PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Analysing pneumococcal invasiveness using Bayesian models of pathogen progression rates

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

1 Department of Infectious Disease Epidemiology, School of Public Health, St. Mary's Campus, Imperial College London, London, United Kingdom, **2** MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, School of Public Health, White City Campus, Imperial College London, London, United Kingdom

* n.croucher@imperial.ac.uk

OPEN ACCESS

Citation: Løchen A, Truscott JE, Croucher NJ (2022) Analysing pneumococcal invasiveness using Bayesian models of pathogen progression rates. PLoS Comput Biol 18(2): e1009389. https://doi.org/10.1371/journal.pcbi.1009389

Editor: Benjamin Althouse, University of Washington, UNITED STATES

Received: August 12, 2021

Accepted: January 28, 2022

Published: February 17, 2022

Data Availability Statement: 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.

Funding: NJC was supported by a Sir Henry Dale Fellowship, jointly funded by Wellcome and the Royal Society (grant no. 104169/Z/14/A; https://wellcome.org/; https://royalsociety.org/). AL was funded by an investigator-initiated grant from GlaxoSmithKline to NJC (https://www.gsk.com/

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.

en-gb/home/). This work was supported by the UK Medical Research Council and Department for International Development (grants MR/R015600/1 and MR/T016434/1; https://mrc.ukri.org/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: AL was funded by an investigator-initiated grant to NJC from GlaxoSmithKline (https://www.gsk.com/en-gb/home/), who manufacture the PCV10 vaccine. NJC has consulted for Pfizer (https://www.pfizer.com/), who manufacture the PCV13 and PCV20 vaccines.

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.

Introduction

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

Progression rates have been extensively studied in the nasopharyngeal commensal and respiratory pathogen *Streptococcus pneumoniae* (the pneumococcus). Over one hundred serotypes have been identified in this species, each corresponding to a structurally-distinct polysaccharide capsule [11,12]. These are clustered into 48 serogroups, based on serological cross-reactivity [13]. This capsule inhibits the clearance of pneumococci by the immune system, and is crucial to their ability to cause invasive pneumococcal disease (IPD) [14,15], infections of normally sterile anatomical sites. Hence the rate of progression from carriage to IPD is referred to as invasiveness [16]. It has long been assumed the capsule is an important determinant of pneumococcal invasiveness, based on differences in serotypes' ability to cause disease in animal experiments [17]. Epidemiological differences are also apparent in human disease. Paediatric serotypes (serotype 14 and serotypes within serogroups 6, 9, 19, 23) were identified as being common in infant disease [18,19]. Epidemic serotypes (such as 1, 2, 5 and 12F) were found to be capable of causing IPD outbreaks in adults [19], and therefore are likely to represent hyperinvasive types of *S. pneumoniae*.

Understanding these differences between serotypes has become crucial for *S. pneumoniae* epidemiology, owing to the introduction of polysaccharide conjugate vaccines (PCVs) for infant immunisation [20]. A seven valent PCV (PCV7) was introduced in the USA in 2000, which has been supplanted by ten- and thirteen-valent formulations (PCV10 and PCV13, respectively). The next generation of higher-valency vaccines (PCV15 and PCV20) will soon be introduced [21,22]. PCVs induce immune responses against a specific subset of serotypes, usually resulting in their elimination from carriage [23]. This contrasts with the 23-valent polysaccharide vaccine (PPV23) administered to older adults, which only protects against

symptomatic disease [20]. However, the overall frequency of *S. pneumoniae* colonisation typically remains stable after any PCV's introduction, owing to the replacement of vaccine serotypes by the plethora of non-vaccine serotypes [24,25]. Nevertheless, PCVs have usually proved highly effective at reducing IPD through facilitating the replacement of the pre-PCV carried population with pneumococcal serotypes associated with lower invasiveness [26,27]. Hence optimal PCVs are those which minimise the overall invasiveness of the carried pneumococcal population [28,29]. Therefore estimation of invasiveness across vaccine-targeted serotypes, and those which may replace them post-PCV, is vital for reducing the incidence of IPD. Such quantification of invasiveness is typically achieved through paired studies of serotypes' prevalence in geographically and temporally matched surveys of pneumococcal carriage and surveillance of IPD.

Multiple methods have been used to estimate invasiveness using such paired case and carrier data. The most common approaches use odds ratios: the ratio of isolates of a given serotype recovered from disease against those recovered from carriage, divided by the same ratio calculated across all other serotypes in the population [30,31]. However, this statistic is intrinsically imperfect, because even if a serotype causes disease at a consistent rate across populations, the mix of serotypes against which it is compared will differ between locations [32,33]. This geographical heterogeneity has been exacerbated by post-PCV serotype replacement, which has driven increasing divergence between countries' serotype compositions [34]. The interpretability of the odds ratios may be improved by standardising them relative to the geometric mean across all odds ratios [35], but this does not resolve the underlying problem, which likely contributes to the substantial heterogeneity observed for some serotypes when odds ratios are combined in meta-analyses [36]. This means carefully-conducted odds ratio analyses can be hampered by having to subsample data, and employ both fixed and random effects analyses, within a single study [37]. One solution has been to estimate invasiveness relative to a standard serotype that is common in both carriage and IPD across studied locations [36,38]. However, these properties mean such serotypes are likely to be targeted by PCVs. Correspondingly, the original standard, serotype 14 [31], has been eliminated by PCV7 in many settings. Similarly, PCV13 has removed the replacement standard serotype used in post-PCV7 studies, 19A [36].

Invasiveness has been more directly characterised as the ratio of disease cases relative to the estimated prevalence in carriage [27]. 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 [2,39]. One approach modelled invasiveness by jointly fitting distributions to carriage and disease data, then used random effects models to estimate invasiveness in different age groups [39]. 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 maximising the available information on each type, which may be infrequently observed in individual studies. This is particularly important in *S. pneumoniae*, as many genotypes that have emerged post-PCV were previously rare, and vary geographically in their prevalence [34,40]. 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 microbial population, or to one standard type. 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 [39,41].

Using a Bayesian approach also enables hypothesis testing through statistical comparison of alternative model structures [2,39]. This is crucial for understanding the host and pathogen factors that affect progression rates. For instance, the correlation between pneumococcal serotypes and invasiveness has not been indisputably established as a causal link. In *Haemophilus influenzae*, changes to an isolate's serotype altered its virulence in an animal model of disease in such a manner that reflected the epidemiology of human disease [42]. While some equivalent experiments in *S. pneumoniae* have replicated observations from human IPD [43,44], others have found alteration of an isolate's serotype did not change its invasiveness in an animal model [45,46], suggesting serotype-independent factors may contribute to this phenotype. Studies of isolates from disease and carriage using comparative genomic hybridisation [47,48] and whole genome sequencing [49,50] have both supported links between non-capsular loci and invasiveness. An unambiguous association cannot even be made between the distinctive phenotypes of the epidemic serotypes and their capsules, as these pneumococci tend to have low genetic diversity [37,51], and therefore their capsule polysaccharide synthesis locus is in linkage disequilibrium with many other variable genomic loci [52,53].

Epidemiological studies using genotyped and serotyped isolates have also found examples of within-serotype differences in invasiveness [35,54,55], although other studies found evidence of serotype being the primary determinant of invasiveness [31]. Genomic analysis by the Global Pneumococcal Sequencing project also identified differences in the invasiveness of strains of the same serotype [33]. However, given the multiple pairwise comparisons conducted in such analyses, it is 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. This is essential for deciding on the most effective methods for ongoing pathogen surveillance, and enabling the most efficient use of such data. Using appropriately-structured models, capable of combining information from multiple studies at the most informative level of resolution, will help identify emerging highly invasive types at the earliest possible opportunity. 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 continual 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 type j, strain k, or the combination of both, j,k. The "type" may be defined by any phenotypic categorisation; analyses of *S. pneumoniae* typically use serotyping. The "strain" can represent any coherent genetic subdivision of the population; for instance, analyses of *S. pneumoniae* can use the Global Pneumococcal Sequence Clusters (GPSCs) [37], but alternatively-defined "clades" or "lineages" could be employed instead. 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 category 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* [39], 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 over a unit of 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:

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

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 corrected 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:

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

All these parameters were assumed to be constant within a study over t_i. As all disease cases were assumed to develop independently, they were modelled as a Poisson process. However, relevant unmodelled variation, or oversampling of local transmission chains in either carriage or disease surveillance [56], could increase 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 equivalent to:

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

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 ($\gamma_i = 1$ for all i). Therefore $d_{i,x}$ is expected to be correlated with $c_{i,x}$:

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

Type-specific model

This model, similar to that Weinberger *et al* applied to a single population [39], allowed for variation in progression rate between types, but still assumed host populations to be homogeneous ($\gamma_i = 1$ for all i). Therefore $d_{i,x}$ depends on both the carriage prevalence and progression rate of x:

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

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:

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

Study-adjusted type-specific model

This model allowed for both variation in the progression rate between types x and study populations i:

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

Joint modelling of strain and type progression rates

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

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

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

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

Strain-determined type-modified progression rate. The second approach to analysing the independent effects of type and genetic background modelled progression rate as being

primarily determined by the strain to which an isolate belonged, which was modified by its type. This was calculated as $\nu_x = \nu_j \nu_k$, where the model priors meant ν_k was less constrained than ν_j. Hence most variation in ν_x should be attributed to the strain k.

Type and strain-determined progression rate. Each combination of type j and strain k was modelled as having a unique progression rate ($\nu_x = \nu_{j,k}$), 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. 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 ($\Sigma_x \rho_{i,x}$) 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* [29], and therefore a uniform prior was placed on the logarithm of ν_x in all models. The lower bound was set at one IPD case per million carriers per unit time, as the largest studies used surveillance of host populations consisting of tens of millions of individuals (S1 Table). The upper bound was ten cases per carrier per unit time, which would effectively correspond to an obligate pathogen for a bacterium with the carriage duration of *S. pneumoniae* [57]:

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

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 [58]:

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

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 burdens [59].

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 type-determined strain-modified, and strain-determined type-modified, progression rate models). The categorisation modelled as determining the progression rate (ν_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 $\mathrm{Y_y}$, sampled from a uniform distribution with narrower boundaries than those for the analogous Γ_i variable:

$$\mathrm{Y}_y \sim U(-1.25, 1.25)$$

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

$$log_{10}\left(v_y\right) = \mu_{mod} + \tau_{mod} tan(\mathrm{Y}_y)$$

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 an exponential distribution. Values of ϕ that are large relative to $\mathbb{E}[d_{i,x}]$ suggest there is little evidence of overdispersion, and therefore these parameter values should be associated with small prior probabilities. As $\mathbb{E}[d_{i,x}]$ must always be close to, or above, one in well-powered case-carrier studies, an exponential rate parameter of one was used to penalise the likelihood of the negative binomial model when ϕ was above this value. Yet this broad, always positive, distribution enabled estimation of smaller, albeit non-zero, ϕ values for representing highly overdispersed data:

$$\phi \sim Exp(1)$$

Model parameters are summarised in Table 1. All models were implemented using R [60] and Rstan [61]. Each model was fitted using two Hamiltonian Markov chain Monte Carlo (MCMC) evaluations for 25,000 iterations, of which half were warmup iterations. Convergence was assessed through analysing the MCMC traces, using the criteria of requiring all $\hat{R}$ values to be below 1.05, and no reports of divergent transitions in the sampling [58]. 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

The ability of models to recover known parameter values from synthetic datasets was tested by simulation, as described in S1 Text. Models were compared using likelihoods, independent of priors, through approximate leave-one-out cross-validation with the loo package [62]. The marginal likelihoods of different models, given the data, were compared with Bayes factors calculated using the bridgesampling package [63]. The same number of iterations were used to calculate the logarithmic marginal likelihoods as were used to fit the models. Evaluation of our results against odds ratio calculations used the metafor package [64], as described previously [29]. Model outputs were analysed and plotted with the tidyverse and ggpubr packages [65,66].

Table 1. Summary of parameters used in the models.

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 parameter	1 or $\log_{10}(\gamma)$~ Cauchy(μ_{study},τ_{study})
ϕ	Precision parameter of negative binomial distribution	ϕ~Exp(1)
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

https://doi.org/10.1371/journal.pcbi.1009389.t001

S. pneumoniae serotype data

The matched carriage and IPD serotype datasets were combined from two meta-analyses [29,36]. Twelve of the reported studies were omitted due to lack of publicly-available data [31,67–72]; difficultly in defining independent cross-sectional carriage samples [73,74]; small sample sizes once stratified by age [75], or study design biased towards particular serotypes [76]. If a PCV were introduced during a study, then samples were divided into pre-PCV and post-PCV datasets. This resulted in 21 systematically-sampled and comprehensively serotyped paired asymptomatic carriage and disease samples (S1 Table). Of these studies, 20 included data on child carriage and child IPD, and five included data on child carriage and adult IPD.

Data that were only resolved to serogroup level, based on the currently-known serotypes, were omitted from the analysis, although the overall number of samples taken (η_i) was not adjusted. However, data on serotypes in historical studies that are now known to correspond to multiple serotypes (e.g. serotype 6A now being resolved into 6A and 6C [77]) were included unaltered. Additionally, isolates of serotypes 15B and 15C were combined into the single 15B/C category.

S. pneumoniae strain data

Analysis of variation of invasiveness between *S. pneumoniae* strains in children used population genomic data from South Africa [37], a mixture of population genomic and genotyped data from the USA [37], and genotyped data from Finland [74], Oxford [31] and Stockholm [78]. The equivalent analysis of strain invasiveness using isolates mainly from adult disease used data from Portugal [54]. Some of the genotyped datasets were not as thoroughly documented as the serotyped datasets, and therefore it was not appropriate to fit models that lacked a study-specific scale factor that reduced their reliance upon accurate carriage sample information (S2 Text).

Results

Bayesian meta-analysis of *S. pneumoniae* serotype invasiveness

Twenty matched IPD case and nasopharyngeal carriage datasets were used to quantify *S. pneumoniae* serotype invasiveness in children (S1 Table). From these, 7,340 carriage isolates and 2,851 disease isolates were extracted, across 72 serotypes (S1 Fig). Multiple models of invasiveness were proposed to analyse these data (see Methods). The null model assumed there was no systematic difference between serotypes' invasiveness across datasets. The type-specific model assumed each serotype had a characteristic invasiveness that was consistent across locations. The study-adjusted model assumed invasiveness did not vary across serotypes, but could differ between studies. Finally, the study-adjusted type-specific model allowed invasiveness to vary between both types and studies. All four approaches to estimating the expected numbers of disease cases and carriage isolates were fitted assuming the number of disease cases would follow a Poisson distribution, and allowing for overdispersion by fitting a negative binomial distribution to these values, resulting in eight models overall. These models were able to accurately recover progression rates (S2 Fig) and infer overdispersion (S3 Fig) from simulated data (S1 Text), with the true values used to generate the synthetic datasets lying within the 95% credibility intervals of the estimates from the fit of the corresponding model.

There was limited power for distinguishing the different model structures when the counts of both carriage and disease isolates were low for a given type in a study. Therefore paired observations from an individual study were only used for the model comparison if either the number of carried, or disease, isolates was at least five. This reduced the dataset to 7,048 carriage isolates and 2,617 disease isolates across 45 serotypes (S4 Fig). The eight models were each fitted to the data using two Hamiltonian MCMC chains, both run for 25,000 iterations. All chains appeared to have converged, based on the traces of the posterior probabilities (S5 Fig) and all $\hat{R}$ values being below 1.001 (S6 Fig). Comparisons of the observed and predicted isolate counts showed all models could accurately reproduce the observations from carriage studies, with the exception of the null and study-adjusted Poisson models (Fig 1). However, only the most complex models, adjusting for both *S. pneumoniae* serotype and study, were able to accurately reproduce the disease case counts.

The deviations between the observed and predicted isolate counts can be quantified using the distributions of absolute residuals (S7 Fig). These show the negative binomial models replicate the carriage values closely, relative to the equivalent Poisson models, with correspondingly greater deviation from the disease counts enabled by the overdispersion permitted by the negative binomial distribution's precision parameter, ϕ. That the ϕ values were below one for the three simplest negative binomial models suggested there was substantial unmodelled variation (S8 Fig). However, for the study-adjusted type-specific model, the ϕ value rose above one, indicating less overdispersion. Correspondingly, the similarly low residuals for the Poisson and

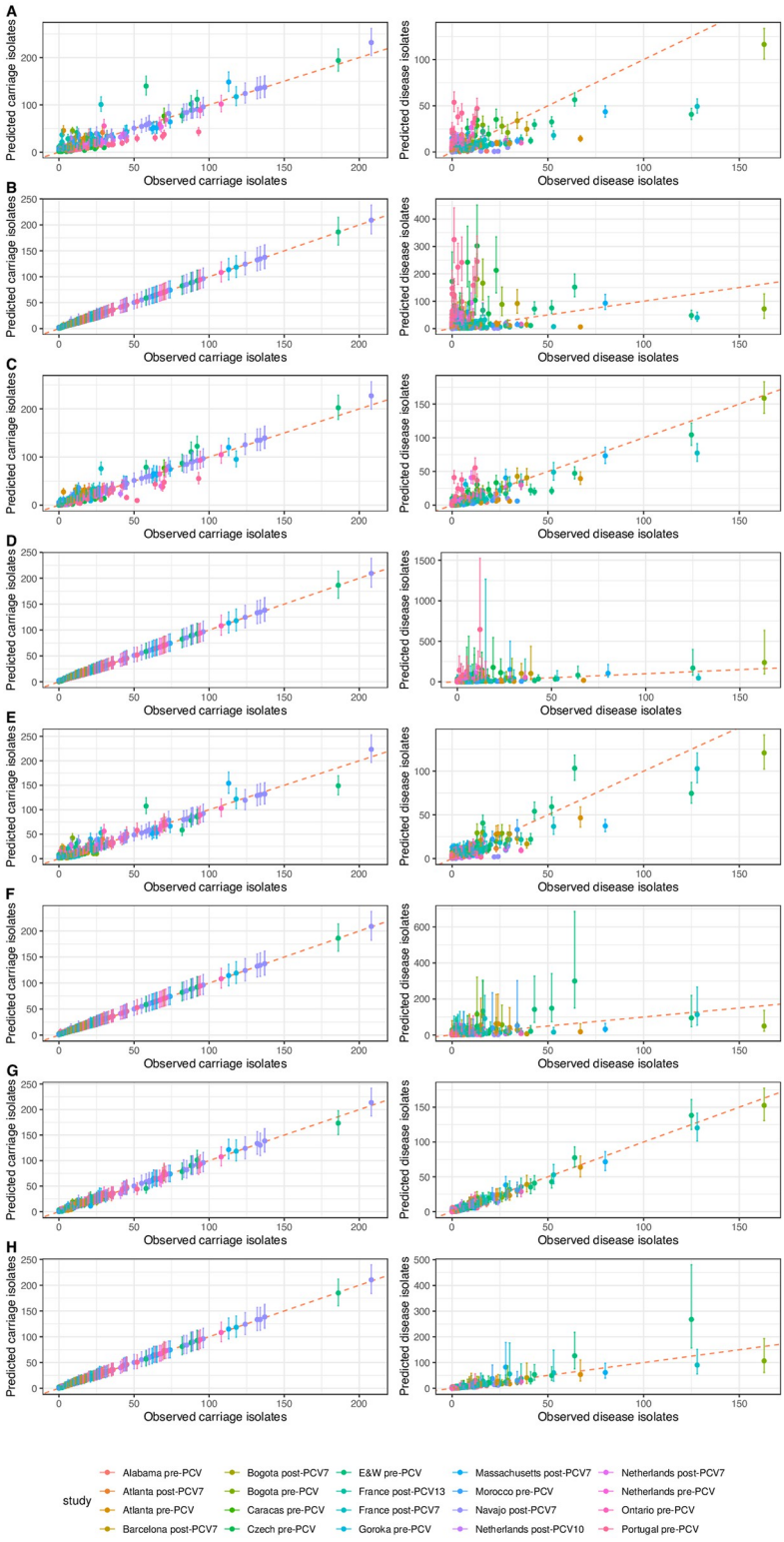

Fig 1. Comparison of observed and predicted counts of each serotype within each study of *S. pneumoniae* invasiveness in children. The points show the observed value on the horizontal axis, and the median predicted value on the vertical axis. The error bars show the 95% credibility intervals. The points and bars are coloured by the study to which they correspond; note that England and Wales is abbreviated to "E&W" (S1 Table). 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.

https://doi.org/10.1371/journal.pcbi.1009389.g001

negative binomial versions of this model suggested adjusting for study and serotype enabled progression rates to be estimated robustly with either distribution (S7 Fig).

The success of the study-adjusted type-specific models in reproducing the observations may represent overfitting by the most complex models. Therefore formal model comparisons were undertaken using leave-one-out cross-validation (LOO-CV) [62]. Although this suggested the study-adjusted type-specific models were the most appropriate, the large number of model parameters meant the Pareto k statistic diagnostic values were too high for these comparisons to be reliable (S2 Table) [79]. This was still true if the models were compared using only the likelihoods calculated from the disease counts, which were more constrained than those calculated from the carriage data (S3 Table). Hence models were instead compared using Bayes factors calculated from bridge sampling [63]. This approach was validated by demonstrating it was able to accurately assign simulated data to the model under which it was generated (S4 Table). The Bayes factors identified the null Poisson model as the most poorly fitting, whereas the study-adjusted type-specific negative binomial model was the most strongly favoured by the data (S5 Table).

Identification of highly invasive non-vaccine serotypes

This study-adjusted type-specific negative binomial model was therefore applied to the full dataset. Both the MCMC traces and $\hat{R}$ statistics indicated the model fit converged after 25,000 iterations (S9 Fig). This enabled the estimation of invasiveness for 72 serotypes (Fig 2 and S1 Dataset), including all 24 included in vaccine formulations that are currently licensed or under development [20–22]. The analysis found the serotypes associated with high invasiveness and narrow credibility intervals were the three serotypes added to PCV7 to generate PCV10 (1, 5 and 7F), which are also present in higher-valency PCVs, and serotype 12F, included in PCV20. Other serotypes not included in current vaccine designs, but likely to be highly invasive, were 9L, 19C, 24B, 24F, 25A, 27, 28A and 46. However, the credibility intervals of some of these estimates were substantial, resulting from small sample sizes.

The serotypes associated with low invasiveness and narrow credible intervals were 11A (included in PCV20), 6C, 21, 23A, 23B, 34, 35B and 35F. Many other serotypes (e.g. 10F, 11F, 19B, 33A, 35A, 36, 39, 42 and 45) were associated with low invasiveness values, but these estimates had wide credibility intervals. The higher uncertainty generally corresponded with lower sample sizes, although some of these serotypes had a total sample size above 10 (e.g. 10F, 35A and 42), but were nevertheless rare in IPD. The absolute estimates of invasiveness relate to the reference population for the model fit, which was the post-PCV7 sample from the Navajo nation [39], as this study contributed the largest number of isolates (S1 Fig). The model fits found the maximal progression rates were above 10^{-2} IPD cases per carrier per year, with the minimum point estimates below 10^{-4} IPD cases per carrier per year. However, the incidence of IPD in the Navajo population is high, with excellent disease surveillance [80], and therefore other studies were expected to report lower disease cases per carriage episode. Correspondingly, the study-adjustment scale factors were below one for most other datasets (S10 Fig), suggesting many locations would expect to detect around 30–50% of the number of IPD cases per carriage episode for the same carriage serotype composition.

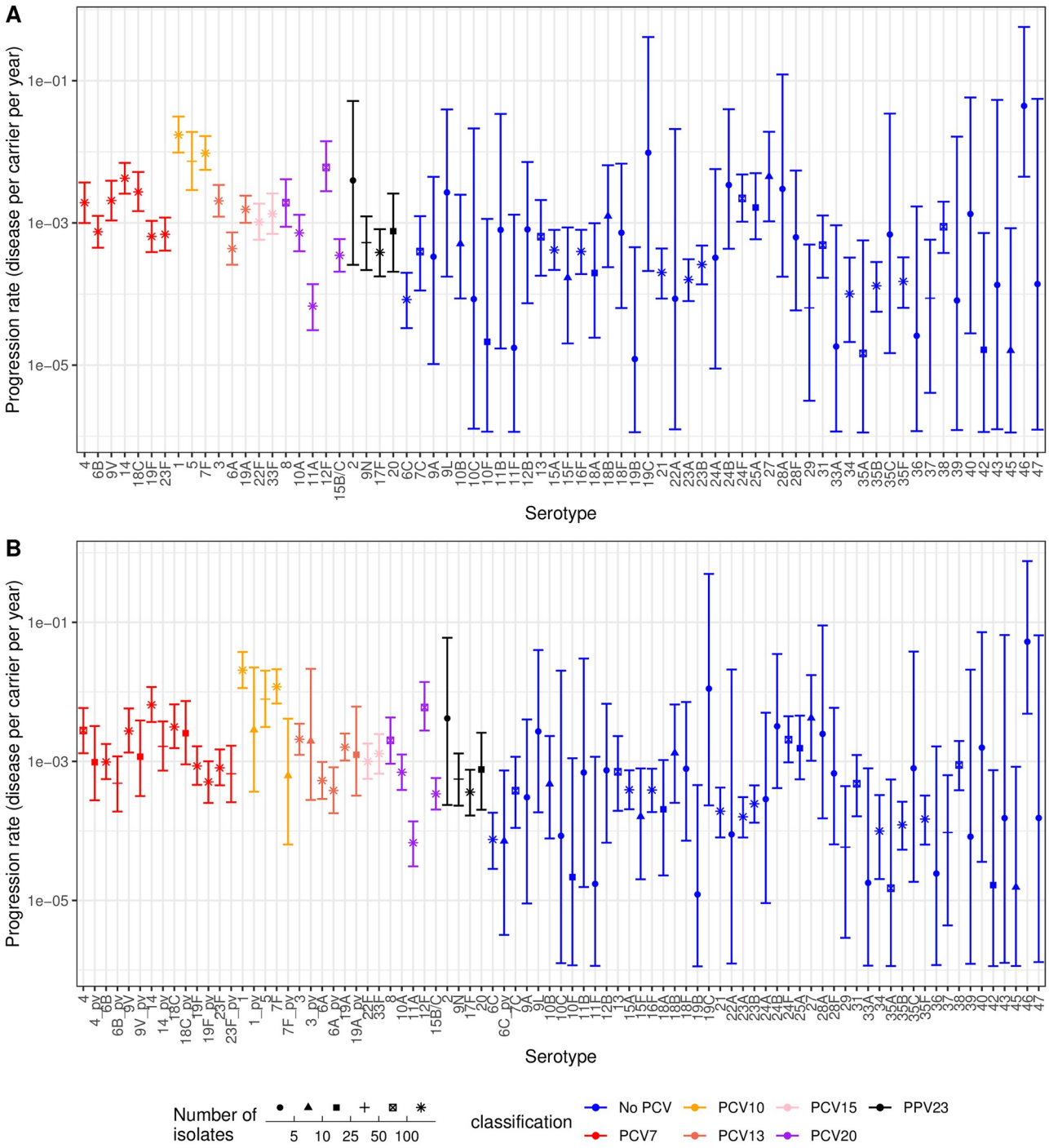

Fig 2. Invasiveness estimates for *S. pneumoniae* serotypes from the study-adjusted type-specific negative binomial model applied to the full dataset of serotype studies from child carriage and disease. Points represent the median estimate, and are coloured according to the vaccine formulation in which 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".

https://doi.org/10.1371/journal.pcbi.1009389.g002

The fitted study-adjustment scale factors were all associated with similarly narrow credibility intervals, demonstrating there were enough shared types between samples to robustly infer these parameters. For four locations (Atlanta, Bogota, the Netherlands and France), both pre-PCV and post-PCV samples were included in the meta-analysis, which might be expected to have similar scale factors, if the host population and disease surveillance were consistent across these eras. However, this was not the case for studies from Bogota. The model's assumption that serotype invasiveness is constant across studies may be violated by a reduction in vaccine-type invasiveness following PCV introduction, if vaccine-induced immunity inhibited progression from carriage to IPD. Therefore the study-adjusted type-specific negative binomial model was refitted, treating vaccine-type serotypes as different types pre- and post-PCV (including 6A as a PCV7 type, and 6C as a PCV13 type [81]; S11 Fig). The point estimates of invasiveness for most vaccine serotypes dropped following PCV introduction, with the evidence strongest for the invasiveness of serotypes 14 and 7F falling after the introduction of PCV7 and PCV10, respectively (Fig 2). This fit further improved the consistency of scale factors across the Netherlands and Atlanta, but did not resolve the discrepancy between samples from Bogota (S12 Fig). These data suggest PCVs can reduce the invasiveness of vaccine types before they are eliminated by herd immunity, and therefore distinguishing between pre- and post-PCV invasiveness may help standardisation across meta-analyses.

Comparison of Bayesian modelling with odds ratios

The outputs of the study-adjusted type-specific negative binomial model were compared to those from estimating pneumococcal invasiveness from the same dataset using odds ratios (Fig 3). Odds ratios were combined across studies using either random (Fig 3A) or fixed effects (Fig 3B) models, and the results disaggregated based on the total number of isolates, summed across carriage and disease, for each serotype. Across all sample sizes, there was a significant positive correlation between the two measures of invasiveness. The range of invasiveness values extended over two orders of magnitude for both methods. Hence the study-adjusted type-specific negative binomial model produced similar relative estimates of invasiveness to odds ratio analyses, while having the advantage of providing both overall, and location-specific, absolute estimates of these progression rates that can be used in quantitative analyses.

Fixed effects analyses assume there is a single invasiveness value across all locations, whereas random effects analyses allow for variation in the underlying invasiveness estimate between studies. The identification of between-location heterogeneity typically suggests random effects models are more appropriate for analyses of pneumococcal invasiveness [29,36,37]. This variation may either reflect genuine differences in serotype behaviour, or arise from standardisation relative to a variable mix of serotypes between locations. By accounting for this variation, random effects models often calculate wider confidence intervals than fixed effect models (Fig 3). Additionally, random effects models are typically not recommended where there are fewer than five studies in which a serotype features [82], and therefore meta-analyses of odds ratios may be a mixture of fixed- and random-effects models [37]. By contrast, this single, consistent Bayesian model can be applied across serotypes, regardless of their distribution across studies. Furthermore, this framework explicitly accounts for the uncertainty associated with overdispersed data through the negative binomial precision parameter, ϕ (Fig 1). This variation is quantified separately from the uncertainty in the progression rate estimate.

Deviations between the methods in the rank order of invasiveness point estimates were most evident at small sample sizes. This was most notable for serotypes 33A and 36, both of which were associated with relatively high logarithmic odds ratios (above zero) and relatively

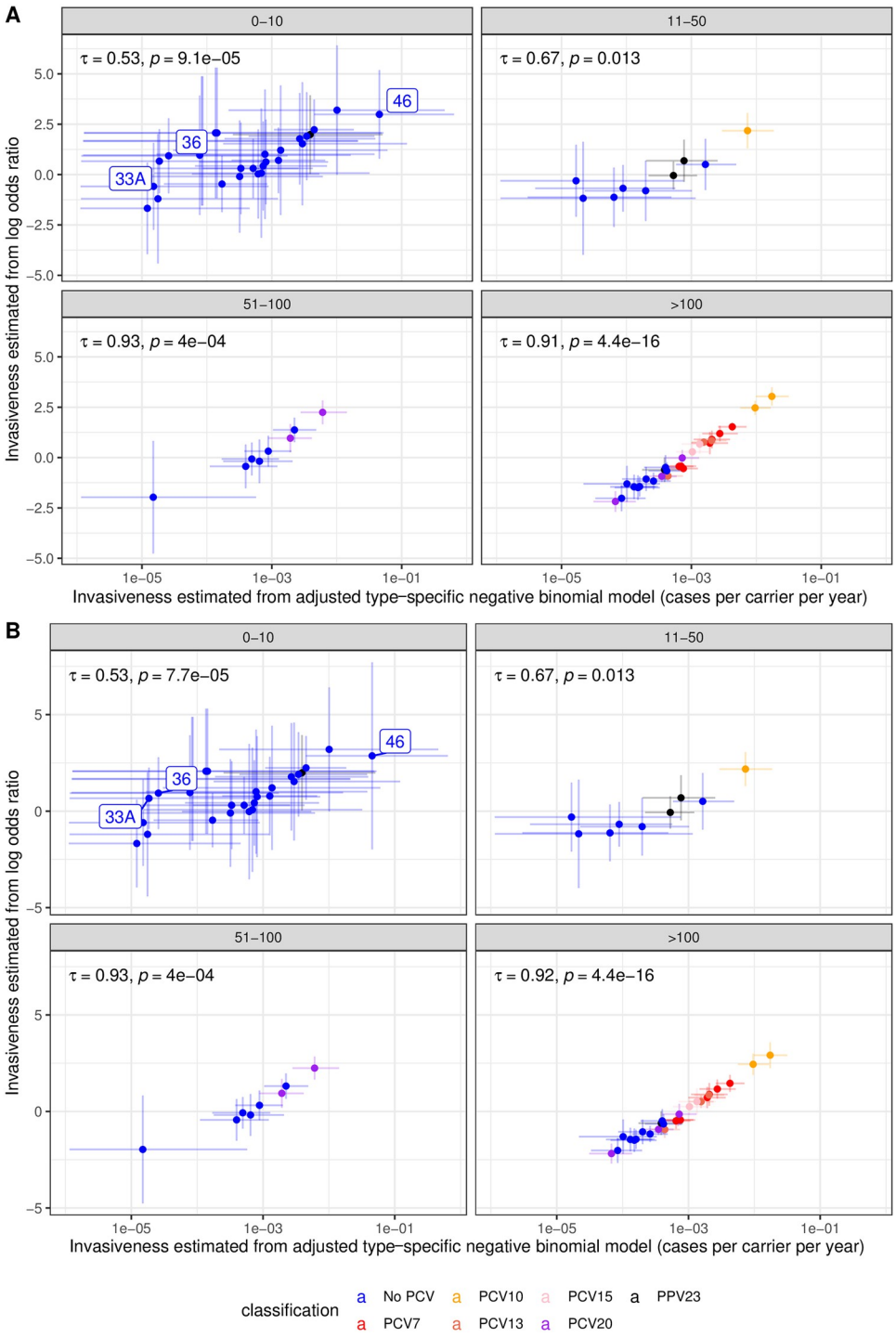

Fig 3. Comparison of the serotype invasiveness estimates from the study-adjusted type-specific negative binomial model with those from meta-analyses of the same dataset calculated as logarithmic odds ratios combined using (A) fixed effects models and (B) random effects models. Points are coloured based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

https://doi.org/10.1371/journal.pcbi.1009389.g003

low Bayesian progression rate estimates (below 10^{-4} cases per carrier per year). Both were observed only in carriage: five isolates across four studies for serotype 33A, and one isolate in each of three studies for serotype 36. Hence the Bayesian model estimates provide a more informative representation of the limited available data.

At the upper end of the scale, the Bayesian model found the most highly invasive serotype to be 46, with credibility intervals indicating this serotype is unlikely to be of intermediate invasiveness [30]. By contrast, the confidence intervals for the odds ratios calculated for serotype 46 were larger relative to the variation between serotypes, with those calculated using random effects spanning the full range of invasiveness point estimates across the species. This serotype was observed in two studies, but never isolated from carriage, and there is further circumstantial evidence that serotype 46 is likely to be highly invasive (see Discussion). Hence this Bayesian framework can use case-carrier studies to provide informative estimates of invasiveness even from small sample sizes.

Odds ratios were previously found to correlate with another measure of 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 [57,83]. 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 (S13 Fig).

Differences in invasiveness between host age demographics

Although *S. pneumoniae* bacteria 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 carriage datasets, to estimate serotype invasiveness in adults. Overall, 3,756 carriage isolates and 3,041 disease isolates were extracted across 53 serotypes (S14 Fig). 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, to improve the power for identifying the best fitting model. This left 3,704 carriage isolates and 2,969 disease isolates across 40 serotypes for this model selection analysis (S15 Fig). As with the child dataset, two 25,000 iteration Hamiltonian MCMCs were used to fit each of the eight models, which converged based on the traces of the posterior probability (S16 Fig) and $\hat{R}$ values being below 1.05 (S17 Fig). 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 (S18 Fig). Bayes factors concurred that the study-adjusted type-specific negative binomial model was the most likely, given the data (S6 Table).

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 (S19 Fig). As with the child analysis, the reference population for the model fit was the post-PCV7 sample from the Navajo nation [39]. The high incidence of IPD in this population is again evident in the adult population, as the other studies had lower study-adjustment scale factors (S20 Fig). Most locations expected to detect 5–50% of 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 and S2 Dataset), including the 24 in vaccine formulations that are currently licensed

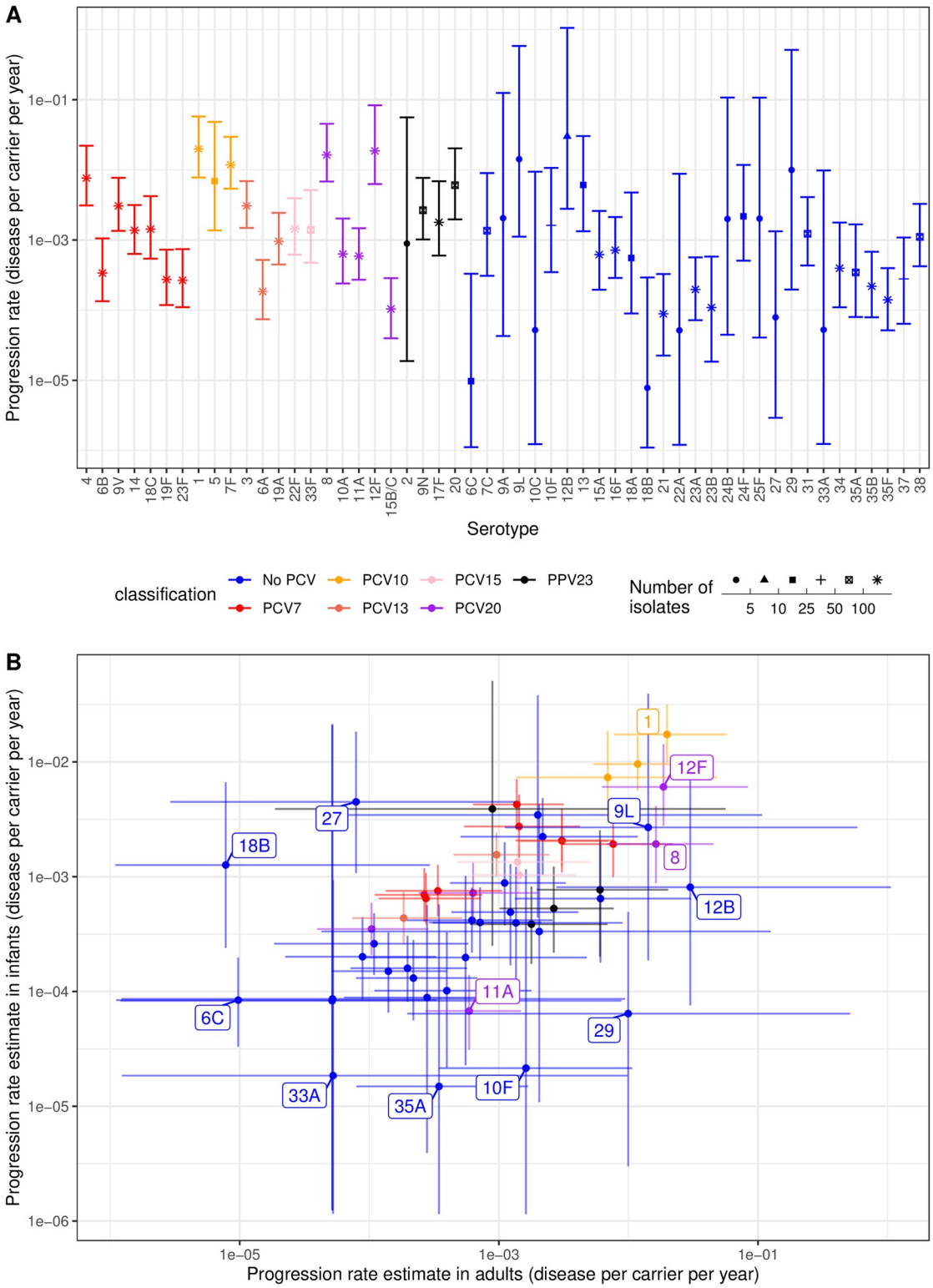

Fig 4. Invasiveness estimates for *S. pneumoniae* serotypes from the study-adjusted type-specific negative binomial model applied to the full dataset of serotype studies from child carriage and adult disease. Points represent the median estimates, and are coloured according to the vaccine formulation 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 serotypes across adults and children.

https://doi.org/10.1371/journal.pcbi.1009389.g004

or under development. The analysis showed the serotypes associated with high invasiveness and narrow credibility intervals not in currently-used PCVs included serotypes 8 and 12F, included in the PCV20 formulation; serotype 20, included in PPV23, and the non-vaccine type serotype 13. Other serotypes not in current vaccine designs, but likely to be highly invasive, included serotypes 9L and 12B. However, the credibility intervals of these estimates were substantial, resulting from small overall sample sizes. The non-PCV serotypes associated with low invasiveness (upper 95% credibility interval estimates $<10^{-3}$ disease cases per carrier per year) were serotypes 6C, 18B, 21, 23A, 23B, 35B and 35F.

Comparisons with the corresponding child invasiveness analysis (Fig 4B) showed many PCV serotypes had similar invasiveness estimates in both age groups [39], with PCV7 and PCV10 types having mid-range and high invasiveness estimates for both children and adults, respectively. The invasiveness of serotype 6C, which is affected by PCV13-induced immunity against the 6A component, was low in both children and adults. The non-PCV serotype 9L appears to be highly invasive in both populations, although this value was associated with considerable uncertainty. However, other non-PCV types showed a discrepancy between the two age groups. For example, serotypes 12B and 29 were estimated to be highly invasive in adults but not children, though the credible interval ranges were large for both.

Serotypes included in the PCV15 and PCV20 formulations appeared to have mid-range invasiveness in both age groups, except for the rarely invasive serotype 11A, and the consistently highly invasive serotypes 8 and 12F. This suggests the removal of these latter two serotypes through herd immunity from higher-valency PCV infant immunisation programmes would likely be beneficial to the adult population [84,85]. However, this would concomitantly reduce the effectiveness of the current PPV23 adult vaccine, due to its overlap in serotype coverage with PCVs meaning it would protect against a lower proportion of the post-vaccine *S. pneumoniae* population [29,86]. PPV23's residual effect would be determined by the prevalence and adult invasiveness of the final non-PCV PPV23 serotypes. Both serotypes 2 and 20 were estimated to be highly invasive for both age groups, albeit with broad credible intervals for adults for the former, due to it only being included in one adult dataset. The remaining non-PCV PPV23 serotypes (9N and 17F) also exhibited elevated invasiveness compared to many non-vaccine types, suggesting that PPV23 may offer protection against a limited number of higher-risk serotypes following the implementation of PCV15 or PCV20 infant vaccination programmes.

Invasiveness varies between pneumococcal strains and serotypes

S. pneumoniae populations are genetically diverse, and can be divided into discrete strains [52,87]. Isolates of a particular serotype are often found across multiple strain backgrounds, and a single strain may be associated with multiple serotypes through switching [53,88]. To test how each of these characteristics affected pneumococcal progression rates, five different models of isolates' invasiveness were fitted using the study-adjusted type-specific model framework (see Methods). These corresponded to the hypotheses that an isolate's invasiveness was determined by its serotype; by its strain background; primarily by its serotype, but modified by strain background; primarily by its strain background, but modified by its serotype; and determined by the combination of its serotype and strain background. All five models were fitted using Poisson and negative binomial distributions.

These ten models were used to conduct a meta-analysis of six studies in which both serotype and strain background could be determined, for at least some isolates (S21 Fig). Three of the studies (post-PCV7 and post-PCV13 South Africa; post-PCV7 USA) were modified versions of a comparison primarily conducted using genomic data [37]. These were combined with pre-PCV genotyped studies from Oxford, UK [31]; Stockholm, Sweden [78], and Finland [74] (see S2 Text and S7 Table). For each dataset i, isolates were grouped by both their serotype j and strain k. As with the child serotype-only analysis, models were fitted to a subset of these samples in which the count of each serotype-strain combination in either carriage or disease in a study ($c_{i,j,k}$ or $d_{i,j,k}$) was at least five. To focus on testing for within-category variation, only strains associated with five different serotypes, and serotypes associated with five different strains, were included in the analysis. The final dataset comprised 11 serotypes, 35 strains, and 46 serotype-strain combinations (S22 Fig). All models converged on stable set of parameter estimates, as inferred from the traces of logarithmic posterior probability values (S23 Fig) and distribution of $\hat{R}$ values (S24 Fig). Few observed frequencies of serotype-strain combinations in either carriage or disease were inconsistent with the 95% credible intervals calculated from any of the fitted models (S25 Fig). There was little evidence of negative binomial distributions improving the fit of these models, and the precision parameters were correspondingly high, albeit with evidence of greater overdispersion when serotype information was omitted from the model fit (S26 Fig). The lowest absolute residuals were associated with the more complex models that accounted for both strain background and serotype (S27 Fig).

Correspondingly, comparisons with LOO-CV found such models that combined genetic and serotype information to be the best-performing (S8 and S9 Tables), although the Pareto k statistic diagnostic values were again too high for this comparison to be reliable. Comparisons using Bayes factors concurred, identifying the most likely model as that in which invasiveness was primarily determined by serotype, but modified by strain, with the count of disease isolates following a Poisson distribution (S10 Table). Some carriage sample sizes had to be approximated in this analysis (for the Oxford and Finland studies in particular; see S2 Text), but the study adjustment factors were robustly estimated and significantly differed across studies, suggesting the model was able to compensate for this aspect of the data (S28 Fig). To test whether these problems could have affected relative model likelihoods, the model comparison was repeated with a 100-fold change in the carriage sample size values for those studies in which the precise number was not reported. The study-adjustment factor meant the comparisons were relatively insensitive to these changes, and the same model was identified as the most likely, given the data (S11 Table). Hence model comparisons demonstrated that neither serotype nor strain background alone determines a genotype's invasiveness.

Identification of invasive pneumococcal genotypes

The best-fitting type-determined strain-modified Poisson model could be reliably fitted to the full dataset (S29 Fig). The invasiveness estimates were plotted to analyse variation by serotype within strains (Fig 5 and S3 Dataset). Some strains expressed serotypes of uniformly low invasiveness (e.g. serotypes 11A, 15A and 20A within GPSC22), whereas others expressed serotypes with consistently higher invasiveness estimates (e.g. serotypes 4 and 19A within GPSC27). By contrast, 19A had an elevated invasiveness relative to 19F within GPSC1, and relative to serotype 15B/C in GPSC4. Similarly, serotypes 8 and 33F were substantially more invasive than 11A within GPSC3. Hence there was considerable variation in estimated invasiveness both between, and within, strains.

Similarly, the factors by which strain backgrounds modified serotypes' invasiveness were plotted, grouped by the serotypes with which they were associated (Fig 6). The meta-analysis

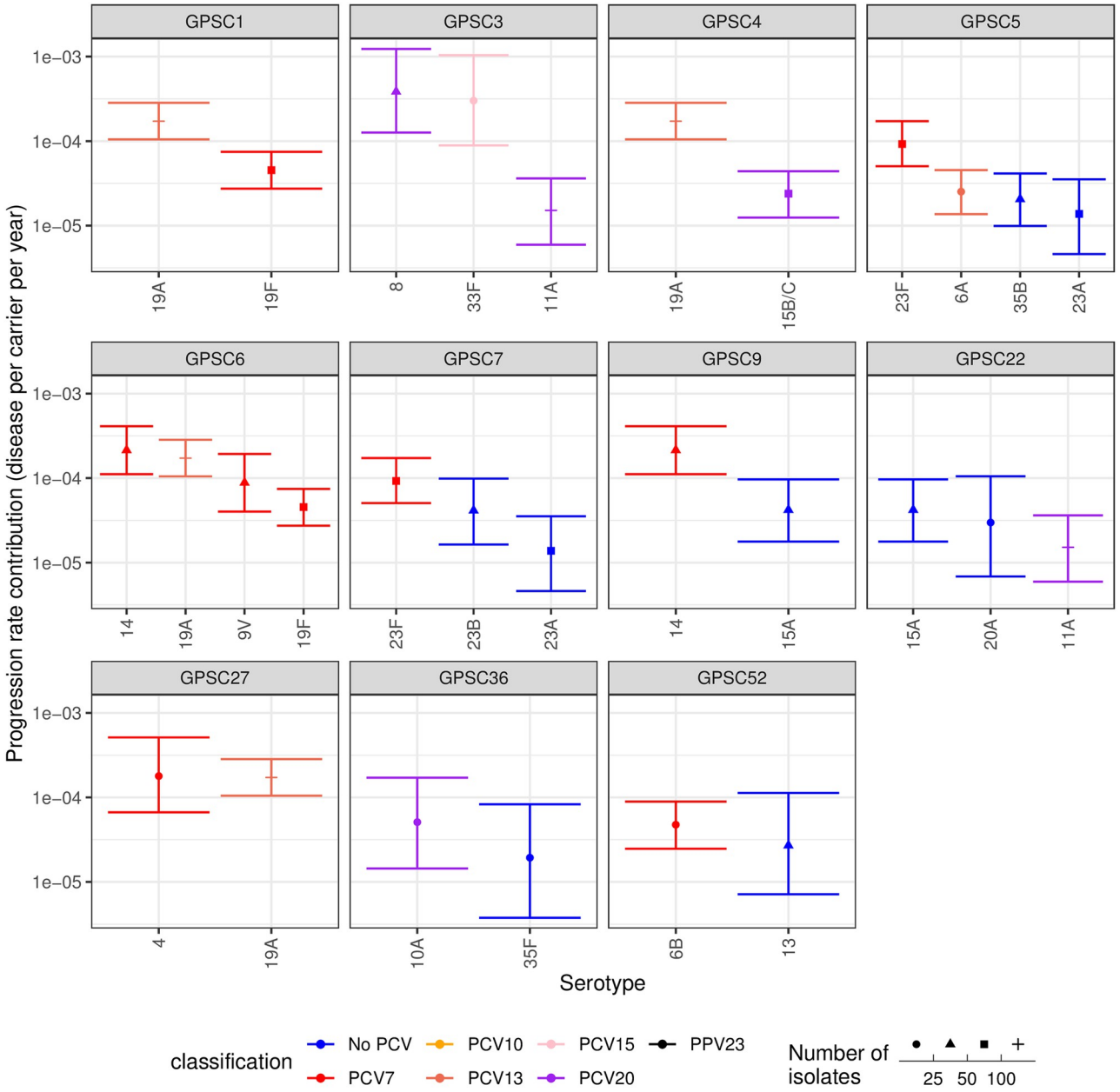

Fig 5. Estimates of the invasiveness associated with serotypes inferred from genotyped isolates using the study-adjusted type-determined strain-modified model. Serotypes are arranged by the strains in which they are found. Data are only shown for strains found in multiple serotype-strain combinations, each represented by at least ten isolates across the studies with genotype information. Points represent the median estimate, and are coloured by the vaccine formulations in which the corresponding serotype is present. The error bars represent the 95% credible intervals. The shape of the point represents the sample size on which the estimate is based.

https://doi.org/10.1371/journal.pcbi.1009389.g005

identified significant evidence to support many of the observations of within-serotype invasiveness variation noted in the individual analyses (S12 Table). These included elevated invasiveness being associated with GPSCs 1, 24 and 41. Heterogeneity was most strongly evident within some of the paediatric serotypes [19], including 6A, 6B, 14, 19F, 19A, 23F and 23A. This may represent these common serotypes being distributed across a wider diversity of strain backgrounds. However, the estimates for other low invasiveness serotypes disseminated across

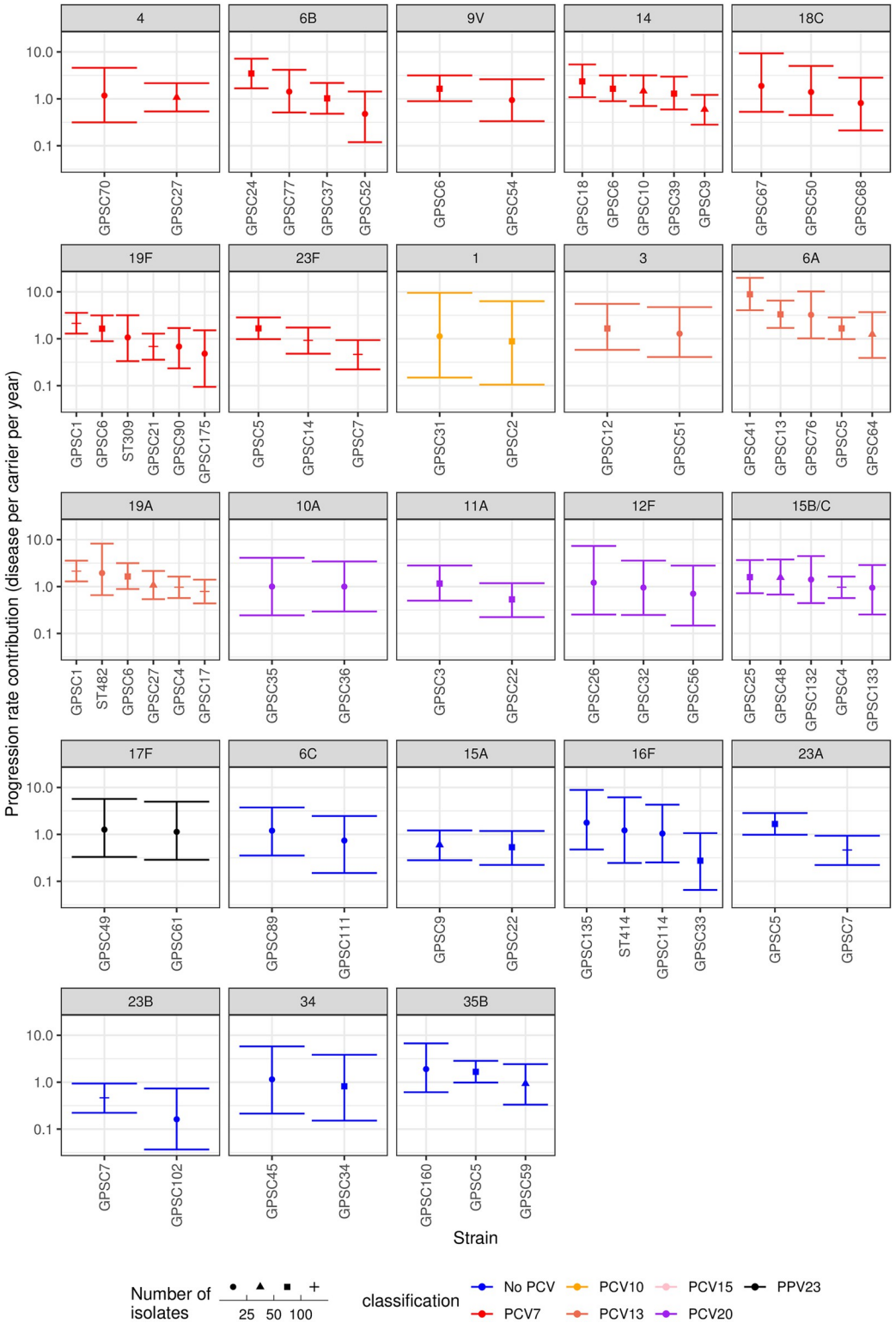

Fig 6. Estimates of the coefficient by which strains modify the invasiveness of their expressed serotype, inferred from genotyped isolates using the study-adjusted type-determined strain-modified model. Strains are arranged by the serotypes with which they are associated. Data are only shown for serotypes found in multiple serotype-strain combinations, each represented by at least ten isolates across the included studies. Points represent the median estimate, and are coloured by the vaccine formulations in which the corresponding serotype is present. The error bars represent the 95% credibililty intervals. The shape of the point represents the sample size on which the estimate is based.

https://doi.org/10.1371/journal.pcbi.1009389.g006

multiple backgrounds, such as 15B/C and 35B, were more consistent. Furthermore, some highly-invasive serotypes (such as 1 and 12F) were similarly invasive in multiple strains. This implies some serotypes may have a stronger effect on an isolate's invasiveness than others.

The overall invasiveness estimate was plotted for each serotype-strain combination in the full dataset, including low frequency types, to identify high-risk genotypes that might emerge as common causes of IPD post-PCV (S30 Fig). The genotypes identified as most concerning in the near-term were typically those expressing serotypes targeted by higher-valency PCVs and regarded as highly invasive, such as 8 and 12F. Yet there were also examples of genotypes with high invasiveness point estimates expressing serotypes not included in any planned PCVs (e.g. 15A, 18B, 25F, 28A, 33A and 33D), albeit these values were inferred from small sample sizes. Most of these genotypes are likely to remain rare, else their invasiveness estimates may regress to the population-wide mean as more data emerge. Ongoing surveillance will be required to identify whether any represent a potentially problematic genotype post-PCV at an early stage in their spread.

An additional genotyped dataset from Portugal (S31 Fig), in which the majority of disease cases were isolated from adults [54], was analysed with the type-determined strain-modified Poisson model (S32 Fig). There were few instances of multiple serotypes being isolated from the same strain, but several serotypes were distributed across multiple genetic backgrounds (S33 Fig). This found evidence of heterogeneity in invasiveness within serotypes 3, 6A and 23F. The elevated invasiveness of serotype 3 was associated with the GPSC6 strain, similarly found to have high invasiveness when expressing multiple serotypes in children (Fig 6). For serotype 6A, this heterogeneity involved strains that did not appear in the child dataset (Fig 6), suggesting further within-serotype variation in invasiveness may emerge as studies include a greater diversity of host age groups.

Discussion

This study describes novel models for using data from cases and carriers to estimate the hazard of an asymptomatically carried pathogen causing a defined disease over a time period, previously defined as a progression rate [2]. This quantity is crucial for determining the efficacy of partial coverage vaccines, such as the PCVs [28,29], and the flexible framework developed here should be applicable to any diverse microbial population. These models calculate progression rates as absolute quantities, rather than dimensionless ratios, as with previous meta-analyses [29,36,38]. These estimates can enable the translation of alterations in microbial population structures to changes in disease incidence, which is crucial for understanding the consequences of strain- or type-specific interventions [29]. Furthermore, given suitable datasets for multiple pathogens, these models could also be used to estimate changes in disease burdens for alterations across microbiota.

When applied to *S. pneumoniae* serotype data, the best-performing model structures were those in which type-specific progression rates were adjusted by dataset-specific scale factors. In principle, these scale factors should correlate with the burden of disease in a given host population, representing a combination of socio-economic factors and co-morbidities that affect IPD incidence [89,90]. Yet in practice, they will also represent the probability of an isolate causing IPD being included in a study dataset. This is affected by multiple factors, including the probability of attending a hospital participating in the study, the efficiency of retrieving bacteria from blood cultures, and the consistency with which cultured isolates are collected by research centres [89].

These model structures facilitate comparisons across populations by standardising progression rate estimates between datasets using all common serotypes or strains shared between

them. Existing methods for meta-analysis typically standardise data relative to a single type, required to be present in all datasets [36,38], or combine estimates derived from ratios calculated relative to variable mixes of types in different locations. The dataset-adjusted type-specific progression rate models are better suited to the extensive international variation in *S. pneumoniae* population structures [37,40], particularly given their ongoing diversification following vaccine introduction [34]. Comparison with odds ratios showed that invasiveness estimates were correlated at large sample sizes, demonstrating both methods converged on similar conclusions when data were available on many isolates. However, this comes with the caveat that 16 of the 21 studies in the meta-analysis originated from North America or Europe, with most collected prior to PCV introduction. Hence there is a pressing need for paired case and carrier studies from post-PCV settings, particularly from Asia and Africa. As more diverse populations are combined, the improved standardisation across studies from these Bayesian analyses will become more important.

At smaller sample sizes, the Bayesian models were able to produce more informative estimates of invasiveness than odds ratios. Information on types that are rare in individual datasets is particularly important for an opportunistic pathogen as diverse as *S. pneumoniae*, in which serotypes can emerge and expand rapidly post-PCV [25,34,84]. A recent meta-analysis using odds ratios highlighted the conclusion that 12F was the only non-PCV13 serotype to have an invasiveness higher than 19A [36], the serotype responsible for much post-PCV7 IPD [34,91]. However, this previous study limited its results to 25 serotypes, omitting less common examples that could rise in frequency when serotype replacement is driven by higher-valency vaccines [34,91]. The Bayesian analysis described in this work identified rare serotypes that may be highly invasive in adults (e.g. 9L, 12B and 29) and children (e.g. 19C, 28A and 46), based on small sample sizes. Serotype 46 has already been observed causing child IPD post-PCV13 when expressed by the strain GPSC26, which is more commonly associated with the highly-invasive serotype 12F [33]. These two capsule types are structurally similar [92], which is common for serotypes expressed by isolates of the same strain [53]. Future post-PCV13 surveillance will likely show whether 12F and 46 are truly similarly invasive. As higher-valency PCVs become available, with the potential to remove commonly-carried serotypes such as 11A and 15B/C [20,93], reducing the uncertainty in invasiveness estimates for remaining non-vaccine types will be of great importance in understanding the vaccine-associated changes in the incidence of IPD.

The need for meta-analysis of multiple datasets is exacerbated by the model comparison in this study that concluded that invasiveness is affected by both strain background and serotype. This conclusion is consistent with the results of some individual studies [37,54,55], although not all [31]. The within-serotype variation in invasiveness caused by differences in isolates' strain background likely explains the superior fit of negative binomial, rather than Poisson, distributions to the serotype-only datasets. Such a result raises the spectre of separately estimating invasiveness for the hundreds of known *S. pneumoniae* strains, many of which are rare in individual populations [37,87], in addition to the over 100 known serotypes. Yet the selected model structure suggests serotype is the primary determinant of invasiveness. Consistent with this conclusion, some serotypes (e.g. 1, 12F) exhibited high invasiveness across multiple strain backgrounds, whereas others were consistently low (e.g. 15B/C, 35F).

By contrast, within some paediatric serotypes (6A, 6B, 23F and 23A), there was strong evidence of strains with distinct invasiveness estimates (defined as estimates with non-overlapping 95% credibility intervals). Weaker evidence for within-serotype variation in progression rates (identified through median point estimates for one strain being outside the 95% credibility intervals of another) was observed in four further paediatric serotypes: 14, 19F, 19A and 23B. This may reflect paediatric serotypes having little effect on invasiveness, or their greater

prevalence in the pre-PCV pneumococcal population providing greater power to detect variation between strains [32]. Heterogeneity has previously been noted in serotype 14, with both GPSC6 (corresponding to PMEN3, or Spain^{9V}-3 [88]) [54,74] and GPSC18 [31,37] identified as highly invasive clones (S12 Table). In this analysis, both were estimated to be more invasive than GPSC9 within the same serotype. This could result in heterogeneous vaccine impacts between regions, as elimination of serotype 14 by PCVs will likely have a greater benefit across Europe and South America, where it is often associated with the more invasive GPSC6 (https://microreact.org/project/gpsGPSC6), rather than Africa and India, where it is often associated with the less invasive GPSC9 (https://microreact.org/project/gpsGPSC9).

Whether these models are useful long-term, or will ultimately be superseded by alternatives based on genome-wide association studies, will depend on the extent to which progression rates are affected by interactions between genetic loci [94]. It may prove possible to explain a high proportion of the variation in invasiveness between microbes through a tractable number of polymorphic loci that each independently contribute to a microbe's invasiveness [47–50]. If so, then the individual estimates for strains can be replaced with a simpler genome-based approach, which would be capable of predicting the invasiveness of previously unseen genotypes. However, if invasiveness depends on combinations of loci [94], it may only be possible to estimate progression rates for extant common types from epidemiological studies. Other alternative model structures include random effects models and multivariate regression analyses, which can incorporate information on the host population characteristics [2,39,95]. However, these models typically require comprehensive surveillance of disease and detailed information on host populations, which limits the range of studies that can be combined in meta-analyses of diverse bacterial populations.

Even using more conventional approaches to estimating invasiveness, whole genome sequencing clearly represents an important tool in future case-carrier studies of opportunistic pathogen invasiveness. In addition to assigning isolates to strains, genomic data is a reliable and cost-effective means of classifying microbes according to their capsular structures [96,97]. Furthermore, it also enables the extraction of additional clinically-relevant information, such as antibiotic resistance loci [88,98], that may be incorporated into future extensions of these models. However, the use of genomic data exacerbates an underlying problem with these models, as high-quality genome assemblies are not necessarily generated from all isolates recovered from carriage. The model assumes there are no false negative swabs in carriage studies, meaning there is an expectation that all carriage samples will be present in the genomic data. Hence the omission of genomic data adds to a false negative rate that, even with serotype data, already reflects the imperfect detection of colonisation by nasopharyngeal swabbing. This can be the consequence of all resident bacteria being missed, or the difficulty of detecting lower prevalence types in instances of multiple carriage [99]. Unless there is variation in the ability to detect carriage of different types, the underestimation of carriage means all progression rates will be uniformly overestimated. However, if only a subset of datasets in a meta-analysis has a lower sensitivity for detecting carriage, the model will adjust for this by increasing the associated scale factors. Hence the model structure is sufficiently flexible to account for such differences. This is necessary, as even with standardised protocols for detecting colonisation, the population-wide estimates for carriage rates for a location will depend on the exact demographics sampled.

Some of these problems with detecting colonisation can be addressed using new techniques with improved sensitivity for detecting multiple serotypes in carriage [100]. However, to simultaneously obtain the genotyping information needed to precisely characterise an isolate's invasiveness using such mixed samples requires deep genome sequencing, or similar high-sensitivity molecular genotyping techniques. Such data will not only improve the input to

epidemiological models, such as those outlined here, but also potential future models based on combinations of genetic loci. Yet meta-analyses of case-carrier studies can also be improved by simple changes to reporting, without any alterations to current procedures. The data and models described in this work are made available for re-use, modification, and application to other multi-strain microbes (see Methods). The fitting of the model to new datasets will require studies to report their raw data in a standard format, and include the total number of swabs included in the carriage study, as well as an informed estimate of the size of the population under surveillance for IPD. To improve comparability between future studies, it would also be ideal for studies to stratify these data by age category, to enable specific subsets to be employed in different meta-analyses studying specific demographics. Therefore, as the repertoire of vaccines against diverse pathogens expands, we can hope to monitor and understand their impact on population-wide microbial invasiveness using improved models, advances in sequencing technology, and greater transparency of reporting.

Supporting information

S1 Text. Validation of model fits using simulated data.
(DOCX)

S2 Text. Sample descriptions for genomic and genotyped datasets.
(DOCX)

S1 Fig. Properties of the serotype data from child carriage and disease. (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.
(PNG)

S2 Fig. Plots of estimated progression rate values from model fits to simulated data. Each panel shows the fit of eight different models, indicated by colour, to data simulated from a single model structure. Points represent the median estimate, and the error bars show the 95% credible intervals. The models from which data were simulated were (A) null Poisson, (B) null negative binomial, (C) type-specific Poisson, (D) type-specific negative binomial, (E) study-adjusted Poisson, (F) study-adjusted negative binomial, (G) study-adjusted type-specific Poisson, and (H) study-adjusted type-specific negative binomial. For panels (A), (B), (E) and (F), all types had the same progression rate, and therefore the dashed horizontal line represents the single true value used in the simulations. For panels (B), (C), (G) and (H), each of the 20 types had a different progression rate, which determines the horizontal position of the point on the graph. In these four plots, the dashed line indicates the line of identity.
(PNG)

S3 Fig. Plot of the estimated values of the precision parameter ϕ from model fits to simulated data. The horizontal position of the point corresponds to the model that was both used to generate the data, and to then infer the parameter value from these data. The vertical position of the point indicates the median estimate of the parameter, and the error bars represent the 95% credibility interval. The horizontal dashed line indicates the true value of the parameter used to simulate data, $\phi = 0.1$.
(PNG)

S4 Fig. Properties of the filtered serotype data from child carriage and disease, after removing datapoints corresponding to observations of fewer than five isolates of a serotype from both carriage and disease in a given study. (A) Stacked bar plot showing the distri-

bution of carriage and disease isolates between studies. (B) Stacked bar plot showing the distribution of carriage and disease isolates between serotypes.
(PNG)

S5 Fig. 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.
(PNG)

S6 Fig. Histograms of $\hat{R}$ 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.
(PNG)

S7 Fig. Violin plots showing the absolute deviations between observed and predicted values across model fits to the filtered serotype data from child carriage and disease. Blue crosses represent the individual observations.
(PNG)

S8 Fig. 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 median estimates from the MCMCs, and the error bars show the 95% credibility intervals.
(PNG)

S9 Fig. Plots validating the fit of the study-adjusted type-specific negative binomial model to the full serotype data from child carriage and disease. (A) Histogram showing the distribution of $\hat{R}$ values. (B) Post-warmup MCMC traces of the log posterior probability.
(PNG)

S10 Fig. Study-adjustment scale factors from the study-adjusted type-specific negative binomial 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 (S1 Fig).
(PNG)

S11 Fig. Plots validating the fit of the study-adjusted type-specific negative binomial 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 $\hat{R}$ values. (B) Post-warmup MCMC traces of the log posterior probability.
(PNG)

S12 Fig. Study-adjustment scale factors from the study-adjusted type-specific negative binomial 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 (S1 Fig).
(PNG)

S13 Fig. 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 assigned a value of zero days.
(PNG)

S14 Fig. Properties of the serotype data from child carriage and adult disease. (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.
(PNG)

S15 Fig. Properties of the filtered serotype data from child carriage and adult disease after removing datapoints corresponding to observations of fewer than five isolates of a serotype 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.
(PNG)

S16 Fig. 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.
(PNG)

S17 Fig. Histograms of $\hat{R}$ 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.
(PNG)

S18 Fig. Comparison of observed and predicted counts of each serotype within each study of adult disease relative to carriage in children. The plots in the left column display data on carriage in children (values of $c_{i,j}$). The plots in the right column display data on disease in adults (values of $d_{i,j}$). The points are coloured by the study to which they correspond, and represent the observed value on the horizontal axis, and the median 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. 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.
(PNG)

S19 Fig. Plots validating the fit of the study-adjusted type-specific negative binomial model to the full serotype data from child carriage and adult disease. (A) Histogram showing the distribution of $\hat{R}$ values. (B) Post-warmup MCMC traces of the log posterior probability.
(PNG)

S20 Fig. Study-adjustment scale factors from the study-adjusted type-specific negative binomial 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 (S15 Fig).
(PNG)

S21 Fig. Properties of the strain and serotype data from child carriage and disease. (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.
(PNG)

S22 Fig. Properties of the filtered strain and serotype data from child carriage and disease after removing datapoints corresponding to observations of fewer than five isolates of a serotype-strain combination from both carriage and disease in a given study, then retaining only serotypes associated with at least five different strains, and strains associated with at least five different serotypes. (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.
(PNG)

S23 Fig. 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 binomially-distributed invasiveness; (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.
(PNG)

S24 Fig. Histograms of $\hat{R}$ 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 binomially-distributed invasiveness; (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.
(PNG)

S25 Fig. 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 median predicted values 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 binomially-distributed invasiveness; (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.
(PNG)

S26 Fig. 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 median estimates from the MCMCs, and the error bars show the 95% credibility interval.
(PNG)

S27 Fig. Violin plots showing the absolute deviations 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.
(PNG)

S28 Fig. Study-adjustment scale factors from the study-adjusted type-determined strain-modified Poisson model fitted to the full serotype and strain data from child carriage and disease. 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 (S2 Text).
(PNG)

S29 Fig. Plots validating the fit of the study-adjusted type-determined strain-modified Poisson model to the full strain and serotype data from child carriage and disease. (A) Histogram showing the distribution of $\hat{R}$ values. (B) Post-warmup MCMC traces of the log posterior probability.
(PNG)

S30 Fig. Point estimates of invasiveness for all strain and serotype combinations in the full dataset from child carriage and disease. The shape of each point represents the number of isolates of each combination observed across carriage and disease samples. The serotype-strain

combinations with the highest combined invasiveness that are not targeted by current PCV designs are labelled.
(PNG)

S31 Fig. Bar plot showing the distribution of carriage and disease isolates between serotypes for the dataset from Portugal comparing child carriage with primarily adult disease.
(PNG)

S32 Fig. Plots validating the fit of the type-determined strain-modified Poisson model to the strain and serotype data from child carriage and primarily adult disease. (A) Histogram showing the distribution of $\hat{R}$ values. (B) Post-warmup MCMC traces of the log posterior probability.
(PNG)

S33 Fig. Invasiveness estimates for strain and serotype combinations represented by at least ten isolates in the study of child carriage and primarily adult disease. Points represent the median estimates, and are coloured by the vaccine formulations in which the corresponding serotype is present. The error bars represent 95% credibility intervals. The shape of the point represents the sample size on which the estimate is based. (A) Estimates of the invasiveness associated with serotypes, arranged by the strains in which they are found. Only strains expressing multiple serotypes are displayed. (B) Estimates of the coefficient by which strains modify the invasiveness of their expressed serotype, arranged by the serotypes with which they are associated. Only serotypes associated with multiple strains are displayed.
(PNG)

S1 Table. Study populations used for the analysis of serotype invasiveness in children and adults.
(DOCX)

S2 Table. Model fit comparison with leave-one-out cross-validation (LOO-CV) using the serotype data from child 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.
(DOCX)

S3 Table. Model fit comparison with LOO-CV using the serotype data from child carriage and disease using a subset of likelihood values. 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.
(DOCX)

S4 Table. Comparison of model fits to simulated data using bridge sampling. Each row corresponds to data simulated from the specified model. All eight models were fitted to each set of simulated data. The table shows the models adjudged to be the first and second best-fitting to each dataset, using bridge sampling. The final column shows the logarithm of the Bayes factor by which the best-fitting model was favoured over the second best-fitting model.
(DOCX)

S5 Table. 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 were calculated for each model relative to that which was found to be the most likely given the data; hence the value is zero for the first row.
(DOCX)

S6 Table. 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 S5.
(DOCX)

S7 Table. Study populations used for the 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.
(DOCX)

S8 Table. Model fit comparison with LOO-CV 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.
(DOCX)

S9 Table. Model fit comparison with LOO-CV using the child strain and serotype data on carriage and disease using a subset of likelihood values. 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.
(DOCX)

S10 Table. Comparison of model fits to strain and serotype data from child carriage and disease using Bayes factors calculated with bridge sampling. The table is displayed as described for Table S5.
(DOCX)

S11 Table. Comparison of model fits to strain and serotype data from child carriage and disease with modified carriage sample sizes 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 the sensitivity of the model comparisons to the uncertainty in this parameter (Text S2). The table is displayed as described for Table S5.
(DOCX)

S12 Table. Comparison of previously-identified within-serotype differences in strain invasiveness with the results of 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 differences between strains were not tested in this analysis.
(DOCX)

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

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

S3 Dataset. Invasiveness estimates for serotypes and strains, calculated from genotyped isolates from child disease and carriage. Each row corresponds to a serotype-strain combination. Both the invasiveness associated with the serotype, and the invasiveness coefficient associated with the strain, are listed, along with the corresponding lower and upper bounds of their 95% credibility intervals.
(XLSX)

Acknowledgments

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.

Author Contributions

Conceptualization: Alessandra Løchen, James E. Truscott, Nicholas J. Croucher.

Data curation: Alessandra Løchen, Nicholas J. Croucher.

Formal analysis: Alessandra Løchen, James E. Truscott, Nicholas J. Croucher.

Funding acquisition: Nicholas J. Croucher.

Investigation: Alessandra Løchen, Nicholas J. Croucher.

Methodology: Alessandra Løchen, James E. Truscott, Nicholas J. Croucher.

Project administration: Nicholas J. Croucher.

Software: Nicholas J. Croucher.

Supervision: James E. Truscott, Nicholas J. Croucher.

Validation: Nicholas J. Croucher.

Visualization: Alessandra Løchen, Nicholas J. Croucher.

Writing – original draft: Alessandra Løchen, Nicholas J. Croucher.

Writing – review & editing: Alessandra Løchen, James E. Truscott, Nicholas J. Croucher.

References

1. Brown SP, Cornforth DM, Mideo N. Evolution of virulence in opportunistic pathogens: generalism, plasticity, and control. Trends Microbiol. 2012; 20: 336–342. https://doi.org/10.1016/j.tim.2012.04.005 PMID: 22564248
2. Lewnard JA, Givon-Lavi N, Weinberger DM, Lipsitch M, Dagan R. Pan-serotype reduction in progression of *Streptococcus pneumoniae* to otitis media after rollout of pneumococcal conjugate vaccines. Clin Infect Dis. 2017; 65: 1853–1861. https://doi.org/10.1093/cid/cix673 PMID: 29020218
3. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet. 2012; 13: 601–612. https://doi.org/10.1038/nrg3226 PMID: 22868263
4. Feil EJ, Enright MC. Analyses of clonality and the evolution of bacterial pathogens. Curr Opin Microbiol. 2004; 7: 308–313. https://doi.org/10.1016/j.mib.2004.04.002 PMID: 15196500

5. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo S, Weiser JN, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res. 2019; 29: 304–316. https://doi.org/10.1101/gr.241455.118 PMID: 30679308

6. Cartman ST, Heap JT, Kuehne SA, Cockayne A, Minton NP. The emergence of "hypervirulence" in *Clostridium difficile*. Int J Med Microbiol. 2010; 300: 387–395. https://doi.org/10.1016/j.ijmm.2010.04.008 PMID: 20547099

7. Fearnley C, Manning G, Bagnall M, Javed MA, Wassenaar TM, Newell DG. Identification of hyperinvasive *Campylobacter jejuni* strains isolated from poultry and human clinical sources. J Med Microbiol. 2008; 57: 570–580. https://doi.org/10.1099/jmm.0.47803-0 PMID: 18436589

8. Knol MJ, Hahné SJM, Lucidarme J, Campbell H, de Melker HE, Gray SJ, et al. Temporal associations between national outbreaks of meningococcal serogroup W and C disease in the Netherlands and England: an observational cohort study. Lancet Public Heal. 2017; 2: e473–e482. https://doi.org/10.1016/S2468-2667(17)30157-3

9. Lynskey NN, Jauneikaite E, Li HK, Zhi X, Turner CE, Mosavie M, et al. Emergence of dominant toxigenic M1T1 *Streptococcus pyogenes* clone during increased scarlet fever activity in England: a population-based molecular epidemiological study. Lancet Infect Dis. 2019; 19: 1209–1218. https://doi.org/10.1016/S1473-3099(19)30446-3 PMID: 31519541

10. Du P, Zhang Y, Chen C. Emergence of carbapenem-resistant hypervirulent *Klebsiella pneumoniae*. Lancet Infect Dis. 2018; 18: 23–24. https://doi.org/10.1016/S1473-3099(17)30625-4 PMID: 29102520

11. Ganaie F, Saad JS, McGee L, van Tonder AJ, Bentley SD, Lo SW, et al. A new pneumococcal capsule type, 10D, is the 100th serotype and has a large *cps* fragment from an oral streptococcus. MBio. 2020; 11: e00937–20. https://doi.org/10.1128/mBio.00937-20 PMID: 32430472

12. Ganaie F, Maruhn K, Li C, Porambo RJ, Elverdal PL, Abeygunwardana C, et al. Structural, genetic, and serological elucidation of *Streptococcus pneumoniae* serogroup 24 serotypes: Discovery of a new serotype, 24C, with a variable capsule structure. J Clin Microbiol. 2021; JCM.00540-21. https://doi.org/10.1128/JCM.00540-21 PMID: 33883183

13. Geno KA, Gilbert GL, Song JY, Skovsted IC, Klugman KP, Jones C, et al. Pneumococcal capsules and their types: Past, present, and future. Clin Microbiol Rev. 2015; 28: 871–899. https://doi.org/10.1128/CMR.00024-15 PMID: 26085553

14. Nelson AL, Roche AM, Gould JM, Chim K, Ratner AJ, Weiser JN. Capsule enhances pneumococcal colonization by limiting mucus-mediated clearance. Infect Immun. 2007; 75: 83–90. https://doi.org/10.1128/IAI.01475-06 PMID: 17088346

15. Hyams C, Camberlein E, Cohen JM, Bax K, Brown JS. The *Streptococcus pneumoniae* capsule inhibits complement activity and neutrophil phagocytosis by multiple mechanisms. Infect Immun. 2009/12/02. 2010; 78: 704–715. IAI.00881-09 [pii] https://doi.org/10.1128/IAI.00881-09 [doi] PMID: 19948837

16. Barker J, Gratten M, Riley M, Lehmann D, Montgomery J, Kajoi M, et al. Pneumonia in children in the eastern highlands of Papua New Guinea: A bacteriologic study of patients selected by standard clinical criteria. J Infect Dis. 1989; 159: 348–352. https://doi.org/10.1093/infdis/159.2.348 PMID: 2783717

17. Webster LT, Clow AD. Intranasal virulence of pneumococci for mice. J Exp Med Med. 1933; 58: 465–483. https://doi.org/10.1084/jem.58.4.465 PMID: 19870209

18. Sleeman KL, Griffiths D, Shackley F, Diggle L, Gupta S, Maiden MC, et al. Capsular serotype-specific attack rates and duration of carriage of *Streptococcus pneumoniae* in a population of children. J Infect Dis. 2006; 194: 682–688. https://doi.org/10.1086/505710 PMID: 16897668

19. Hausdorff WP, Feikin DR, Klugman KP. Epidemiological differences among pneumococcal serotypes. Lancet Infect Dis. 2005; 5: 83–93. https://doi.org/10.1016/S1473-3099(05)01280-6 PMID: 15680778

20. Croucher NJ, Løchen A, Bentley SD. Pneumococcal Vaccines: Host Interactions, Population Dynamics, and Design Principles. Annu Rev Microbiol. 2018; 72: 521–549. https://doi.org/10.1146/annurev-micro-090817-062338 PMID: 30200849

21. Stacey HL, Rosen J, Peterson JT, Williams-Diaz A, Gakhar V, Sterling TM, et al. Safety and immunogenicity of 15-valent pneumococcal conjugate vaccine (PCV-15) compared to PCV-13 in healthy older adults. Hum Vaccines Immunother. 2019; 15: 530–539. https://doi.org/10.1080/21645515.2018.1532249 PMID: 30648919

22. Hurley D, Griffin C, Young M, Scott DA, Pride MW, Scully IL, et al. Safety, Tolerability, and Immunogenicity of a 20-Valent Pneumococcal Conjugate Vaccine (PCV20) in Adults 60 to 64 Years of Age. Clin Infect Dis. 2020; ciaa1045. https://doi.org/10.1093/cid/ciaa1045 PMID: 32716500

23. Miller E, Andrews NJ, Waight PA, Slack MPE, George RC. Herd immunity and serotype replacement 4 years after seven-valent pneumococcal conjugate vaccination in England and Wales: An observational cohort study. Lancet Infect Dis. 2011; 11: 760–68. https://doi.org/10.1016/S1473-3099(11)70090-1 PMID: 21621466

24. Weinberger DM, Malley R, Lipsitch M. Serotype replacement in disease after pneumococcal vaccination. The Lancet. 2011. pp. 1962–1973. https://doi.org/10.1016/S0140-6736(10)62225-8 PMID: 21492929

25. Lewnard JA, Hanage WP. Making sense of differences in pneumococcal serotype replacement. Lancet Infect Dis. 2019. https://doi.org/10.1016/S1473-3099(18)30660-1 PMID: 30709666

26. Devine VT, Cleary DW, Jefferies JMC, Anderson R, Morris DE, Tuck AC, et al. The rise and fall of pneumococcal serotypes carried in the PCV era. Vaccine. 2017; 35: 1293–1298. https://doi.org/10.1016/j.vaccine.2017.01.035 PMID: 28161425

27. Southern J, Andrews N, Sandu P, Sheppard CL, Waight PA, Fry NK, et al. Pneumococcal carriage in children and their household contacts six years after introduction of the 13-valent pneumococcal conjugate vaccine in England. PLoS One. 2018; 13: e0195799. https://doi.org/10.1371/journal.pone.0195799 PMID: 29799839

28. Nurhonen M, Auranen K. Optimal Serotype Compositions for Pneumococcal Conjugate Vaccination under Serotype Replacement. PLoS Comput Biol. 2014; 10: e1003477. https://doi.org/10.1371/journal.pcbi.1003477 PMID: 24550722

29. Colijn C, Corander J, Croucher NJ. Designing ecologically optimized pneumococcal vaccines using population genomics. Nat Microbiol. 2020; 5: 473–485. https://doi.org/10.1038/s41564-019-0651-y PMID: 32015499

30. Smith T, Lehmann D, Montgomery J, Gratten M, Riley ID, Alpers MP. Acquisition and invasiveness of different serotypes of *Streptococcus pneumoniae* in young children. Epidemiol Infect. 1993; 111: 27–39. https://doi.org/10.1017/s0950268800056648 PMID: 8348930

31. Brueggemann AB, Griffiths DT, Peto T, Meats E, Crook DW, Spratt BG. Clonal Relationships between Invasive and Carriage *Streptococcus pneumoniae* and Serotype- and Clone-Specific Differences in Invasive Disease Potential. J Infect Dis. 2003; 187: 1424–32. https://doi.org/10.1086/374624 PMID: 12717624

32. Johnson HL, Deloria-Knoll M, Levine OS, Stoszek SK, Hance LF, Reithinger R, et al. Systematic evaluation of serotypes causing invasive pneumococcal disease among children under five: The pneumococcal global serotype project. PLoS Med. 2010; 7: e1000348. https://doi.org/10.1371/journal.pmed.1000348 PMID: 20957191

33. Lo SW, Gladstone RA, van Tonder AJ, Lees JA, du Plessis M, Benisty R, et al. Pneumococcal lineages associated with serotype replacement and antibiotic resistance in childhood invasive pneumococcal disease in the post-PCV13 era: an international whole-genome sequencing study. Lancet Infect Dis. 2019; 19: 759–769. https://doi.org/10.1016/S1473-3099(19)30297-X PMID: 31196809

34. Løchen A, Croucher NJ, Anderson RM. Divergent serotype replacement trends and increasing diversity in pneumococcal disease in high income settings reduce the benefit of expanding vaccine valency. Sci Rep. 2020; 10: 18977. https://doi.org/10.1038/s41598-020-75691-5 PMID: 33149149

35. Hanage WP, Auranen K, Syrjanen R, Herva E, Makela PH, Kilpi T, et al. Ability of pneumococcal serotypes and clones to cause acute otitis media: implications for the prevention of otitis media by conjugate vaccines. Infect Immun. 2004; 72: 76–81. https://doi.org/10.1128/IAI.72.1.76-81.2004 PMID: 14688083

36. Balsells E, Dagan R, Yildirim I, Gounder PP, Steens A, Muñoz-Almagro C, et al. The relative invasive disease potential of *Streptococcus pneumoniae* among children after PCV introduction: A systematic review and meta-analysis. J Infect. 2018; 77: 368–378. https://doi.org/10.1016/j.jinf.2018.06.004 PMID: 29964140

37. Gladstone RA, Lo SW, Lees JA, Croucher NJ, van Tonder AJ, Corander J, et al. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. EBioMedicine. 2019; 43: 338–346. https://doi.org/10.1016/j.ebiom.2019.04.021 PMID: 31003929

38. Brueggemann AB, Peto TEA, Crook DW, Butler JC, Kristinsson KG, Spratt BG. Temporal and Geographic Stability of the Serogroup-Specific Invasive Disease Potential of *Streptococcus pneumoniae* in Children. J Infect Dis. 2004/09/04. 2004; 190: 1203–1211. https://doi.org/10.1086/423820 PMID: 15346329

39. Weinberger DM, Grant LR, Weatherholtz RC, Warren JL, O'Brien KL, Hammitt LL. Relating Pneumococcal Carriage among Children to Disease Rates among Adults before and after the Introduction of Conjugate Vaccines. Am J Epidemiol. 2016; 183: 1055–62. https://doi.org/10.1093/aje/kwv283 PMID: 27188949

40. Corander J, Fraser C, Gutmann MU, Arnold B, Hanage WP, Bentley SD, et al. Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. Nat Ecol Evol. 2017; 1: 1950–1960. https://doi.org/10.1038/s41559-017-0337-x PMID: 29038424

41. Weinberger DM, Bruden DT, Grant LR, Lipsitch M, O'Brien KL, Pelton SI, et al. Using pneumococcal carriage data to monitor postvaccination changes in invasive disease. Am J Epidemiol. 2013; 178: 1488–1495. https://doi.org/10.1093/aje/kwt156 PMID: 24013204

42. Zwahlen A, Winkelstein JA, Moxon ER. Surface Determinants of *Haemophilus influenzae* Pathogenicity: Comparative Virulence of Capsular Transformants in Normal and Complement-Depleted Rats. J Infect Dis. 1983; 148: 385–394. https://doi.org/10.1093/infdis/148.3.385 PMID: 6604759

43. Kelly T, Dillard JP, Yother J. Effect of genetic switching of capsular type on virulence of *Streptococcus pneumoniae*. Infect Immun. 1994; 62: 1813–1819. Available: https://iai.asm.org/content/iai/62/5/1813.full.pdf https://doi.org/10.1128/iai.62.5.1813-1819.1994 PMID: 8168944

44. Hathaway LJ, Grandgirard D, Valente LG, Täuber MG, Leib SL. *Streptococcus pneumoniae* capsule determines disease severity in experimental pneumococcal meningitis. Open Biol. 2016; 6: 150269. https://doi.org/10.1098/rsob.150269 PMID: 27009189

45. Hu FZ, Eutsey R, Ahmed A, Frazao N, Powell E, Hiller NL, et al. In vivo capsular switch in *Streptococcus pneumoniae*—Analysis by whole genome sequencing. PLoS One. 2012; 7: e47983–e47983. https://doi.org/10.1371/journal.pone.0047983 PMID: 23144841

46. Mizrachi Nebenzahl Y, Porat N, Lifshitz S, Novick S, Levi A, Ling E, et al. Virulence of *Streptococcus pneumoniae* may be determined independently of capsular polysaccharide. FEMS Microbiol Lett. 2004; 233: 147–152. https://doi.org/10.1016/j.femsle.2004.02.003 PMID: 15043881

47. Obert C, Sublett J, Kaushal D, Hinojosa E, Barton T, Tuomanen EI, et al. Identification of a candidate *Streptococcus pneumoniae* core genome and regions of diversity correlated with invasive pneumococcal disease. Infect Immun. 2006; 74: 4766–4777. https://doi.org/10.1128/IAI.00316-06 PMID: 16861665

48. Blomberg C, Dagerhamn J, Dahlberg S, Browall S, Fernebro J, Albiger B, et al. Pattern of Accessory Regions and Invasive Disease Potential in *Streptococcus pneumoniae*. J Infect Dis. 2009; 199: 1032–1042. https://doi.org/10.1086/597205 PMID: 19203261

49. Lees JA, Ferwerda B, Kremer PHC, Wheeler NE, Serón MV, Croucher NJ, et al. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. Nat Commun. 2019; 10: 2176. https://doi.org/10.1038/s41467-019-09976-3 PMID: 31092817

50. Cremers AJ, Mobegi F, van der Gaast–de Jongh C, van Weert M, van Opzeeland F, Vehkala M, et al. The contribution of genetic variation of *Streptococcus pneumoniae* to the clinical manifestation of invasive pneumococcal disease. Clin Infect Dis. 2019; 68: 61–69. https://doi.org/10.1093/cid/ciy417 PMID: 29788414

51. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli S V, et al. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol. 2010; 11: R107–R107. https://doi.org/10.1186/gb-2010-11-10-r107 PMID: 21034474

52. Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD, Hanage WP. Diversification of bacterial genome content through distinct mechanisms over different timescales. Nat Commun. 2014; 5: 5471. https://doi.org/10.1038/ncomms6471 PMID: 25407023

53. Croucher NJ, Kagedan L, Thompson CM, Parkhill J, Bentley SD, Finkelstein JA, et al. Selective and Genetic Constraints on Pneumococcal Serotype Switching. PLoS Genet. 2015; 11: e1005095. https://doi.org/10.1371/journal.pgen.1005095 PMID: 25826208

54. Sá-Leao R, Pinto F, Aguiar S, Nunes S, Carriço JAJA, Frazao N, et al. Analysis of invasiveness of pneumococcal serotypes and clones circulating in Portugal before widespread use of conjugate vaccines reveals heterogeneous behavior of clones expressing the same serotype. J Clin Microbiol. 2011; 49: 1369–75. https://doi.org/10.1128/JCM.01763-10 PMID: 21270219

55. Browall S, Norman M, Tångrot J, Galanis I, Sjöström K, Dagerhamn J, et al. Intraclonal Variations Among *Streptococcus pneumoniae* Isolates Influence the Likelihood of Invasive Disease in Children. J Infect Dis. 2014; 209: 377–388. https://doi.org/10.1093/infdis/jit481 PMID: 24009156

56. Fraser C, Hanage WP, Spratt BG. Neutral microepidemic evolution of bacterial pathogens. Proc Natl Acad Sci U S A. 2005; 102: 1968–1973. https://doi.org/10.1073/pnas.0406993102 PMID: 15684071

57. Lees JA, Croucher NJ, Goldblatt D, Nosten F, Parkhill J, Turner C, et al. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. Elife. 2017; 6: e26255. https://doi.org/10.7554/eLife.26255 PMID: 28742023

58. Stan Development Team. Stan User's Guide, Version 2.27. In: Interaction Flow Modeling Language [Internet]. 2020 p. 23.7. https://mc-stan.org/docs/2_27/stan-users-guide/reparameterization-section.html

59. Wahl B, O'Brien KL, Greenbaum A, Majumder A, Liu L, Chu Y, et al. Burden of *Streptococcus pneumoniae* and *Haemophilus influenzae* type b disease in children in the era of conjugate vaccines: global, regional, and national estimates for 2000–15. Lancet Glob Heal. 2018; 6: E744–E757. https://doi.org/10.1016/S2214-109X(18)30247-X PMID: 29903376

60. R Core Team. R: A Language and Environment for Statistical Computing. R Found Stat Comput. 2019.

61. Stan Development Team. RStan: the R interface to Stan. 2020.

62. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput. 2017; 27: 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

63. Gronau QF, Singmann H, Wagenmakers EJ. Bridgesampling: An R package for estimating normalizing constants. J Stat Softw. 2020; 92: 1–29. https://doi.org/10.18637/jss.v092.i10

64. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. J Stat Softw. 2015. https://doi.org/10.18637/jss.v036.i03

65. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 2019; 4: 1686. https://doi.org/10.21105/joss.01686

66. Kassambara A. ggpubr: "ggplot2" Based Publication Ready Plots. 2020. https://cran.r-project.org/package=ggpubr

67. Valles X, Roca A, Lozano F, Morais L, Suarez B, Casals F, et al. Serotype-specific pneumococcal disease may be influenced by mannose-binding lectin deficiency. Eur Respir J. 2010; 36: 856–863. https://doi.org/10.1183/09031936.00171409 PMID: 20150204

68. Principi N, Marchisio P. Epidemiology of *Streptococcus pneumoniae* in Italian children. Acta Paediatr Suppl. 2000; 89: 40–43. https://doi.org/10.1111/j.1651-2227.2000.tb00782.x PMID: 11194797

69. Smart LE, Platt DJ, Timbury MC. A comparison of the distribution of pneumococcal types in systemic disease and the upper respiratory tract in adults and children. Epidemiol Infect. 1987; 98: 203–209. https://doi.org/10.1017/s0950268800061926 PMID: 3556447

70. Mogdasy MC, Camou T, Fajardo C, Hortal M. Colonizing and invasive strains of *Streptococcus pneumoniae* in Uruguayan children: type distribution and patterns of antibiotic resistance. Pediatr Infect Dis J. 1992; 11: 648–652. PMID: 1523077

71. Scott J, Hall A, Hannington A, Edwards R, Mwarumba S, Lowe B, et al. Serotype distribution and prevalence of resistance to benzylpenicillin in three representative populations of *Streptococcus pneumoniae* isolates from the coast of Kenya. Clin Infect Dis. 1998; 27: 1442–50. https://doi.org/10.1086/515013 PMID: 9868658

72. Levidiotou S, Vrioni G, Tzanakaki G, Pappa C, Gesouli H, Gartzonika C, et al. Serotype distribution of *Streptococcus pneumoniae* in north-western Greece and implications for a vaccination programme. FEMS Immunol Med Microbiol. 2006; 48: 179–82. https://doi.org/10.1111/j.1574-695X.2006.00126.x PMID: 17064274

73. Azzari C, Cortimiglia M, Nieddu F, Moriondo M, Indolfi G, Mattei R, et al. Pneumococcal serotype distribution in adults with invasive disease and in carrier children in Italy: Should we expect herd protection of adults through infants' vaccination? Hum Vaccin Immunother. 2016; 12: 344–350. https://doi.org/10.1080/21645515.2015.1102811 PMID: 26647277

74. Hanage WP, Kaijalainen TH, Syrjänen RK, Auranen K, Leinonen M, Mäkelä PH, et al. Invasiveness of serotypes and clones of *Streptococcus pneumoniae* among children in Finland. Infect Immun. 2005; 73: 431–5. https://doi.org/10.1128/IAI.73.1.431-435.2005 PMID: 15618181

75. Cullotta AR, Kalter HD, Delgado J, Gilman RH, Facklam RR, Velapatino B, et al. Antimicrobial susceptibilities and serotype distribution of *Streptococcus pneumoniae* isolates from a Low socioeconomic area in Lima, Peru. Clin Diagn Lab Immunol. 2002; 9: 1328–1331. https://doi.org/10.1128/cdli.9.6.1328-1331.2002 PMID: 12414769

76. Takala AK, Vuopio-Varkila J, Tarkka E, Leinonen M, Musser JM. Subtyping of common pediatric pneumococcal serotypes from invasive disease and pharyngeal carriage in Finland. J Infect Dis. 1996; 173: 128–135. https://doi.org/10.1093/infdis/173.1.128 PMID: 8537649

77. In HP, Pritchard DG, Cartee R, Brandao A, Brandileone MCC, Nahm MH. Discovery of a new capsular serotype (6C) within serogroup 6 of *Streptococcus pneumoniae*. J Clin Microbiol. 2007; 45: 1225–1233. https://doi.org/10.1128/JCM.02199-06 PMID: 17267625

78. Browall S, Backhaus E, Naucler P, Galanis I, Sjöström K, Karlsson D, et al. Clinical manifestations of invasive pneumococcal disease by vaccine and non-vaccine types. Eur Respir J. 2014; 44: 1646–57. https://doi.org/10.1183/09031936.00080814 PMID: 25323223

79. Vehtari A, Simpson D, Gelman A, Yao Y, Gabry J. Pareto Smoothed Importance Sampling. arXiv. 2021; 1507.02646v7.

80. O'Brien KL, Shaw J, Weatherholtz R, Reid R, Watt J, Croll J, et al. Epidemiology of invasive *Streptococcus pneumoniae* among Navajo children in the era before use of conjugate pneumococcal vaccines, 1989–1996. Am J Epidemiol. 2004; 160: 270–8. https://doi.org/10.1093/aje/kwh191 PMID: 15258000

81. Ladhani SN, Collins S, Djennad A, Sheppard CL, Borrow R, Fry NK, et al. Rapid increase in non-vaccine serotypes causing invasive pneumococcal disease in England and Wales, 2000–17: a prospective national observational cohort study. Lancet Infect Dis. 2018; 18: 441–451. https://doi.org/10.1016/S1473-3099(18)30052-5 PMID: 29395999

82. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, et al. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol. 2009; 24: 127–135. https://doi.org/10.1016/j.tree.2008.10.008 PMID: 19185386

83. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014; 46: 305–309. https://doi.org/10.1038/ng.2895 PMID: 24509479

84. Feikin DR, Kagucia EW, Loo JD, Link-Gelles R, Puhan MA, Cherian T, et al. Serotype-Specific Changes in Invasive Pneumococcal Disease after Pneumococcal Conjugate Vaccine Introduction: A Pooled Analysis of Multiple Surveillance Sites. PLoS Med. 2013; 10: e1001517. https://doi.org/10.1371/journal.pmed.1001517 PMID: 24086113

85. Shiri T, Datta S, Madan J, Tsertsvadze A, Royle P, Keeling MJ, et al. Indirect effects of childhood pneumococcal conjugate vaccination on invasive pneumococcal disease: a systematic review and meta-analysis. Lancet Glob Heal. 2017; 5: e51–e59. https://doi.org/10.1016/S2214-109X(16)30306-0 PMID: 27955789

86. José RJ, Brown JS. Adult pneumococcal vaccination: advances, impact, and unmet needs. Curr Opin Pulm Med. 2017; 23.

87. Harrow GL, Lees JA, Hanage WP, Lipsitch M, Corander J, Colijn C, et al. Negative frequency-dependent selection and asymmetrical transformation stabilise multi-strain bacterial population structures. ISME J. 2021; 15: 1523–1538. https://doi.org/10.1038/s41396-020-00867-w PMID: 33408365

88. D'Aeth JC, van der Linden MPG, McGee L, de Lencastre H, Turner P, Song J-H, et al. The role of interspecies recombination in the evolution of antibiotic-resistant pneumococci. Elife. 2021; 10: e67113. Available: https://doi.org/10.7554/eLife.67113 PMID: 34259624

89. Fedson DS, Anthony J, Scott G. The burden of pneumococcal disease among adults in developed and developing countries: what is and is not known. Vaccine. 1999; 17: S11–S18. https://doi.org/10.1016/S0264-410X%2899%2900122-X

90. Torres A, Blasi F, Dartois N, Akova M. Which individuals are at increased risk of pneumococcal disease and why? Impact of COPD, asthma, smoking, diabetes, and/or chronic heart disease on community-acquired pneumonia and invasive pneumococcal disease. Thorax. 2015; 70: 984–989. https://doi.org/10.1136/thoraxjnl-2015-206780 PMID: 26219979

91. Kyaw MH, Lynfield R, Schaffner W, Craig AS, Hadler J, Reingold A, et al. Effect of Introduction of the Pneumococcal Conjugate Vaccine on Drug-Resistant *Streptococcus pneumoniae*. N Engl J Med. 2006; 354: 1455–1463. https://doi.org/10.1056/NEJMoa051642 PMID: 16598044

92. Mavroidi A, Aanensen DM, Godoy D, Skovsted IC, Kaltoft MS, Reeves PR, et al. Genetic relatedness of the *Streptococcus pneumoniae* capsular biosynthetic loci. J Bacteriol. 2007; 189: 7841–7855. https://doi.org/10.1128/JB.00836-07 PMID: 17766424

93. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet. 2013; 45: 656–663. https://doi.org/10.1038/ng.2625 PMID: 23644493

94. Skwark MJ, Croucher NJ, Puranen S, Chewapreecha C, Pesonen M, Xu YY, et al. Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. PLoS Genet. 2017; 13: e1006508. https://doi.org/10.1371/journal.pgen.1006508 PMID: 28207813

95. Lewnard JA, Givon-Lavi N, Tähtinen PA, Dagan R. Pneumococcal Phenotype and Interaction with Nontypeable *Haemophilus influenzae* as Determinants of Otitis Media Progression. Infect Immun. 2018; 86: e00727–17. https://doi.org/10.1128/IAI.00727-17 PMID: 29378791

96. Epping L, van Tonder AJ, Gladstone RA, Bentley SD, Page AJ, Keane JA. SeroBA: rapid high-throughput serotyping of *Streptococcus pneumoniae* from whole genome sequence data. Microb Genomics. 2018; 4: https://doi.org/10.1099/mgen.0.000186 PMID: 29870330

97. Kapatai G, Sheppard CL, Al-Shahib A, Litt DJ, Underwood AP, Harrison TG, et al. Whole genome sequencing of *Streptococcus pneumoniae*: development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline. PeerJ. 2016; 4: e2477. https://doi.org/10.7717/peerj.2477 PMID: 27672516

98. Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE, Walker H, et al. Penicillin-Binding Protein Transpeptidase Signatures for Tracking and Predicting β-Lactam Resistance Levels in *Streptococcus pneumoniae*. MBio. 2016; 7: e00756–16. https://doi.org/10.1128/mBio.00756-16 PMID: 27302760

99. Huebner RE, Dagan R, Porath N, Wasas AD, T M, Klugman KP. Lack of utility of serotyping multiple colonies for detection of simultaneous nasopharyngeal carriage of different pneumococcal serotypes.

Pediatr Infect Dis J. 2000; 19: 1017–1020. https://doi.org/10.1097/00006454-200010000-00019 PMID: 11055610

100. Turner P, Hinds J, Turner C, Jankhot A, Gould K, Bentley SD, et al. Improved detection of nasopharyngeal cocolonization by multiple pneumococcal serotypes by use of latex agglutination or molecular serotyping by microarray. J Clin Microbiol. 2011; 49: 1784–1789. https://doi.org/10.1128/JCM.00157-11 PMID: 21411589