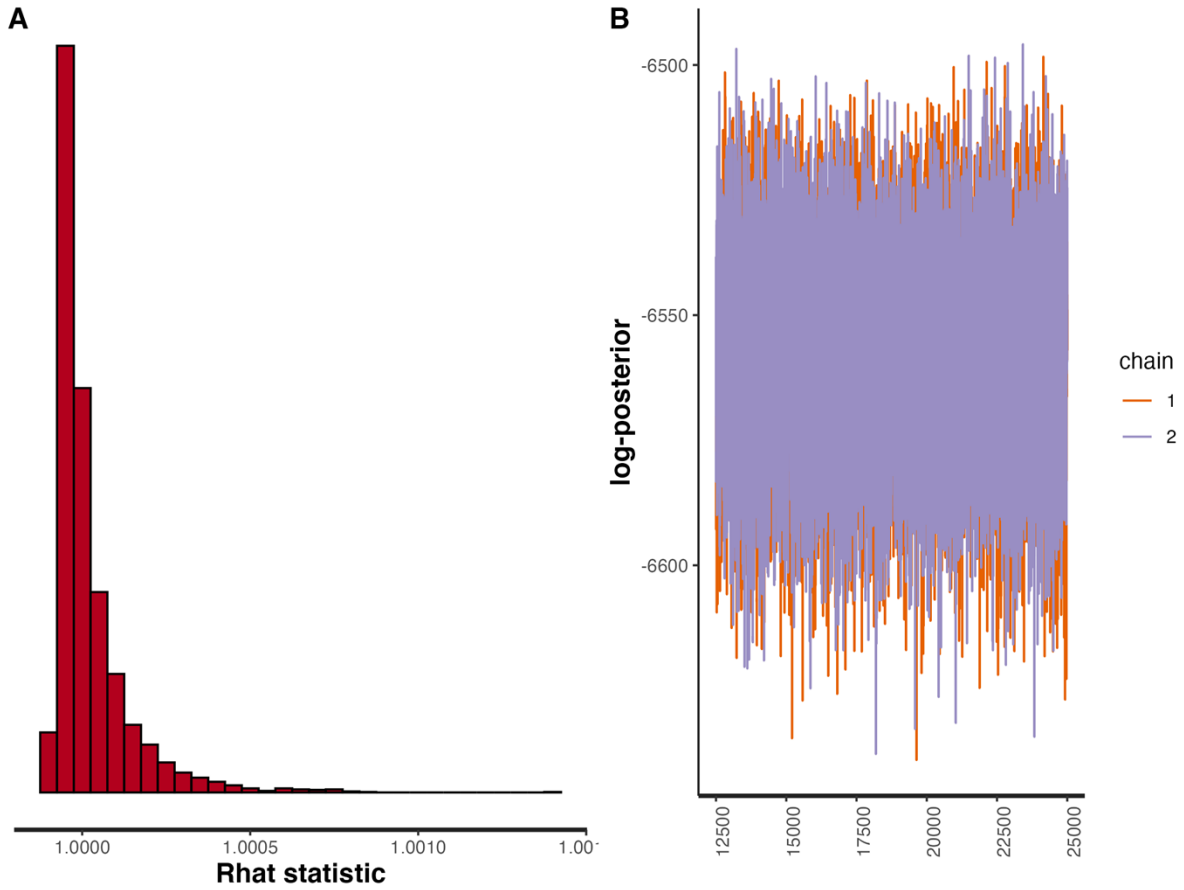


Figure S9 Plots validating the fit of the study-adjusted type-specific Poisson model to the full serotype data from child carriage and disease, when distinguishing between vaccine-type serotypes pre- and post-PCV. (A) Histogram showing the distribution of Rhat values. (B) Post-warmup MCMC traces of the log posterior probability.

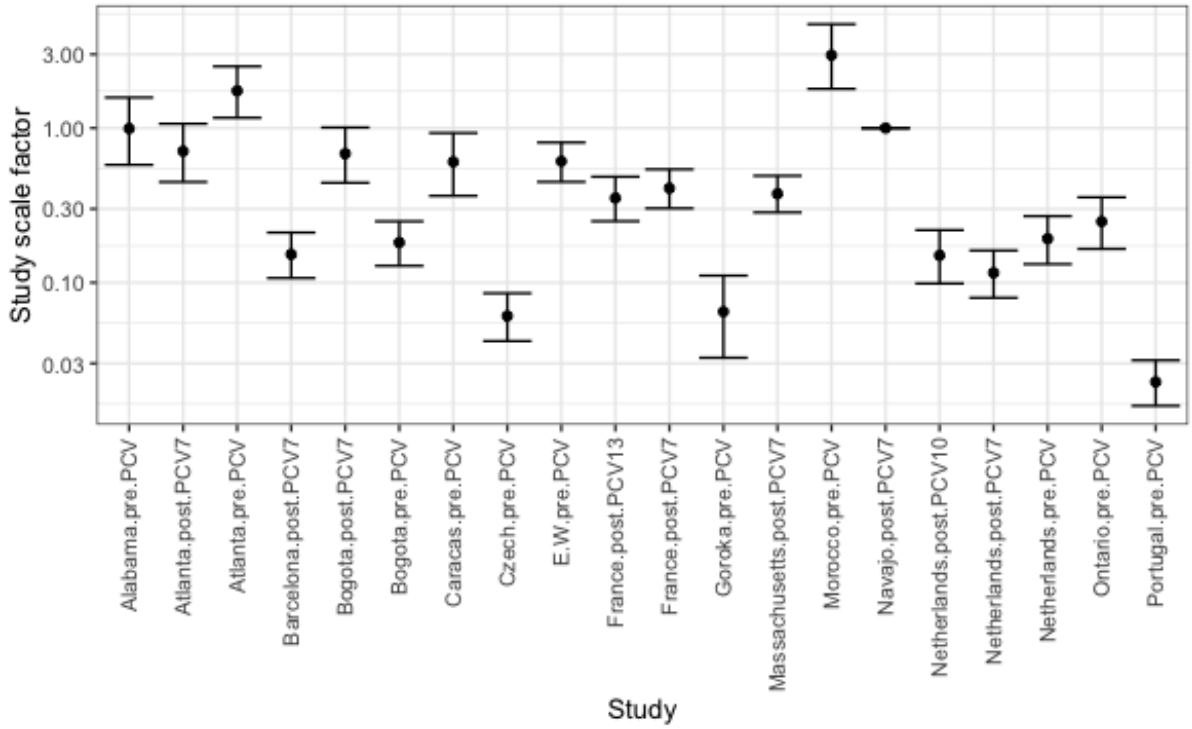


Figure S10 Study-adjustment scale factors from the study-adjusted type-specific Poisson model fitted to the full serotype data from child carriage and disease, when distinguishing between vaccine-type serotypes pre- and post-PCV. The reference study, for which the value was fixed at one, was the Navajo post-PCV7 dataset, which had the greatest sample size in this meta-analysis (Fig. S1).

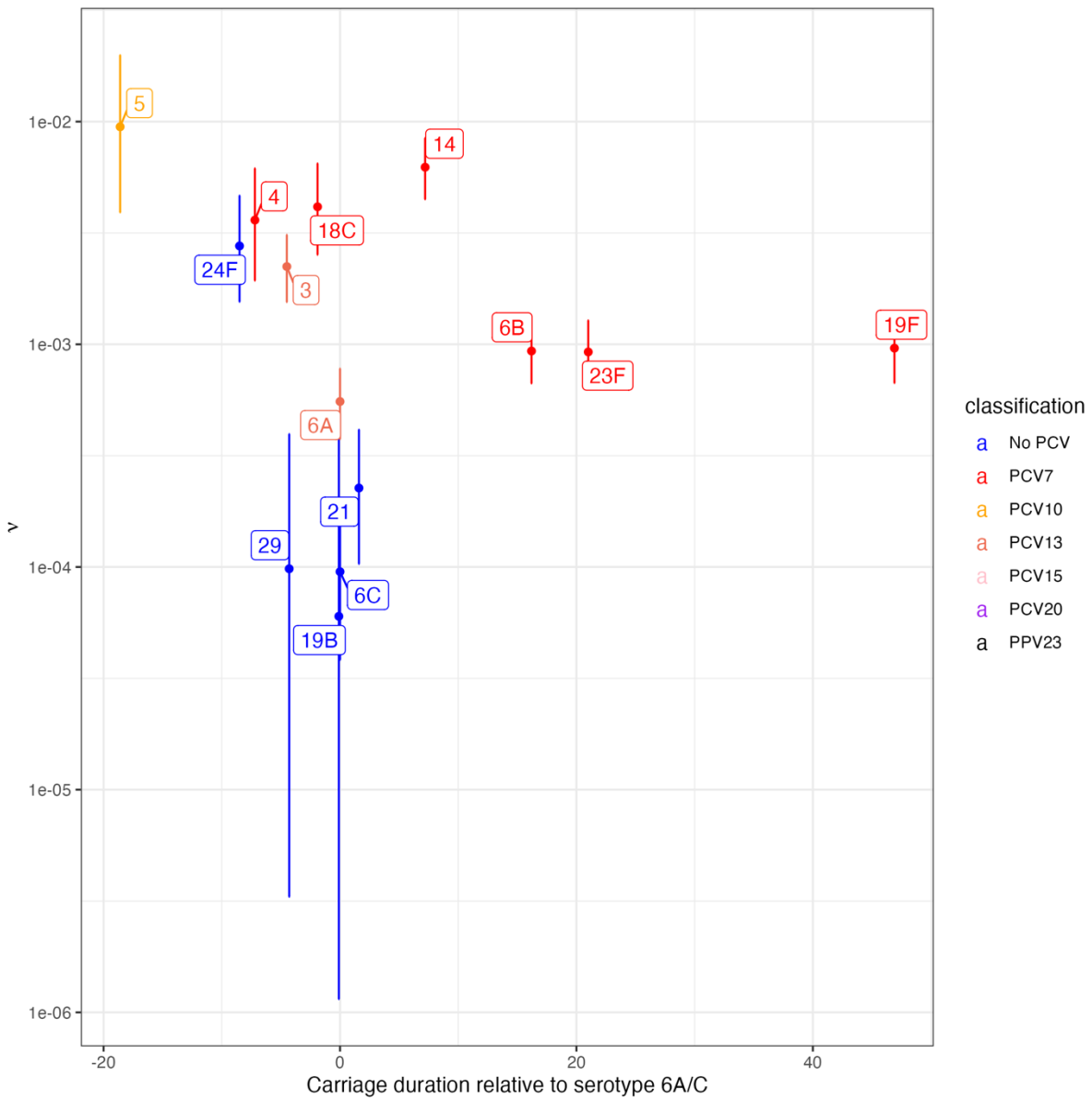


Figure S11 Relationship between serotype invasiveness in children (as shown in Fig. 2) and carriage duration in a study of infant colonisation in Maela, Thailand. The carriage duration estimates were derived from a multi-variate lasso regression that included both serotype and antibiotic resistance phenotypes. Values were available for 14 serotypes, all relative to the carriage duration of serotype 6A/C, which was therefore assigned a value of zero days.

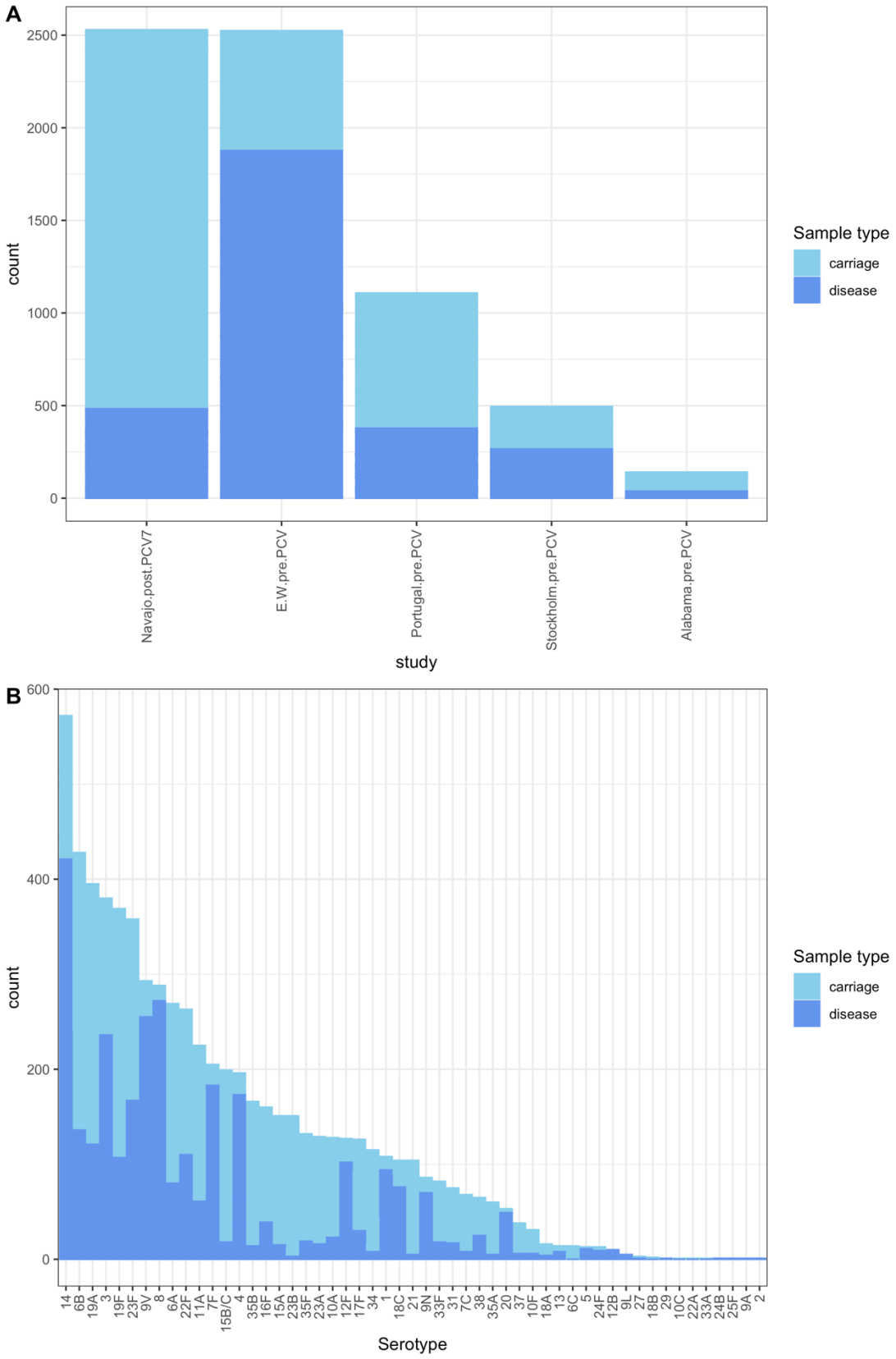


Figure S12 Properties of the adult serotype dataset. (A) Stacked bar plot showing the distribution of carriage and disease isolates between studies. (B) Stacked bar plot showing the distribution of carriage and disease isolates between serotypes.

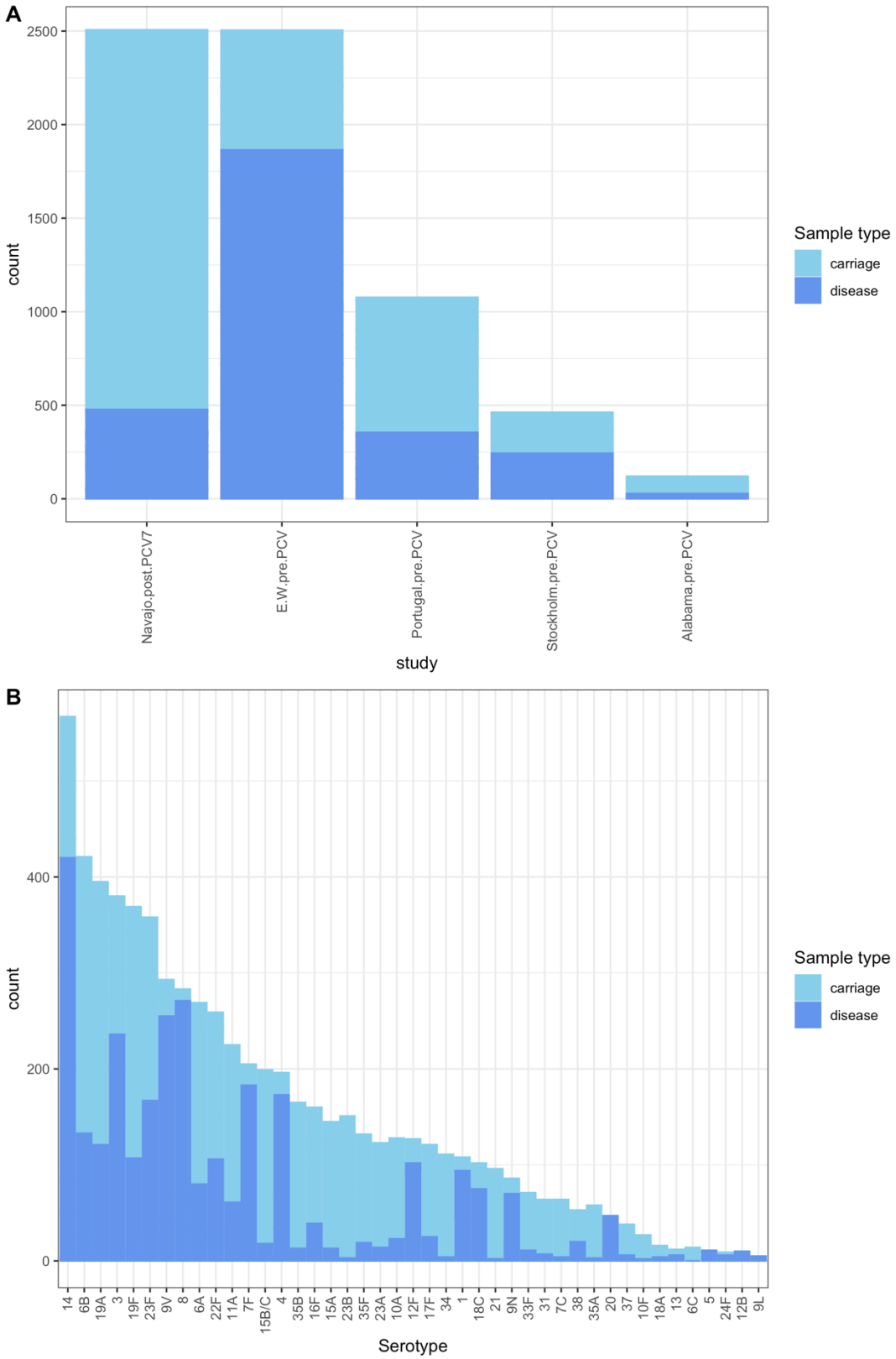


Figure S13 Properties of the adult serotype dataset after removing datapoints in which fewer than five isolates of a serotype were recorded from both carriage and disease in a given study. (A) Stacked bar plot showing the distribution of carriage and disease isolates between studies. (B) Stacked bar plot showing the distribution of carriage and disease isolates between serotypes.

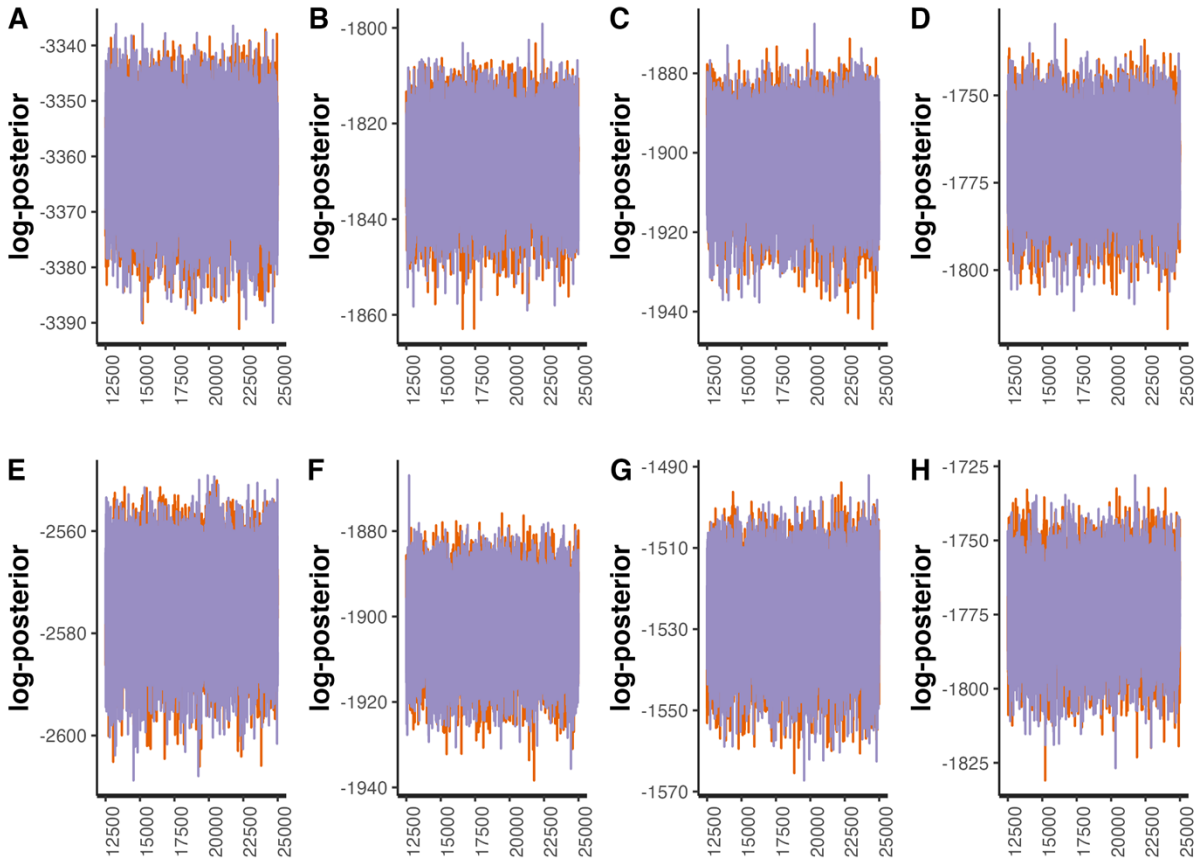


Figure S14 Line plots showing the post-warmup MCMC traces for the logarithmic posterior probabilities across two independent chains for models fitted to the filtered serotype data from child carriage and adult disease. The horizontal axis shows the generation of the MCMC, with values for the two chains shown by orange and purple lines. Each panel corresponds to a different model: (A) null Poisson model; (B) null negative binomial model; (C) type-specific Poisson model; (D) type-specific negative binomial model; (E) study-adjusted Poisson model; (F) study-adjusted negative binomial model; (G) study-adjusted type-specific Poisson model; (H) study-adjusted type-specific negative binomial model.

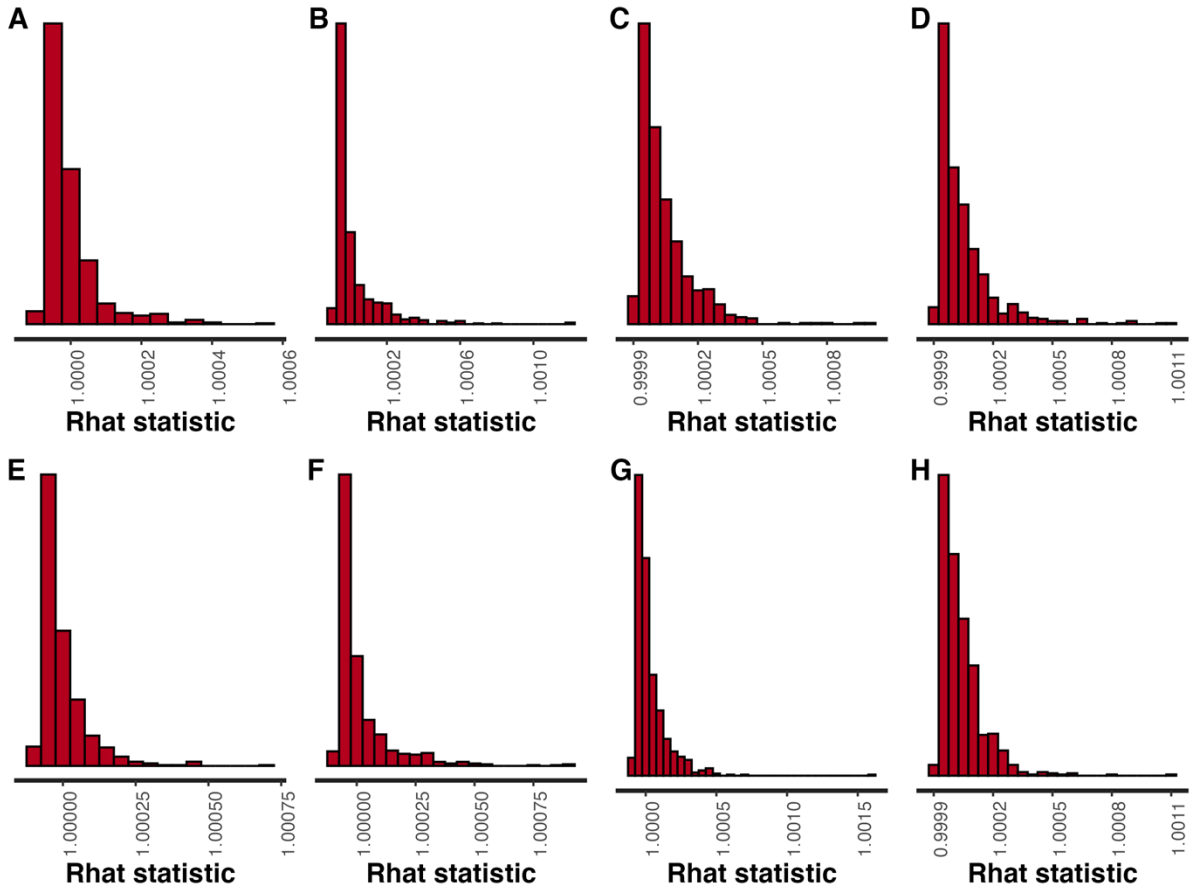
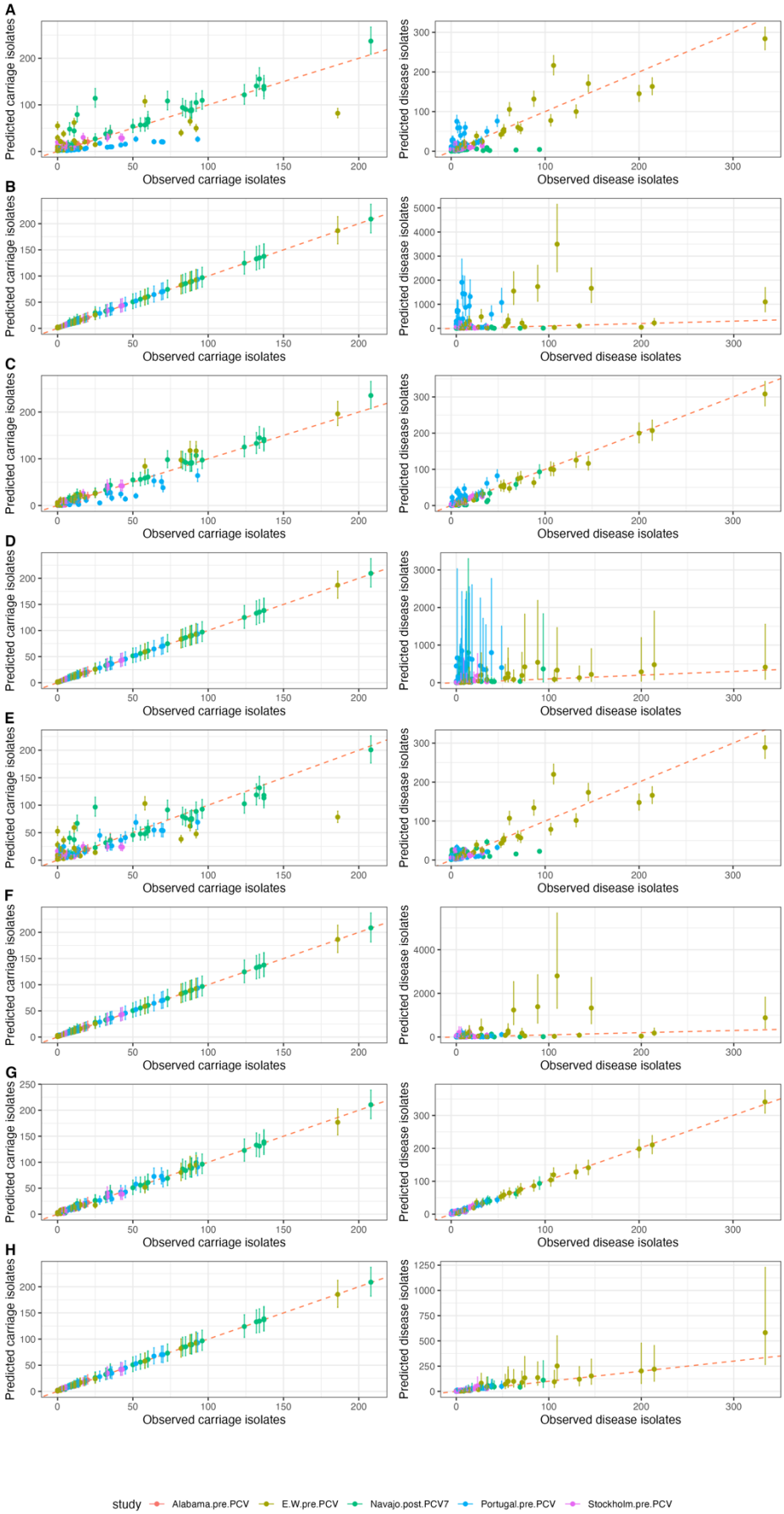


Figure S15 Histograms of Rhat values calculated from paired MCMCs for models fitted to the filtered serotype data from child carriage and adult disease. Each panel corresponds to a different model: (A) null Poisson model; (B) null negative binomial model; (C) type-specific Poisson model; (D) type-specific negative binomial model; (E) study-adjusted Poisson model; (F) study-adjusted negative binomial model; (G) study-adjusted type-specific Poisson model; (H) study-adjusted type-specific negative binomial model.


A
B
C
D
E
F
G
H
Predicted carriage isolates
Observed carriage isolates
Predicted disease isolates
Observed disease isolates
study
Alabama.pre.PCV
E.W.pre.PCV
Navajo.post.PCV7
Portugal.pre.PCV
Stockholm.pre.PCV

1

Figure S16 Comparison of observed and predicted counts of each serotype within each adult study. The points are coloured by the study to which they correspond, and represent the observed value on the horizontal axis, and the mean predicted value on the vertical axis. The error bars show the 95% credibility intervals. The red dashed line shows the line of identity, corresponding to a perfect match between prediction and observation. The left column shows the correspondence for carriage data (values of $c_{i,j}$), and the right column shows the correspondence for disease isolates (values of $d_{i,j}$). Each row corresponds to a different model: (A) null Poisson model; (B) null negative binomial model; (C) type-specific Poisson model; (D) type-specific negative binomial model; (E) study-adjusted Poisson model; (F) study-adjusted negative binomial model; (G) study-adjusted type-specific Poisson model; (H) study-adjusted type-specific negative binomial model.

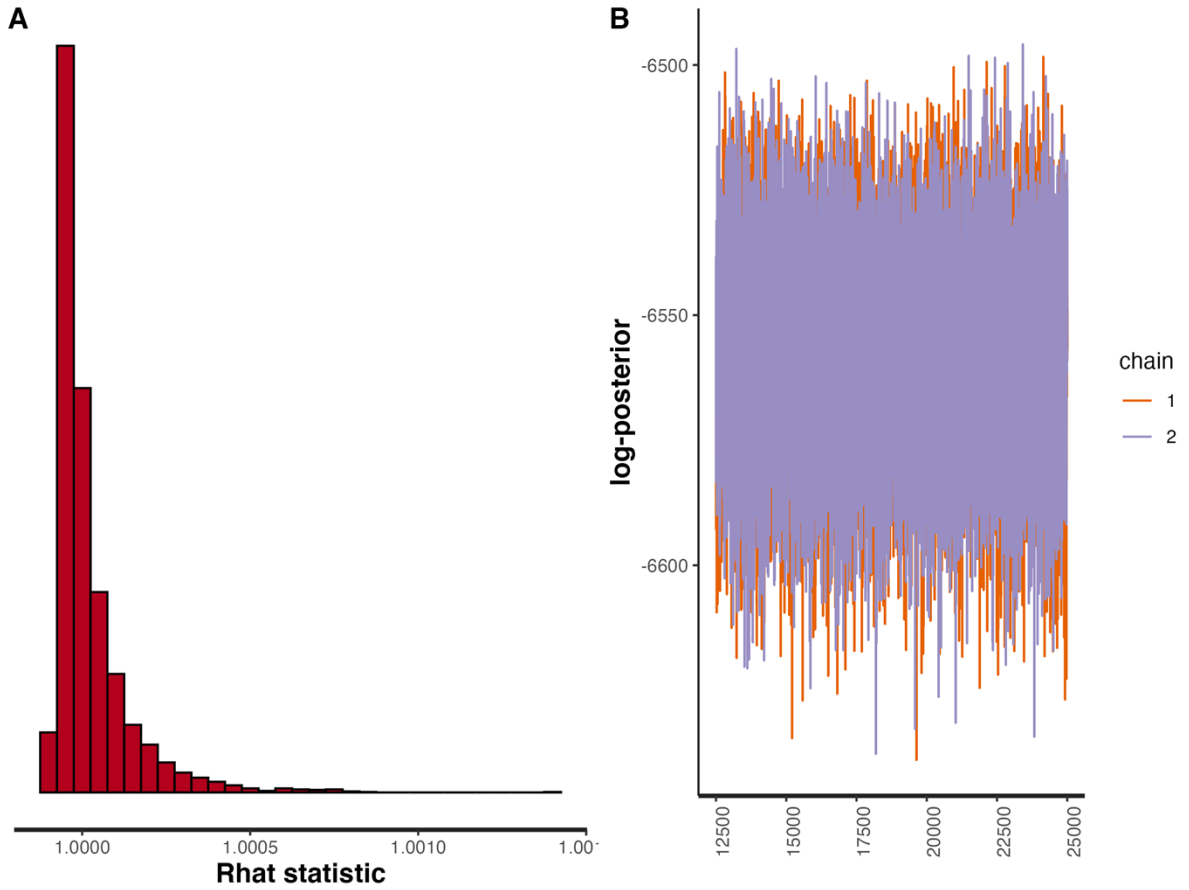


Figure S17 Plots validating the fit of the study-adjusted type-specific Poisson model to the full serotype data from child carriage and adult disease. (A) Histogram showing the distribution of Rhat values. (B) Post-warmup MCMC traces of the log posterior probability.

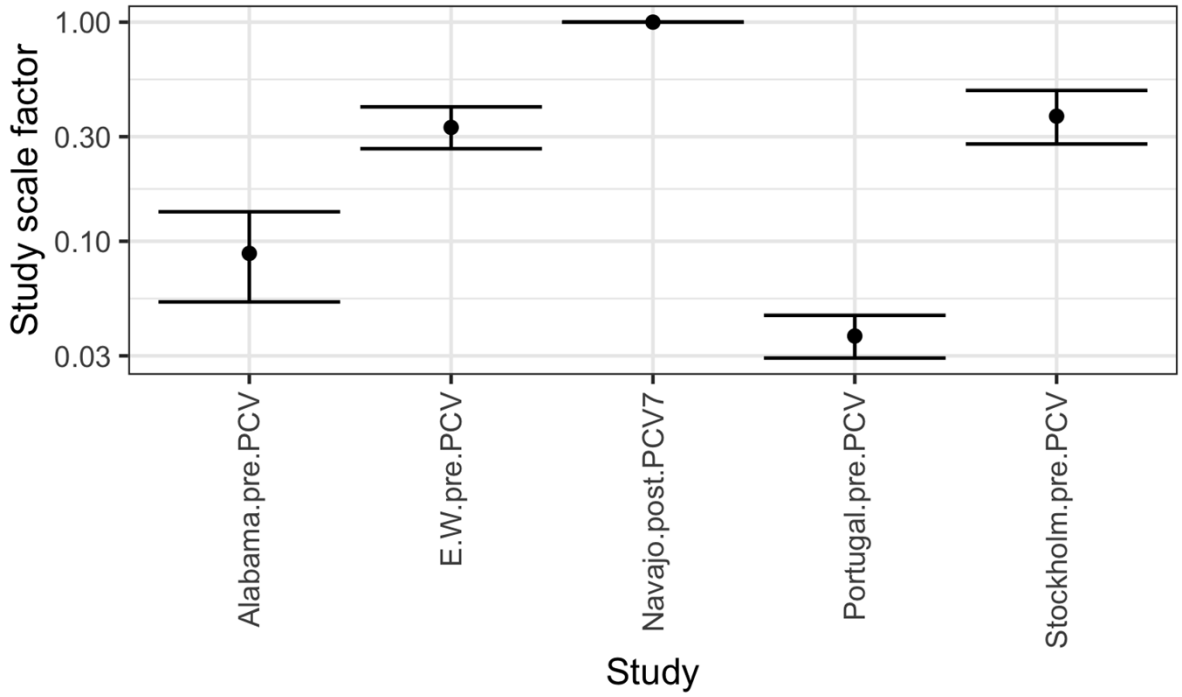


Figure S18 Study-adjustment scale factors from the study-adjusted type-specific Poisson model fitted to the full serotype data from child carriage and adult disease. The reference study, for which the value was fixed at one, was the Navajo post-PCV7 dataset, which had the greatest sample size in this meta-analysis (Fig. S11).

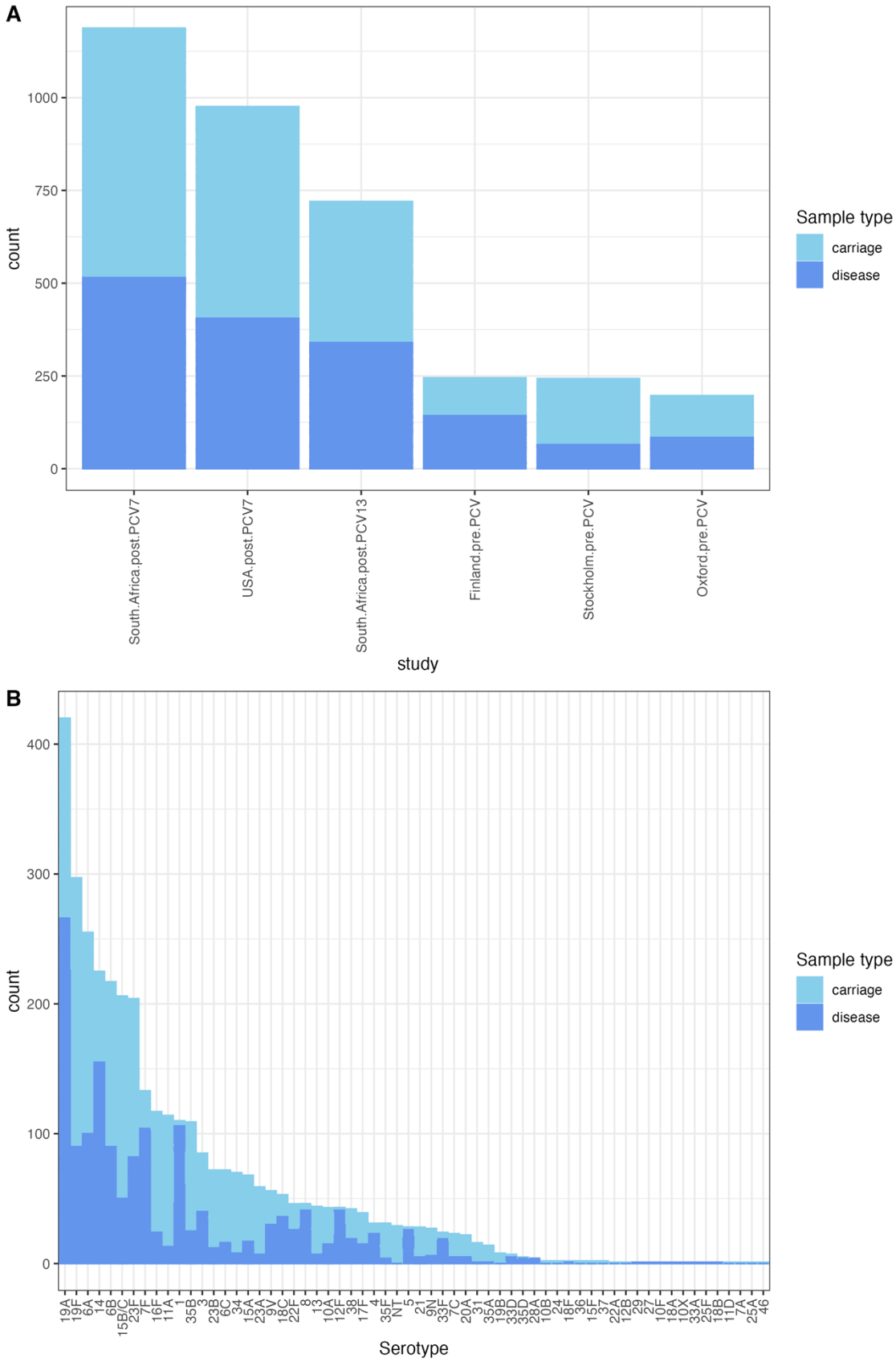


Figure S19 Properties of the child serotype and strain dataset. (A) Stacked bar plot showing the distribution of carriage and disease isolates between studies. (B) Stacked bar plot showing the distribution of carriage and disease isolates between serotypes.

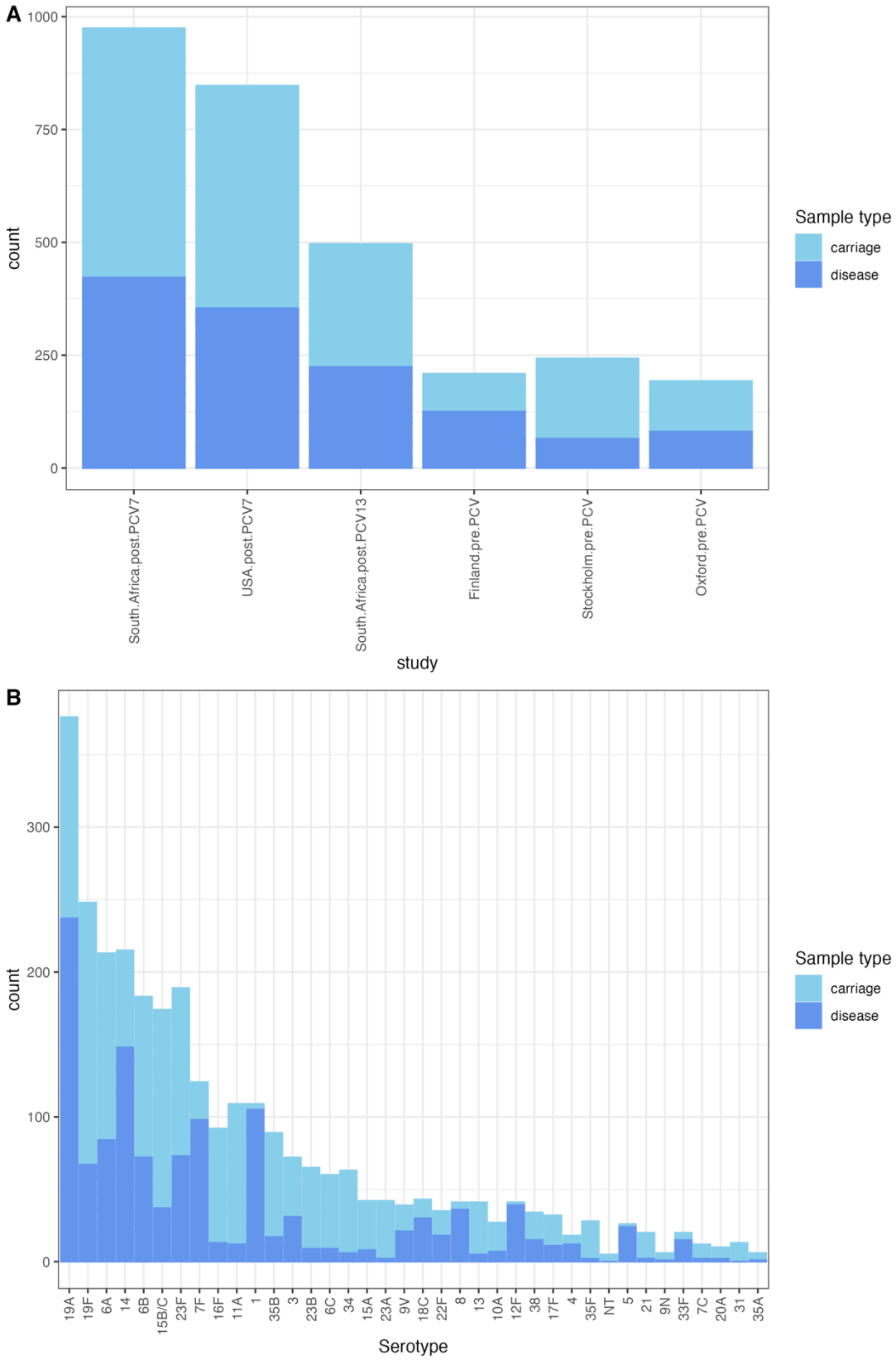


Figure S20 Properties of the child serotype and strain dataset after removing datapoints in which fewer than five isolates of a serotype were recorded from both carriage and disease in a given study. (A) Stacked bar plot showing the distribution of carriage and disease isolates between studies. (B) Stacked bar plot showing the distribution of carriage and disease isolates between serotypes.

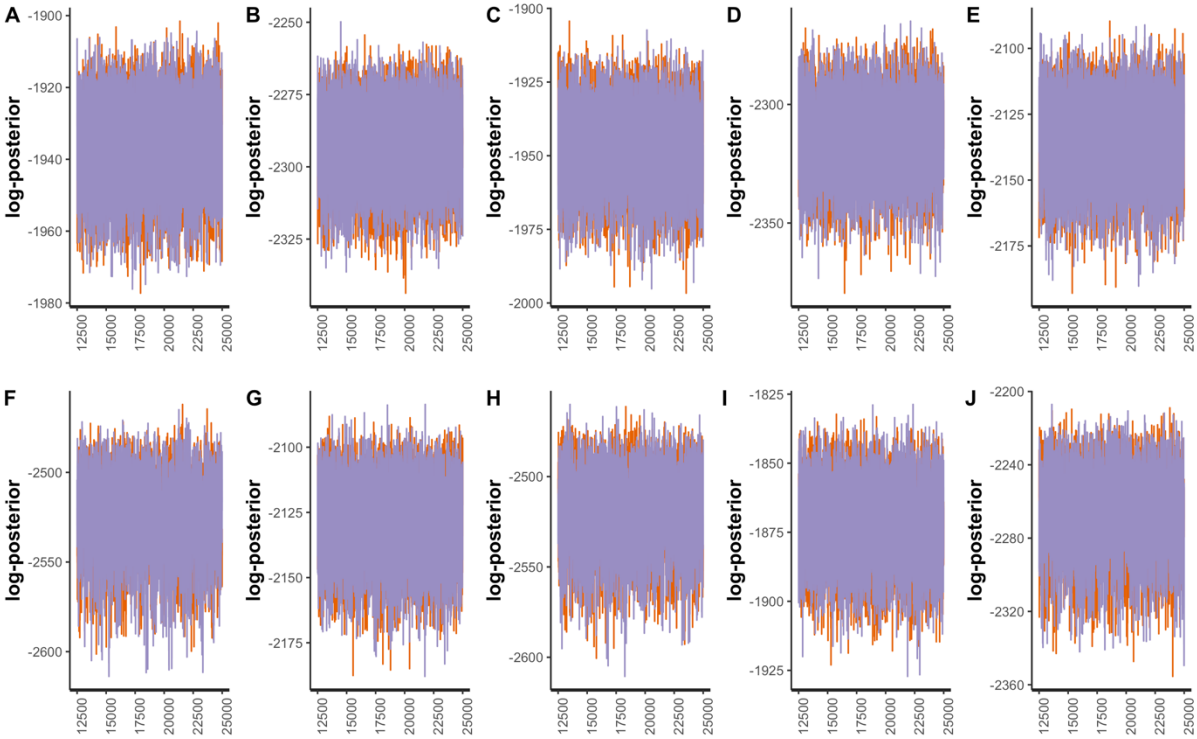


Figure S21 Line plots showing the post-warmup MCMC traces for the logarithmic posterior probabilities across two independent chains for models fitted to the filtered strain and serotype data from child carriage and disease. Each panel corresponds to a model with a different method of associating isolates with an invasiveness estimate: (A) serotype-determined, Poisson-distributed invasiveness; (B) serotype-determined, negative binomially-distributed invasiveness; (C) strain-determined, Poisson-distributed invasiveness; (D) strain-determined, negative binomially-distributed invasiveness; (E) serotype-determined, strain-modified Poisson-distributed invasiveness; (F) serotype-determined, strain-modified negative binomial model; (G) strain-determined, serotype-modified Poisson-distributed invasiveness; (H) strain-determined, serotype-modified negative binomially-distributed invasiveness; (I) strain- and serotype-determined Poisson-distributed invasiveness; (J) strain- and serotype-determined negative binomially-distributed invasiveness.

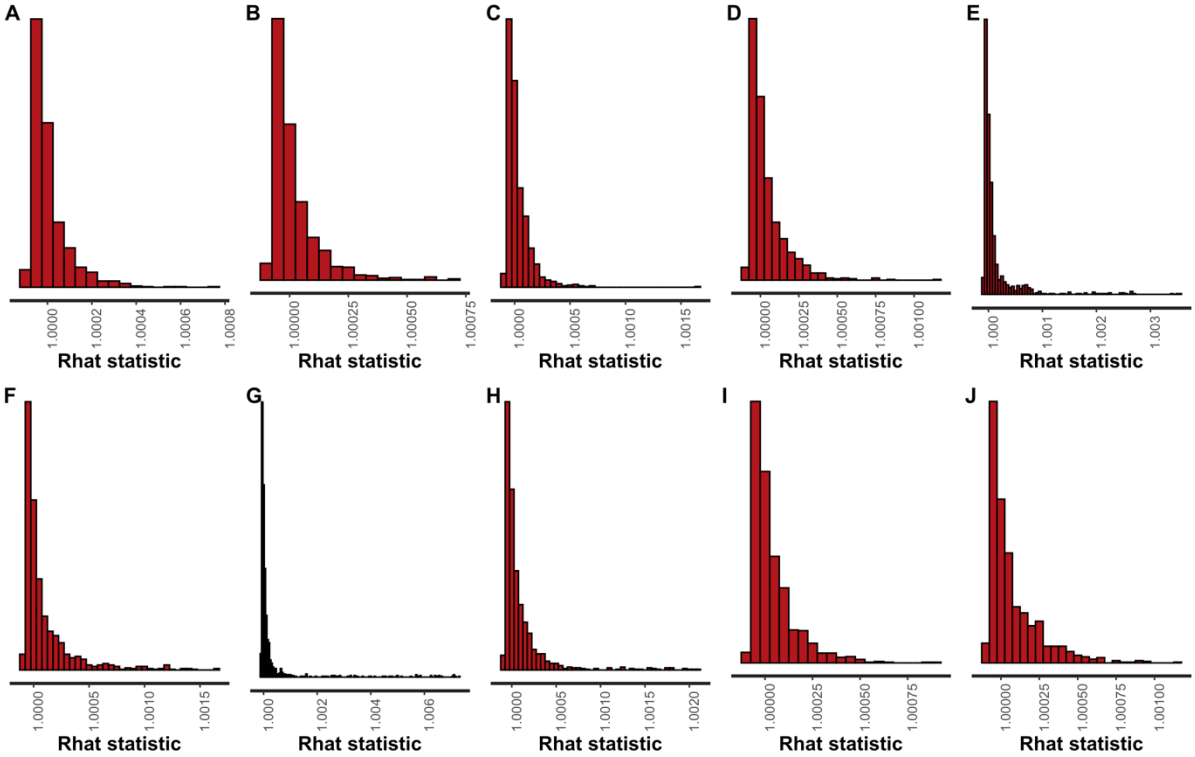
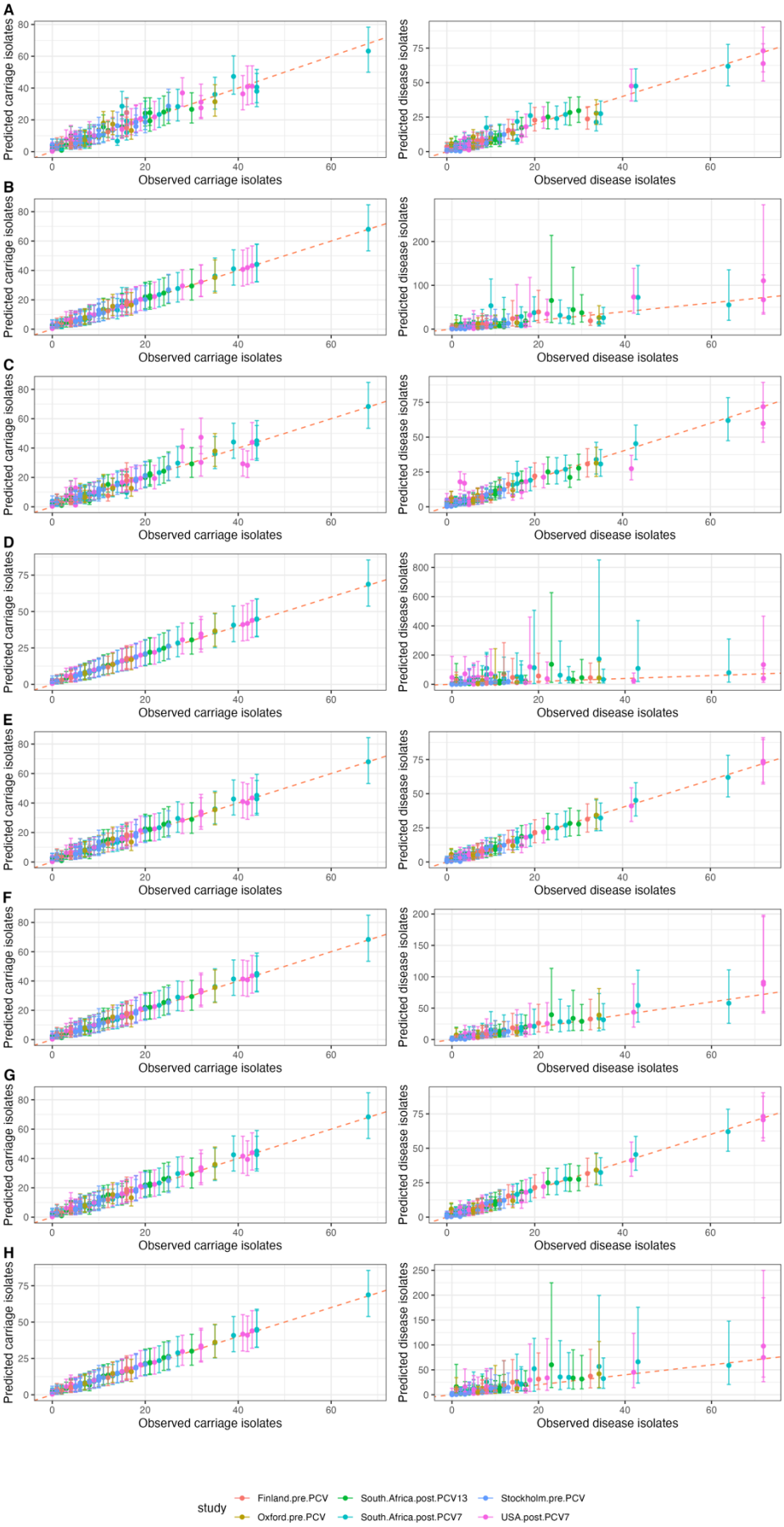


Figure S22 Histograms of Rhat values calculated from paired MCMCs for models fitted to the filtered strain and serotype data from child carriage and disease. Each panel corresponds to a different model: (A) serotype-determined, Poisson-distributed invasiveness; (B) serotype-determined, negative binomially-distributed invasiveness; (C) strain-determined, Poisson-distributed invasiveness; (D) strain-determined, negative binomially-distributed invasiveness; (E) serotype-determined, strain-modified Poisson-distributed invasiveness; (F) serotype-determined, strain-modified negative binomial model; (G) strain-determined, serotype-modified Poisson-distributed invasiveness; (H) strain-determined, serotype-modified negative binomially-distributed invasiveness; (I) strain- and serotype-determined Poisson-distributed invasiveness; (J) strain- and serotype-determined negative binomially-distributed invasiveness.


A
B
C
D
E
F
G
H
Predicted carriage isolates
Observed carriage isolates
Predicted disease isolates
Observed disease isolates
study
Finland.pre.PCV
Oxford.pre.PCV
South.Africa.post.PCV13
South.Africa.post.PCV7
Stockholm.pre.PCV
USA.post.PCV7

1

Figure S23 Comparison of observed and predicted counts of isolates, categorised by both their serotype and strain background, within each study. The points are coloured by the study to which they correspond, and represent the observed value on the horizontal axis, and the mean predicted value on the vertical axis. The error bars show the 95% credibility intervals. The red dashed line shows the line of identity, corresponding to a perfect match between prediction and observation. The left column shows the correspondence for carriage data (values of $c_{i.j.k}$), and the right column shows the correspondence for disease isolates (values of $d_{i.j.k}$). Each row corresponds to a different model: (A) serotype-determined, Poisson-distributed invasiveness; (B) serotype-determined, negative binomially-distributed invasiveness; (C) strain-determined, Poisson-distributed invasiveness; (D) strain-determined, negative binomially-distributed invasiveness; (E) serotype-determined, strain-modified Poisson-distributed invasiveness; (F) serotype-determined, strain-modified negative binomial model; (G) strain-determined, serotype-modified Poisson-distributed invasiveness; (H) strain-determined, serotype-modified negative binomially-distributed invasiveness; (I) strain- and serotype-determined Poisson-distributed invasiveness; (J) strain- and serotype-determined negative binomially-distributed invasiveness.

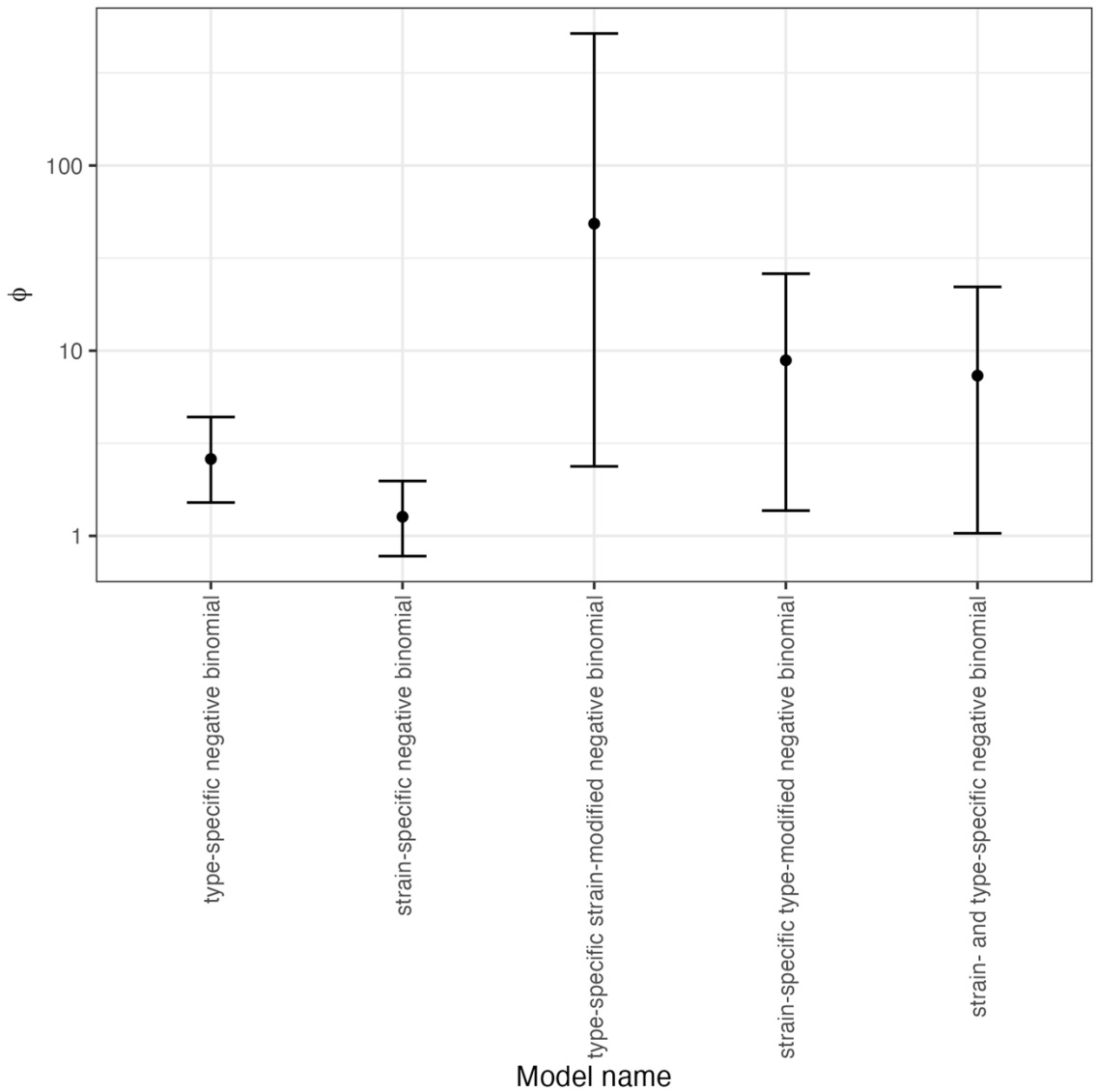


Figure S24 Graph showing the values of the negative binomial distribution's precision parameter, ϕ, from model fits to the filtered strain and serotype data from child carriage and disease. The points represent the mean estimate from the MCMCs, and the error bars show the 95% credibility interval.

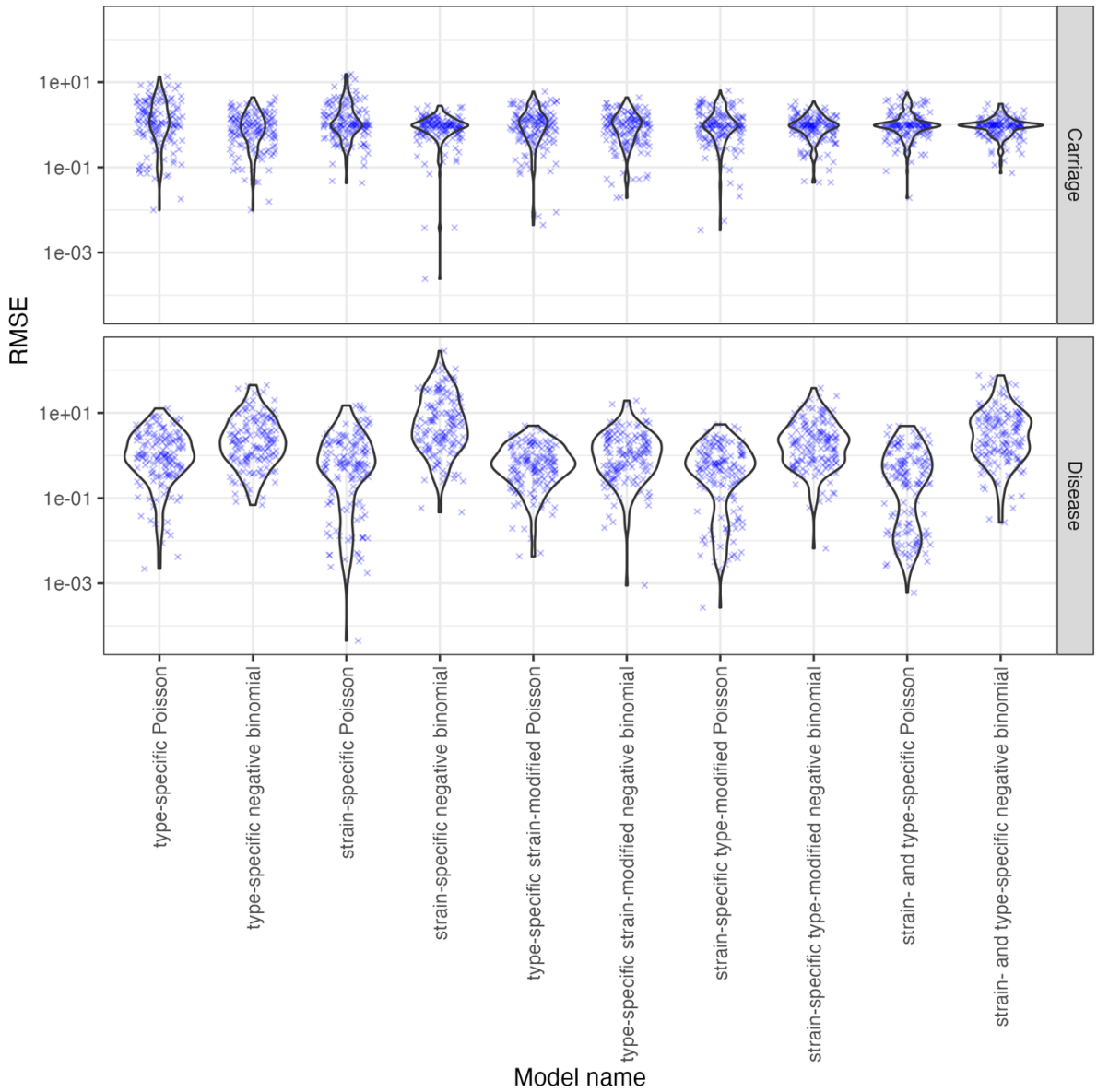


Figure S25 Violin plots showing the root mean square error between observed and predicted values across model fits to the filtered strain and serotype data from child carriage and disease. Blue crosses represent the individual observations.

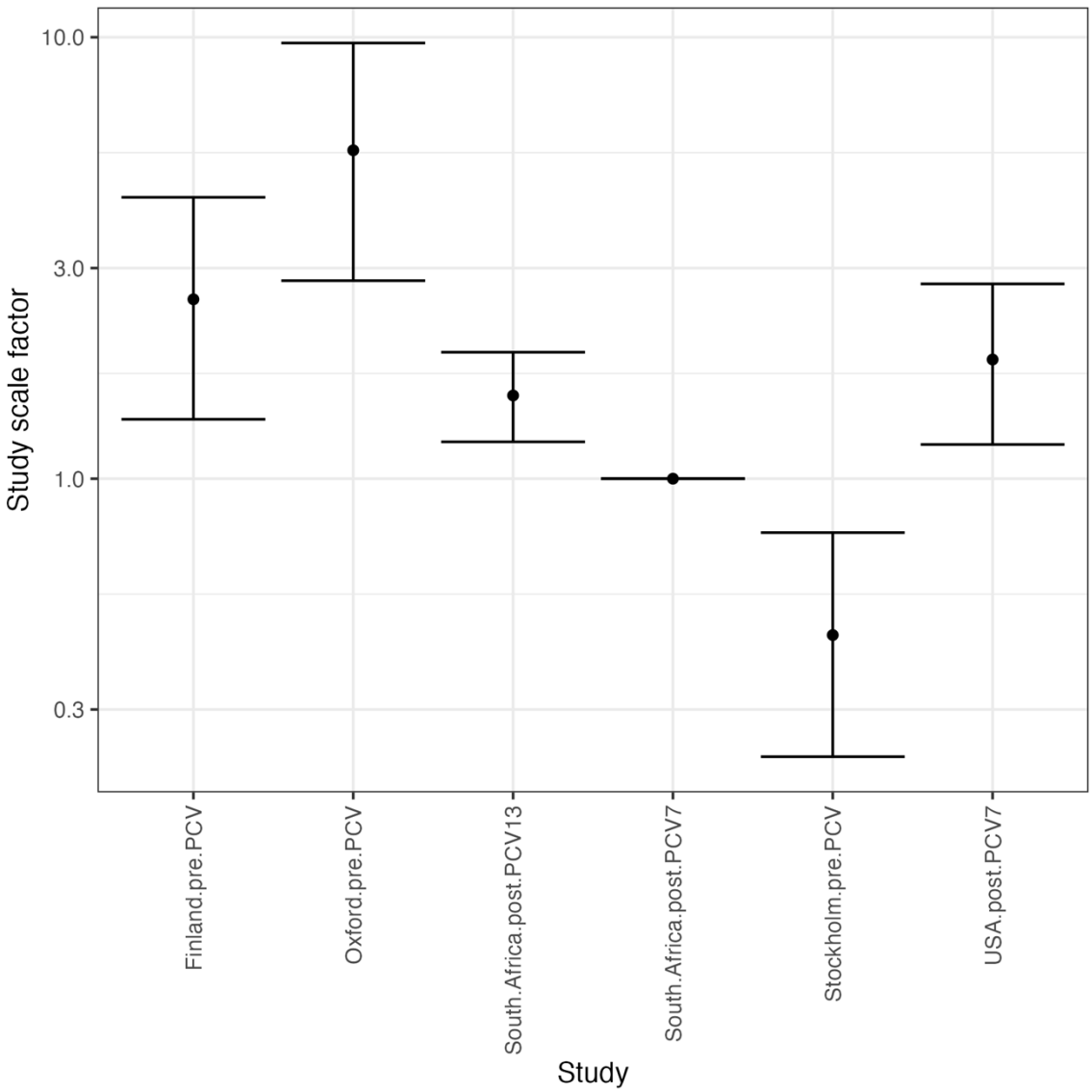


Figure S26 Study-adjustment scale factors from the study-adjusted type-specific Poisson model fitted to the full serotype and strain data from child carriage and disease, with the combination of both serotype and strain determining an isolate's invasiveness. The reference study, for which the value was fixed at one, was the South Africa post-PCV7 dataset, which had the greatest sample size in this meta-analysis, and was associated with a known carriage sample size (Text S1).

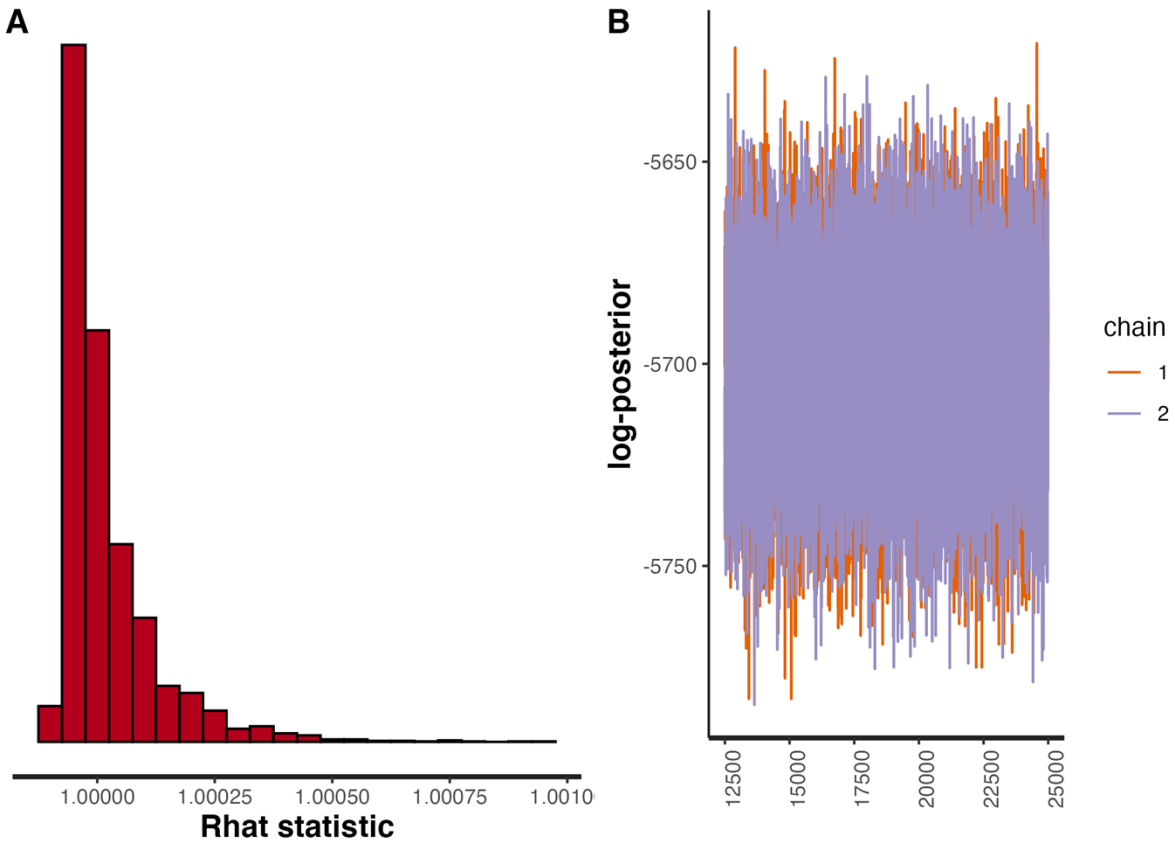


Figure S27 Plots validating the fit of the study-adjusted type-specific Poisson model to the full strain and serotype data from child carriage and disease, with the combination of both serotype and strain determining an isolate's invasiveness. (A) Histogram showing the distribution of Rhat values. (B) Post-warmup MCMC traces of the log posterior probability.

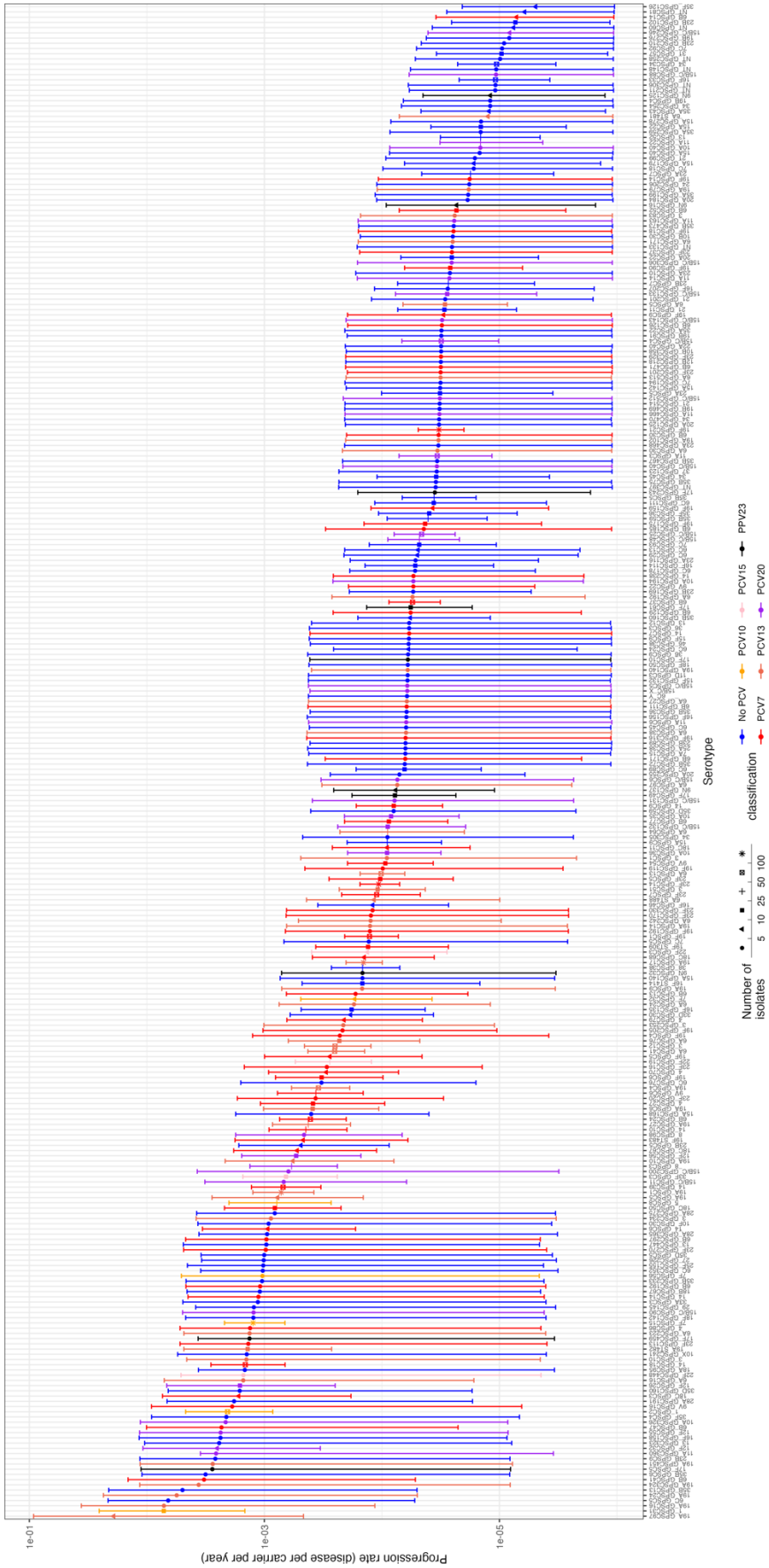


Figure S28 Invasiveness estimates for all strain and serotype combinations in the full strain and serotype dataset from child carriage and disease.

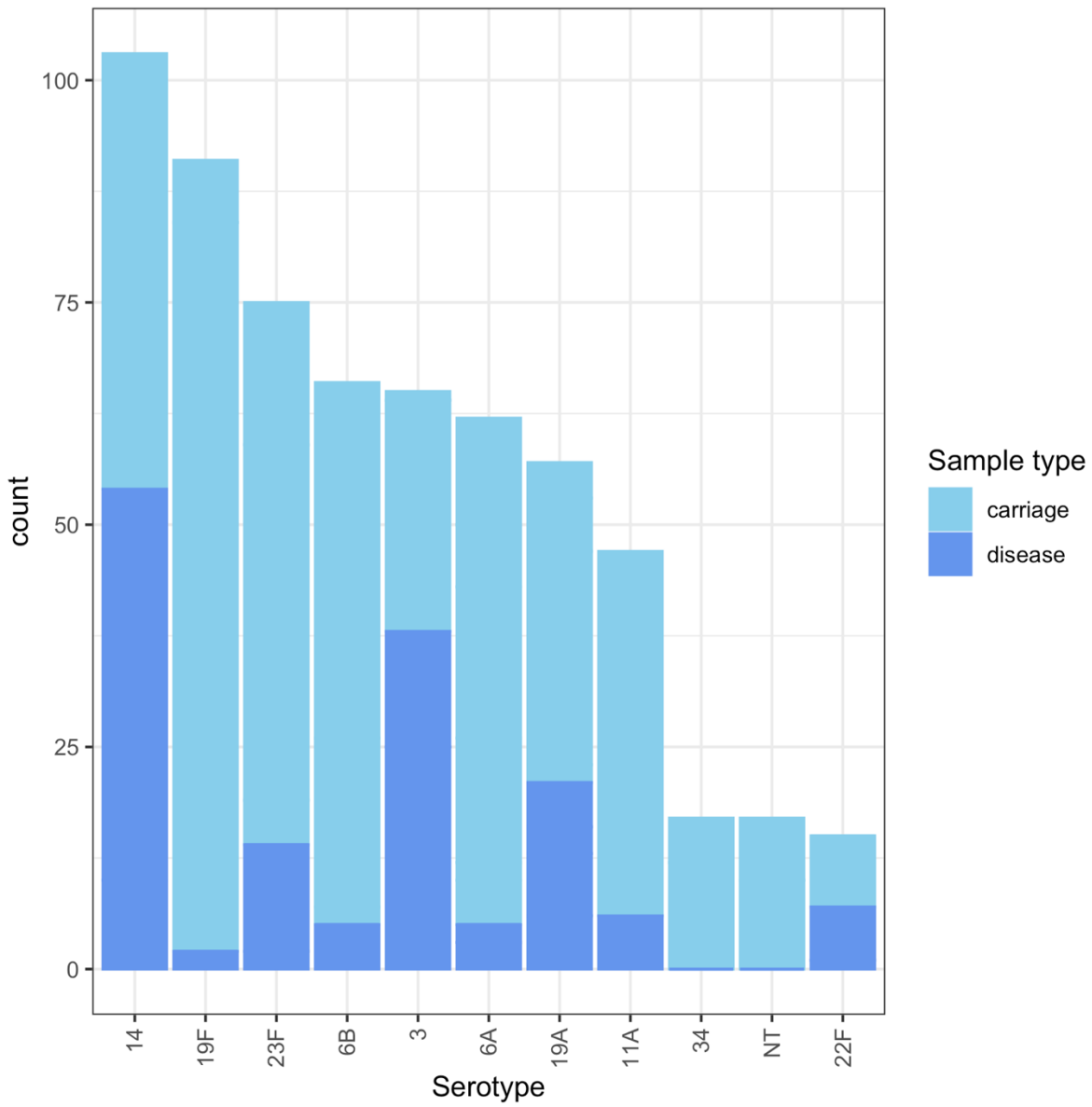


Figure S29 Bar plot showing the distribution of carriage and disease isolates between serotypes for the dataset from Portugal comparing child carriage with primarily adult disease.

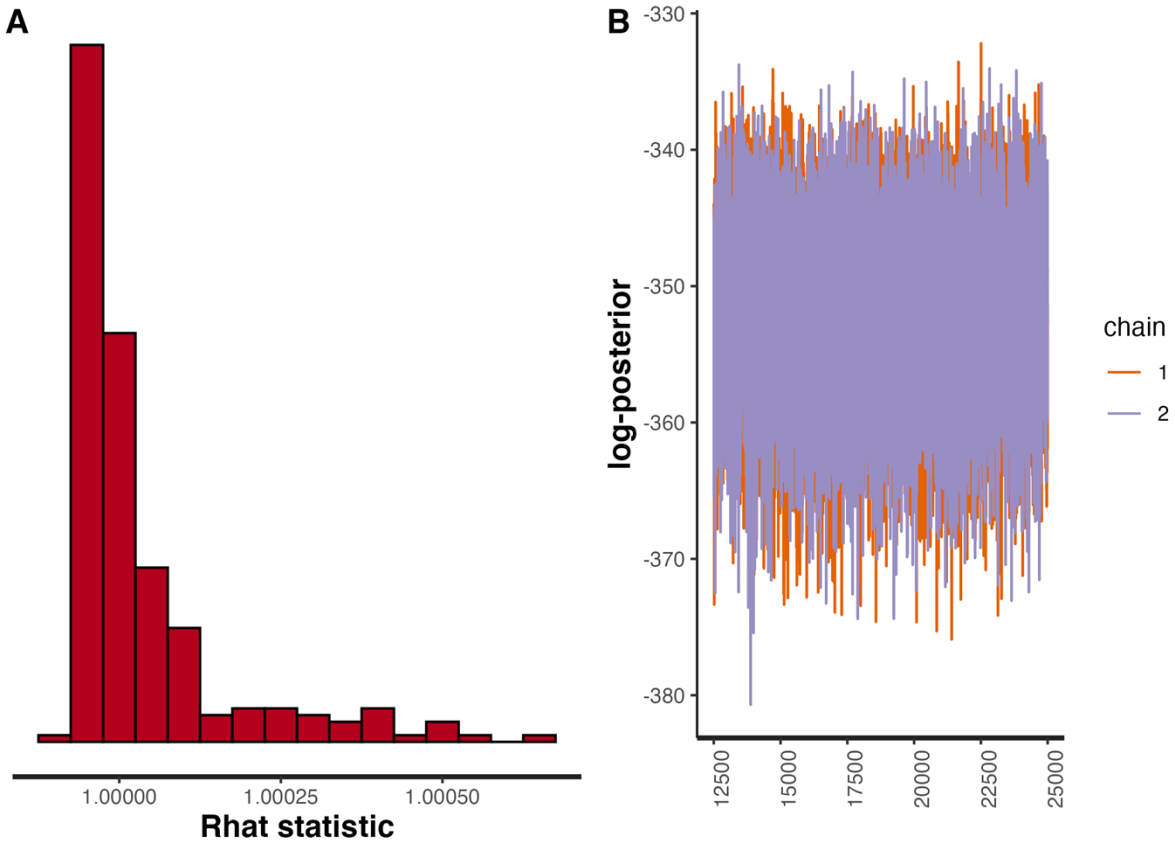


Figure S30 Plots validating the fit of the type-specific Poisson model to the strain and serotype data from child carriage and adult disease, with the combination of both serotype and strain determining an isolate's invasiveness. (A) Histogram showing the distribution of Rhat values. (B) Post-warmup MCMC traces of the log posterior probability.

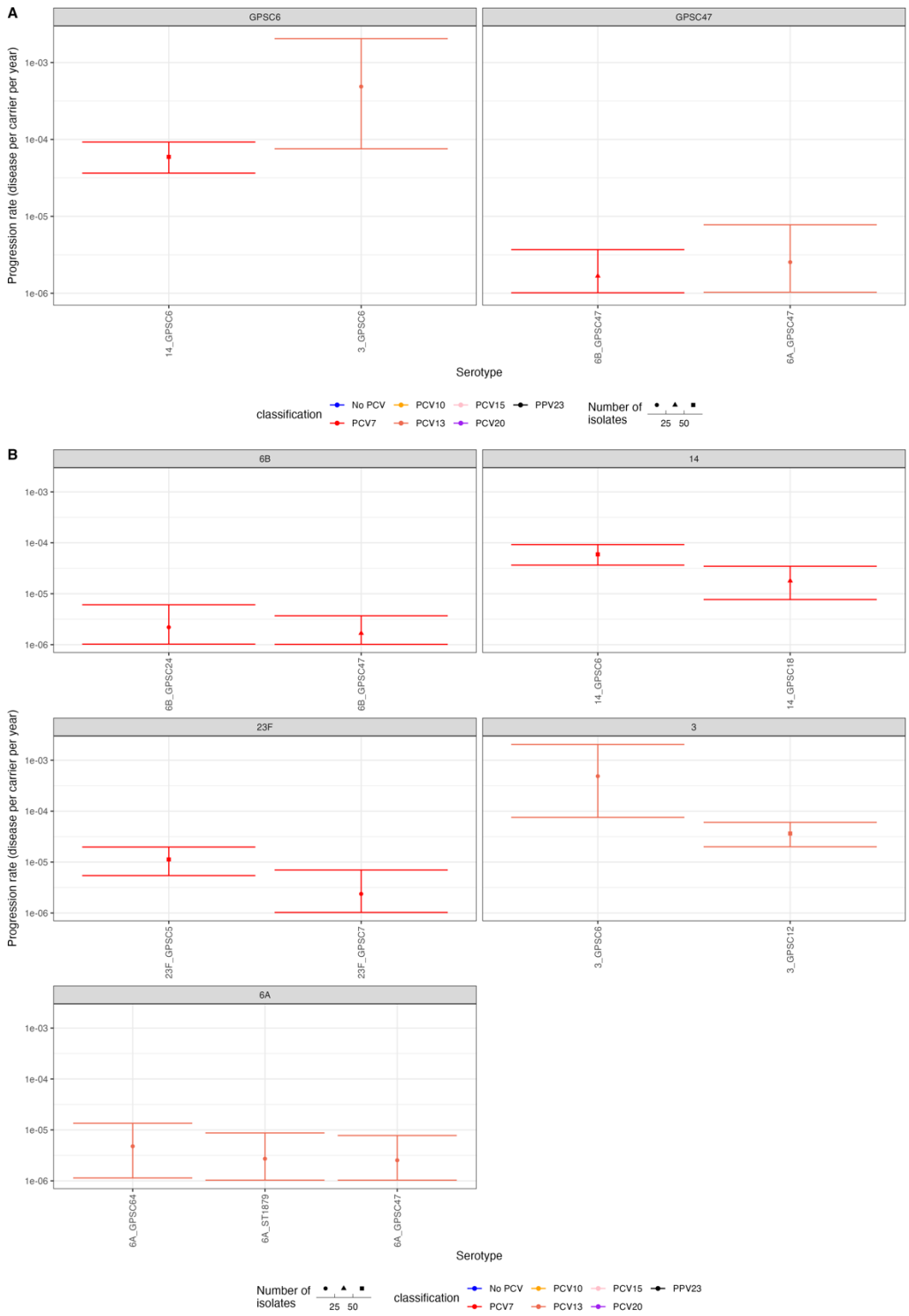


Figure S31 Invasiveness estimates for strain and serotype combinations represented by at least ten isolates in the study of child carriage and adult disease. Points represent the mean estimate, and are coloured by the vaccine formulations in which the corresponding serotype is present. The shape of the point represents the sample size on which the estimate is based. (A) Estimates are shown grouped by strain. Only strains for which estimates of multiple strain and serotype combinations were possible are displayed. (B) Estimates are shown grouped by serotype. Only serotypes for which estimates of multiple strain and serotype combinations were possible are displayed.

1 **Supplementary Tables**

2

Population	Vaccine period	Carriage study time interval	Carriage isolate source	Disease isolate source	No. of swabs	Population of children	No. disease isolates from children	Population of adults	No. disease isolates from adults	References
Alabama	Pre-PCV	July 1975 – December 1978	Unvaccinated sick and healthy children <18 years old from Alabama	Unvaccinated children with IPD <18 years old from Alabama and unvaccinated adults hospitalised with pneumonia or IPD in Alabama	827	19,316	114	232,373	86	[114,115]
Atlanta	Pre-PCV	January – December 1995	Unvaccinated children with recent URI in Atlanta <5 years old	Unvaccinated children <5 years old with IPD in Atlanta	231	204,680	202	-	-	[116]
Bogota	Pre-PCV	May 2005 – November 2006	Healthy unvaccinated children <18 months old in Bogota	IPD in children <2 years old in Bogota	197	357,200	353	-	-	[117]
Caracas	Pre-PCV	December 2006 – January 2008	Unvaccinated healthy children <6 years old in Caracas	Unvaccinated children with IPD <6 years old in Caracas	1,004	146,125	36	-	-	[118,119]
Czech	Pre-PCV	1996 – 2005	Unvaccinated healthy children 3 - 5 years old across the Czech Republic	Unvaccinated children <6 years old with IPD across the Czech Republic	425	478,177	138	-	-	[120,121]
England & Wales	Pre-PCV	July 1996 – June 2006	Unvaccinated healthy children	Unvaccinated children <5	3,752	3,091,000	461	48,702,414	1,876	[122]

			<5 years old in Hertfordshire	years old across and predominantly unvaccinated individuals >4 years old across England & Wales						
Goroka	Pre-PCV	1981 – 1987	Unvaccinated children attending clinics near Goroka Town	Unvaccinated children with IPD in Goroka hospital	2,844	96,207	56	-	-	[28,123]
Morocco	Pre-PCV	November 2010 – December 2011	Unvaccinated healthy children in Rabat	Children hospitalised with severe pneumonia in Rabat <5 years old	200	212,566	118	-	-	[124,125]
Netherlands	Pre-PCV	June 2004 – May 2006	Unvaccinated healthy children <2 years old in Noord-Holland, Zuid-Holland and Utrecht	Children <5 years with IPD across Netherlands	321	250,924	100	-	-	[126,127]
Ontario	Pre-PCV	1995	Unvaccinated healthy children primarily <4 years old in Toronto	Unvaccinated children <18 years old with IPD in Toronto and surrounding area	1,139	580,507	89	-	-	[128]
Portugal	Pre-PCV	January 2001 – December 2003	Unvaccinated healthy children <7 years old in Lisbon and Oeiras	Unvaccinated children (<18 years) and adults (>18 years) across Portugal	1,170	2,071,223	90	8,284,894	378	[52,129]

Stockholm	Pre-PCV	1997	Unvaccinated healthy children <7 years old in Stockholm	Unvaccinated adults (and a small number of children) hospitalised with IPD in Stockholm	611	-	-	2,004,152	273	[130,131]
Atlanta	Post-PCV7	June 2008 – May 2009	PCV7-vaccinated sick Georgia-resident children <5 years old	PCV7-vaccinated children <5 years old with IPD in Atlanta	451	298,831	47	-	-	[116]
Barcelona	Post-PCV7	2007 – 2011	PCV7-vaccinated healthy children <7 years old in Barcelona	PCV7-vaccinated children <7 years old with IPD in Barcelona	209	228,000	159	-	-	[132]
Bogota	Post-PCV7	June – November 2011	PCV7-vaccinated healthy children <18 months old in Bogota	IPD in children <2 years old in Bogota	246	357,200	91	-	-	[117]
France	Post-PCV7	January 2008 – December 2009	PCV7-vaccinated healthy children <2 years old across France	PCV7-vaccinated children with IPD <2 years old across France	1,212	838,866	388	-	-	[133,134]
Massachusetts	Post-PCV7	2001 – April 2009	PCV7-vaccinated children visiting physicians <7 years old across Massachusetts	PCV7-vaccinated children <7 years with IPD old across Massachusetts	2,969	820,000	206	-		[135]
Navajo	Post-PCV7	March 2006 - March 2008	PCV7-vaccinated Navajo or White Mountain	Active IPD surveillance of PCV7-vaccinated	6,541	65,048	132	201,553	514	[37,39,136]

			Apache native American children <9 years old	Navajo or White Mountain Apache native American children <7 years old and people at least 18 years old						
Netherlands	Post-PCV7	June 2008 – May 2012	Vaccinated healthy children <2 years old in Noord-Holland, Zuid-Holland and Utrecht	IPD in children (<5 years) across Netherlands	660	232,251	73	-	-	[126,127]
France	Post-PCV13	January 2012 – December 2013	PCV13-vaccinated healthy children <2 years old across France	PCV13-vaccinated children with IPD <2 years old across France	1,212	842,076	181	-	-	[133,134]
Netherlands	Post-PCV10	June 2012 – May 2016	Vaccinated healthy children <2 years old in Noord-Holland, Zuid-Holland and Utrecht	IPD in children (<5 years) across Netherlands	659	222,671	47	-	-	[126,127]

Table S1: Study populations used for the joint analysis of strain and serotype invasiveness.

Model	ELPD difference	ELPD difference standard error
study-adjusted type-specific Poisson	0	0
study-adjusted type-specific negative binomial	-11.61731	13.49535
type-specific negative binomial	-179.09155	23.2183
study-adjusted negative binomial	-239.06206	25.67848
null negative binomial	-283.58518	25.5161
type-specific Poisson	-825.94846	138.34479
study-adjusted null Poisson	-920.16588	112.08258
null Poisson	-1774.3198	213.90718

Table S2 Model fit comparison with leave-one-out (LOO) cross-validation using the child serotype data on carriage and disease. These values were generated using the logarithm of the likelihoods calculated within the models. Models are ranked by their expected log pointwise predictive density (ELPD), calculated from the individual pointwise log predictive densities across all observed data points. The ELPD difference column shows the difference between the ELPD of a model and that of the most likely model (this value is zero for the first row, corresponding to the most likely model given the data). The ELPD difference standard error is calculated from the distribution of individual pointwise log predictive densities from the same comparison.

Model	ELPD difference	ELPD difference standard error
study-adjusted type-specific negative binomial	0	0
study-adjusted type-specific Poisson	-16.68564	10.50462
type-specific negative binomial	-148.19801	13.72573
study-adjusted negative binomial	-203.23604	16.36764
null negative binomial	-249.59283	17.21276
type-specific Poisson	-729.33726	112.9739
study-adjusted null Poisson	-821.7134	96.65528
null Poisson	-1638.1251	190.51644

Table S3 Model fit comparison with LOO cross-validation using the child serotype data on carriage and disease. These values were generated using the logarithm of the likelihoods calculated for the observations of isolates from disease only, as these were more constrained than the modelling of isolate counts from carriage. The table is displayed as described for Table S2.

Model	Log(Bayes factor) relative to most likely model
study-adjusted type-specific Poisson	0.0000
type-specific Poisson	-406.5216
type-specific negative binomial	-459.7201
null negative binomial	-600.7347
study-adjusted type-specific negative binomial	-666.8982
study-adjusted null Poisson	-848.4240
study-adjusted negative binomial	-876.0923
null Poisson	-1445.7665

Table S4 Comparison of model fits to child carriage and disease serotype data using Bayes factors calculated with bridge sampling. Models are ranked by their logarithmic marginal likelihoods. The logarithmic Bayes factors are calculated for each model relative to the most likely given the data; hence the value is zero for the first row.

Model	Log(Bayes factor) relative to most likely model
study-adjusted type-specific Poisson	0.0000
type-specific negative binomial	-131.7431
study-adjusted type-specific negative binomial	-185.9657
null negative binomial	-205.1582
study-adjusted negative binomial	-286.0869
type-specific Poisson	-362.1535
study-adjusted null Poisson	-1010.5423
null Poisson	-1787.1696

Table S5 Comparison of model fits to child carriage and adult disease serotype data using Bayes factors calculated with bridge sampling. The table is displayed as described for Table S4.

1

Population	Vaccine period	Carriage study time interval	Carriage isolate source	Disease isolate source	No. of swabs	Population of children	No. disease isolates from children	Population of adults	No. disease isolates from adults	References
Finland	Pre-PCV	1994 - 1996	Unvaccinated children <2 years old in Tampere	Unvaccinated children <2 years across Finland	329	120,238	143	-	-	[71,137]
Oxford	Pre-PCV	1994 - 2001	Unvaccinated healthy children <5 years old in Oxford	Unvaccinated children <5 years old with IPD in Oxford	639	37,467	84	-	-	[29,138]
Portugal	Pre-PCV	January 2001 - December 2003	Unvaccinated healthy children <7 years old in Lisbon and Oeiras	Unvaccinated children (<18 years) and adults (>18 years) across Portugal	1,170	2,071,223	-	8,284,894	152	[52,129]
Stockholm	Pre-PCV	1997 - 2004	Unvaccinated children <7 years old attending day care centres in Stockholm County	Unvaccinated children <18 years old from the Stockholm area	1,330	397,289	65	-	-	[75,111,112, 130]
South Africa	Post-PCV7	2009 - 2010	HIV negative children <13 years old in Soweto and Agincourt	HIV negative children <7 years old in South Africa	2,674	7,187,314	515	-	-	[35,102,104, 105]
USA	Post-PCV7	2006 - 2009	Children <7 years of age in Massachusetts	Children <7 years old in the USA ABCS regions	1,983	1,931,331	405	-	-	[35,91,107,1 39,140]

South Africa	Post-PCV13	2011 - 2013	HIV negative children <13 years old in Soweto and Agincourt	HIV negative children <7 years old in South Africa	2,023	7,329,006	340	-	-	[35,102,104, 105]

Table S6: Study populations used for the joint analysis of strain and serotype invasiveness. The disease isolates from Portugal came from a mixture of infants and adults, but are tabulated based on them primarily arising from the latter age category.

Model	ELPD difference	ELPD difference standard error
strain- and serotype-specific Poisson	0	0
strain-specific serotype-modified Poisson	-3.7248	5.477322
type-specific strain-modified Poisson	-6.287393	5.468617
serotype-specific Poisson	-37.42351	18.306771
serotype-specific negative binomial	-46.514338	12.630727
strain-specific serotype-modified negative binomial	-56.069404	7.98463
type-specific strain-modified negative binomial	-57.304676	7.837755
strain- and serotype-specific negative binomial	-67.860004	7.923821
strain-specific Poisson	-90.523468	29.006827
strain-specific negative binomial	-98.704763	13.503281

Table S7 Model fit comparison with LOO cross-validation using the child strain and serotype data on carriage and disease. These values were generated using the logarithm of the likelihoods calculated for the observations of isolates from carriage and disease. The table is displayed as described for Table S2.

Model	ELPD difference	ELPD difference standard error
strain- and serotype-specific Poisson	0	0
strain-specific serotype-modified Poisson	-6.069704	4.036103
type-specific strain-modified Poisson	-6.667989	4.181882
serotype-specific negative binomial	-33.721585	10.603748
serotype-specific Poisson	-43.800528	15.81467
type-specific strain-modified negative binomial	-48.519279	6.091823
strain-specific serotype-modified negative binomial	-50.379496	6.308592
strain- and serotype-specific negative binomial	-58.830455	6.314365
strain-specific Poisson	-78.341963	24.643501
strain-specific negative binomial	-83.515938	11.810553

Table S8 Model fit comparison with LOO cross-validation using the child strain and serotype data on carriage and disease. These values were generated using the logarithm of the likelihoods calculated for the observations of isolates from disease only. The table is displayed as described for Table S2.

Model	Log(Bayes factor) relative to best-fitting model
strain- and serotype-specific Poisson	0
serotype-specific Poisson	-77.5832
strain-specific Poisson	-86.35369
strain-specific serotype-modified Poisson	-258.18324
type-specific strain-modified Poisson	-260.87589
strain- and serotype-specific negative binomial	-334.62845
strain-specific negative binomial	-371.46834
serotype-specific negative binomial	-388.66057
type-specific strain-modified negative binomial	-575.30246
strain-specific serotype-modified negative binomial	-575.70213

Table S9 Comparison of model fits to child carriage and disease strain and serotype data using Bayes factors calculated with bridge sampling. The table is displayed as described for Table S4.

Model	Log(Bayes factor) relative to best-fitting model
strain- and serotype-specific Poisson	0.00000
serotype-specific Poisson	-77.67605
strain-specific Poisson	-86.35044
type-specific strain-modified Poisson	-257.60470
strain-specific serotype-modified Poisson	-267.16176
strain- and serotype-specific negative binomial	-334.11318
strain-specific negative binomial	-371.16642
serotype-specific negative binomial	-388.88025
type-specific strain-modified negative binomial	-574.35959
strain-specific serotype-modified negative binomial	-577.58130

Table S10 Comparison of model fits to child carriage and disease strain and serotype data using Bayes factors calculated with bridge sampling. In this analysis, the carriage sample size for three studies (Finland pre-PCV, Oxford pre-PCV and Stockholm pre-PCV) was increased 100-fold, to test for the sensitivity of the model comparisons to the uncertainty in this parameter (Text S1). The table is displayed as described for Table S4.

1

Previous publication	Serotype	Low invasiveness genotype	Low invasiveness strain	High invasiveness genotype	High invasiveness strain	Results from this study
Gladstone *et al* 2019	6A	CC172	GPSC5	CC1094	GPSC41	GPSC41-6A invasiveness significantly higher than GPSC5-6A
Gladstone *et al* 2019	14	CC63	GPSC9	CC15	GPSC18	GPSC18-14 invasiveness significantly higher than GPSC9-14
Gladstone *et al* 2019	16F	CC4088	GPSC33	CC30	GPSC46	Not tested
Gladstone *et al* 2019	19F	CC347	GPSC21	CC320	GPSC1	GPSC1-19F invasiveness significantly higher than GPSC21-19F
Gladstone *et al* 2019	23B	CC439	GPSC7	CC172	GPSC5	Not tested
Sá-Leão *et al* 2011	3	ST180	GPSC12	ST156	GPSC6	GPSC6 found to have significantly higher invasiveness
Sá-Leão *et al* 2011	6A	ST460, ST315, ST1879	GPSC47, GPSC64, ST1879	-	-	GPSC64 does not exhibit significantly divergent invasiveness; others not tested
Sá-Leão *et al* 2011	6B	ST176, ST315	GPSC24, GPSC47	-	-	Other GPSCs of this serotype not tested
Sá-Leão *et al* 2011	11A	ST62	GPSC3	-	-	Not tested
Sá-Leão *et al* 2011	14	-	-	ST156	GPSC6	More invasive than GPSC9-14
Sá-Leão *et al* 2011	19A	ST81	GPSC16	ST193, ST230	GPSC11, GPSC10	Not tested
Sá-Leão *et al* 2011	19F	ST177	GPSC44	-	-	Not tested
Sá-Leão *et al* 2011	22F	-	-	ST443	ST443	Not tested
Sá-Leão *et al* 2011	23F	ST176, ST439	GPSC24, GPSC7	-	-	GPSC7-23F has lower invasiveness than GPSC5-23F
Sá-Leão *et al* 2011	34	ST1439	GPSC45	-	-	Not tested
Brueggemann *et al* 2003	14	-	-	ST9	GPSC18	More invasive than GPSC9-14

Brueggemann *et al* 2003	14	-	-	ST124	GPSC39	More invasive than GPSC9
Brueggemann *et al* 2003	18C	-	-	ST113	GPSC50	Higher invasiveness than GPSC68-18C, but with overlapping credibility intervals
Hanage *et al* 2005	14	-	-	ST156	GPSC6	More invasive than GPSC9-14
Hanage *et al* 2005	19A	-	-	ST482	ST482	Higher invasiveness than GPSC17-19A
Hanage *et al* 2005	7F	-	-	ST191	GPSC15	Not tested
Hanage *et al* 2005	14	-	-	ST124	GPSC39	More invasive than GPSC9-14
Hanage *et al* 2005	6B	-	-	ST138	GPSC24	More invasive than other 6B GPSCs
Hanage *et al* 2005	19F	ST485	GPSC175	-	-	Low invasiveness, with overlapping credibility intervals with other GPSCs
Hanage *et al* 2005	11A	ST62	GPSC3	-	-	Similar to GPSC22

Table S11 Comparison of previously-identified within-serotype differences in strain invasiveness with those supported by this analysis. The exclusion of strain and serotype combinations represented by fewer than ten isolates, summed across carriage and disease, meant that some within-serotype tests between strains were not undertaken in this analysis.

Supplementary Datasets

Dataset S1 Invasiveness estimates for serotypes in children. The mean estimates, and lower and upper bounds of the 95% credibility interval, are listed for each serotype.

Dataset S2 Invasiveness estimates for serotypes in adults. The mean estimates, and lower and upper bounds of the 95% credibility interval, are listed for each serotype.

Dataset S3 Invasiveness estimates for serotype-strain combinations in children. The mean estimates, and lower and upper bounds of the 95% credibility interval, are listed for each serotype-strain combination.