Variation in pneumococcal invasiveness metrics is driven by serotype carriage duration and initial risk of disease

Benjamin J. Metcalf[a,b,c,*], **Kristofer Wollein Waldetoft**[a,b,d], **Bernard W. Beall**[c], **Sam P. Brown**[a,b,*]

[a]School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia

[b]Center for Microbial Dynamics and Infection, Georgia Institute of Technology, Atlanta, Georgia

[c]Centers for Disease Control and Prevention, Atlanta, Georgia

[d]Torsby Hospital, Torsby, Sweden

Abstract

Streptococcus pneumoniae is an opportunistic pathogen that, while usually carried asymptomatically, can cause severe invasive diseases like meningitis and bacteremic pneumonia. A central goal in *S. pneumoniae* public health management is to identify which serotypes (immunologically distinct strains) pose the most risk of invasive disease. The most common invasiveness metrics use cross-sectional data (*i.e.*, invasive odds ratios (IOR)), or longitudinal data (*i.e.*, attack rates (AR)). To assess the reliability of these metrics we developed an epidemiological model of carriage and invasive disease. Our mathematical analyses illustrate qualitative failures with the IOR metric (*e.g.*, IOR can decline with increasing invasiveness parameters). Fitting the model to both longitudinal and cross-sectional data, our analysis supports previous work indicating that invasion risk is maximal at or near time of colonization. This pattern of early invasive disease risk leads to substantial (up to 5-fold) biases when estimating underlying differences in invasiveness from IOR metrics, due to the impact of carriage duration on IOR. Together, these results raise serious concerns with the IOR metric as a basis for public health decision-making and lend support for multiple alternate metrics including AR.

Author summary

Streptococcus pneumoniae (the pneumococcus) is an opportunistic pathogen comprised of immunologically distinct serotypes that can cause severe invasive disease. Thus, reliable

[*]Corresponding authors at: School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia. ycm6@cdc.gov (B.J. Metcalf), sam.brown@biology.gatech.edu (S.P. Brown).

Declaration of Competing Interest
None.

CRediT authorship contribution statement
Benjamin J. Metcalf: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing. **Kristofer Wollein Waldetoft:** Conceptualization, Methodology, Investigation, Formal analysis. Bernard W. Beall: Investigation, Formal analysis. **Sam P. Brown:** Conceptualization, Methodology: Investigation, Formal analysis, Writing – original draft, Writing – review & editing.

Appendix A. Supporting information
Supplementary data associated with this article can be found in the online version at doi:10.1016/j.epidem.2023.100731.

metrics of serotype invasiveness are essential for research and infection management. We show mathematically that the primary invasiveness metric (the invasive odds ratio, IOR) is a logically flawed metric, as IOR can decrease given increases in intrinsic invasiveness of a serotype. Fitting our models to longitudinal and cross-sectional epidemiological data we show that invasion risk is maximal at or near time of colonization and IOR is confounded by rapid invasion and variation in serotype carriage duration. This confounding can lead to substantial (up to 5-fold) biases when estimating underlying differences in invasiveness from IOR metrics. In contrast, we find that an Attack Rate (AR) metric calculated from longitudinal data and alternate cross-sectional metrics do not suffer from these limitations. Our analysis suggests that alternate invasiveness metrics, including AR, may be more appropriate for pneumococcal public health management and invasive disease research.

Keywords

Streptococcus pneumoniae; Invasive disease; Virulence; Compartmental model; Carriage duration

1. Introduction

Streptococcus pneumoniae (Spn) is a gram positive opportunistic pathogen and a major cause of childhood bacterial pneumonia, meningitis, and sepsis globally (O'Brien et al., 2009). Although an important human pathogen, in most cases it colonizes the nasopharynx asymptomatically (referred to as carriage). Symptomatic *S. pneumoniae* usually presents as non-invasive diseases like otitis media and non-bacteremic pneumonia, but, on rare occasions, it can also spread into normally sterile sites and cause severe infections known as invasive pneumococcal disease (IPD).

Epidemiological studies of pneumococcal disease are typically organized by serotype, with each serotype distinguished by an immunologically distinct polysaccharide capsule. To date over 90 unique serotypes have been identified, although only about 40 are known to commonly cause pneumococcal disease. The propensity of an individual serotype to cause invasive disease can be measured from longitudinal data, by estimating the ratio of IPD incidence to carriage acquisition rate. This quantity is known as the attack rate (AR) and varies across serotypes from zero (for serotypes not known to cause IPD) to 75 per 100,000 carriage acquisitions (serotypes 1 and 5, (Sleeman et al., 2006)). Unfortunately, obtaining serotype-specific AR metrics of invasiveness is challenging, due to the requirement of detailed longitudinal studies needed to measure carriage acquisition rates (Sleeman et al., 2006).

Due to the challenges of acquiring longitudinal data, an alternate metric of invasiveness based on cross-sectional data is more widely used (Supplement) (Løchen et al., 2022). The Invasive Odds Ratio (IOR) is calculated from cross-sectional prevalences of carriage and invasive disease and has demonstrated a positive correlation with capsular-specific AR (Sleeman et al., 2006). Specifically, IOR is the number of invasive disease cases over the number of carriage cases for a specific serotype, referenced against either a particular serotype or all other serotypes. To illustrate, if a is the number of invasive cases for the focal serotype, b is the number of carriage cases for the focal serotype, c is the number of invasive cases for the reference, and d is the number of carriage cases for the reference, then $IOR = (ad)/(bc)$. Using serotype 14 as a common reference, Brueggemann *et al.* estimated that the high AR strains 1 and 5 also have high IOR (4.5 and 6.0 respectively, *i.e.*, 4–6 times more invasive than serotype 14) (Brueggemann et al., 2004).

The observed variation in invasiveness (indicated by both IOR and AR) across strains poses the critical question of mechanism: Why are some strains more likely to produce invasive disease? Much of the research that has sought an explanation has focused, quite naturally, on variation in virulence factors (molecular determinants of disease, (Allen et al., 2014)) across Spn serotypes. Some studies have taken a targeted experimental approach, for instance Hyams *et al.* identify serotype-specific interactions with innate immune components that associate with variation in AR invasiveness (Hyams et al., 2013). Other studies have taken a more global, genome-wide association study (GWAS) approach comparing IPD cases against carriage controls that have also found genetic factors outside of serotype playing a role in predicting invasiveness (Lees et al., 2019). However, it’s important to note that the GWAS analysis still found 50% of the variation in invasiveness was attributable to serotype.

In addition to variation in invasiveness with serotype, a growing literature points to variation in the risk of invasive disease with time since colonization. For example, a longitudinal study by Grey *et al.* found that 74% of pneumococcal infections (both invasive and non-invasive) occurred within the first time point (one month) after acquisition (Gray et al., 1980). Indirect evidence of a rapid IPD progression following acquisition was also documented in epidemiological studies that observed annual spikes in adult IPD between December 24 and January 7 which closely overlapped with winter holiday festivities (Dowell et al., 2003; Walter et al., 2009). Finally, work on SPN transmission clusters found that IPD outbreaks are marked by short timespans and high genetic relatedness between connected infections, indicating a limited amount of time to transmit and accumulate within-host genetic diversity (Metcalf et al., 2021).

While the work of Hyams *et al.* (Hyams et al., 2013) and Lees et al. (Lees et al., 2019) highlight the potential importance of virulence factors in driving Spn invasion, they don’t fully explain the variation in invasiveness across serotypes. One simple alternate ecological hypothesis is that the variation in invasive metrics is driven by variation in the duration of serotype carriage. In the case of IOR the logic is simple; serotypes vary in carriage duration (Sleeman et al., 2006), and increased carriage duration will (all else being equal) increase carriage prevalence, therefore reducing IOR. We use an epidemiological modeling approach to mathematically map how different epidemiological assumptions on the timing of invasive disease translate into differing relationships between carriage duration and invasive disease metrics, while fixing the ‘virulence factor’ parameters to be constant across serotypes. We then use existing longitudinal and cross-sectional datasets (Brueggemann et al., 2004; Sleeman et al., 2006) to parameterize our alternate model structures, supporting prior experimental and observational research indicating that invasive disease risk is maximal during the initial stages of colonization (Domínguez-Hüttinger et al., 2017; Dowell et al., 2003; Gray et al., 1980; Walter et al., 2009). Together, our model and data analysis highlight the limits of widely used IOR metrics to identify differences in strain virulence and support the conclusion that invasive risk is greatest during the initial stages of carriage.

2. Results

2.1. Analyzing the impact of invasive disease timing on metrics of invasive disease (IOR and AR)

The susceptible-carrier-invasive-recovered (SCIR) compartmental model used in the following analyses is presented in Fig. 1 and described in greater detail in Materials and Methods (defined by Eq. 1 and Table 1). We begin by addressing from first principles whether IOR and AR metrics can serve as effective measures of the underlying invasive disease processes that are governed by parameters p (probability of transition to invasive state on initial acquisition) and d (rate of invasive disease progression from carriage state). Given the assumption that the system of ordinary differential equations (ODEs) (Eq. 1) reaches an endemic infection equilibrium, we can derive simple analytical expressions for IOR and AR (see supplementary for derivation), namely that

$$\mathrm{IOR} = \frac{\tau_0(d\tau + p)}{\tau(d\tau_0 + p)},$$

$$\mathrm{AR} = \frac{d\tau + p}{(1 - p)(d\tau + 1)}. \tag{2}$$

Both invasive disease metrics center on calculating ratios of epidemiological quantities, which leads to the canceling out of many parameters in the full dynamical system (Eq. 1), leaving only the invasive disease parameters (d, p) and measures of carriage duration (serotype specific duration τ and reference serotype duration τ_0). To assess whether AR and IOR serve as effective metrics of underlying potential disease parameters d and p, we next assess whether AR and IOR are increasing functions of d and p (see Supplementary Information (SI)) (Fig. 2). The AR metric passes this test, as AR is positively associated with both p and d (Fig. 2A; the gradients $\mathrm{AR'}(p)$ and $\mathrm{AR'}(d)$ are always positive functions). In contrast, IOR fails in an either/or manner. If a long carriage duration serotype is used as the reference (*i.e.*, $\tau_0 > \tau$), then IOR becomes negatively associated with d (Fig. 2B; $\mathrm{IOR'}(d) < 0$). In contrast if $\tau_0 < \tau$, then IOR becomes negatively associated with p (Fig. 2C; $\mathrm{IOR'}(p) < 0$).

The continued presence of carriage duration parameters in both AR and IOR metrics points to additional problems with a reliance on AR or IOR as a tool to estimate intrinsic invasive disease risk, as changes in these metrics could follow solely from changes in carriage duration τ across strains (Fig. 3). Indeed, prior studies report a clear negative relationship between serotype IOR and carriage C (Brueggemann et al., 2004), and a marginal negative relationship between AR and τ (Sleeman et al., 2006). We next ask what values of d and p are consistent with these qualitative patterns? In the case of IOR, we find that a negative relationship between IOR and τ (*i.e.* the gradient $\mathrm{IOR}'(\tau) < 0$) is possible if and only if $p > 0$ (SI for details). In the case of AR, we find that a negative relationship (*i.e.* $\mathrm{AR}'(\tau) < 0$) is not possible for any combination of d and p. The gradient $\mathrm{AR}'(\tau)$ is minimized at zero if and only if $d = 0$ (Fig. 3B, SI for details). Together these analytical considerations lend support for a model where invasive disease risk is associated with initial strain acquisition only (*i.e.*, $p > 0$, $d = 0$). Given the additional assumption that p is small (<< 0.5), $\mathrm{AR} = p/(1 - p)$ becomes an effective metric, capturing the underlying key biological disease process p, without contamination from other epidemiological parameters including τ. In contrast, IOR now simplifies to $\tau_0 < \tau$ and so bears no relationship to p and is entirely defined by carriage durations.

In this section we have used qualitative properties in existing epidemiological data to guide our model assumptions. We next use this existing data in a model fitting approach to further assess whether our support for early invasive disease risk ($p > 0$, $d = 0$) is warranted.

2.2. Fitting the SCIR compartmental model to the Spn carriage and invasive disease data

The most direct model fitting approach is to fit the expressions for AR and IOR (Eq. 2 above) to existing AR and IOR data for serotypes with defined carriage durations (Fig. 3). For the IOR calculations, following Brueggemann *et al.* (Brueggemann et al., 2004). we used serotype 14 as the reference strain (therefore defining $\tau_0 = 14$ weeks). Fitting the IOR and AR expressions simultaneously (see methods) yields parameter estimates of $p = 2.9 \times 10^{-4}$ (95%CI: 1.3–4.6 × 10^{-4}) and $d = 0.0$ (CI: 0–1.4 *10^{-5}), lending support for the qualitative conclusion above that invasive disease risk is associated with strain acquisition ($p > 0$) and not with ongoing carriage ($d = 0$). These results allow simplification of the IOR and AR relationships to carriage duration τ to $\mathrm{IOR} = \tau_0/\tau$, and $\mathrm{AR} = p/(1 - p)$, see fitted lines in Fig. 3. While the best fit model found no influence of carriage duration on attack rate (Fig. 3B) there does appear to be a negative trend when comparing all serotypes catalogued in the Sleeman *et al.* (Sleeman et al., 2006) longitudinal data (Spearman rho = − 0.33, p = 0.085) which may reflect variation in invasiveness not attributable to life-history traits.

To harness additional epidemiological data provided in the work by Sleeman *et al.* (Sleeman et al., 2006) and Brueggemann *et al.* (Brueggemann et al., 2004) (specifically carriage prevalence, carriage acquisition rate and invasive incidence), we next fit the endemic equilibrium state of the entire epidemiological model (Fig. 1; endemic equilibrium equations are defined in SI). In this approach the parameter estimates will not only be informed by the AR and IOR data but also how parameters effect other aspects of Spn disease and transmission. Specifically, we now simultaneously fit the endemic equilibrium solution of the SCIR model with AR and IOR data, plus both the incidence and prevalence data. Because the longitudinal and cross-sectional data does not include length of invasive infection, we fix the invasive clearance rate h at 0.5/week based on previous work by Baldo *et al.* (Baldo et al., 2015). For IOR calculations, we again use serotype 14 as the reference strain. In agreement with our analyses above, we estimate $p = 2.9 \times 10^{-4}$ (95%CI: 1.3–4.5 × 10^{-4}) and $d = 0.0$ (CI: 0–1.4 × 10^{-5}). In addition, we now simultaneously estimate transmission ($\beta = 0.24$ (CI: 0.06–0.41; in line with previous work (Domenech de Cellès et al., 2011; Melegaro et al., 2004)) and immunity loss rate ($f = 3.1 \times 10^{-3}$ (CI: 1.9 × 10^{-4} to 5.9 × 10^{-3}). The Brueggemann *et al.* (Brueggemann et al., 2004) paper used carriage prevalence instead of carriage duration, but a similar relationship holds when prevalence is substituted for duration (Fig. S1).

Expanding the data to incorporate additional epidemiological measures did not change our model fits to AR or IOR data (still captured by Fig. 3, fitted to AR and IOR data only). In Fig. 4 we assess the full dataset fitted model against the additional epidemiological data on carriage prevalence, carriage acquisition rate and invasive incidence. Together, Figs. 3 and 4 illustrate that while we capture the overall pneumococcal transmission dynamics by minimizing model fitting error across multiple datasets, we can see biases in the model fit to individual datasets. In particular, the fitted model tends to overshoot the invasive incidence relationship over longer carriage durations (Fig. 4C). This is also reflected in the attack rate fit shown in Fig. 3B.

A key assumption built into this model design is that the duration of carriage follows an exponential distribution with a constant clearance rate. Given that this is a decreasing probability density function, the mode is 0 which means, in this context, that most infections will be cleared immediately. To rule out the possibility that our results concerning rapid invasion are not simply artifacts of the model design, we show that a separate two-stage carriage model which generates a non-zero mode of duration yields the same result as the SCIR model with a constant clearance rate (Supplement).

As a final check on our model inference, we fit two variations of our SCIR model representing either an initial risk only ($d = 0$; the model supported above) or constant risk only ($p = 0$) progression to invasive disease. In agreement with our model fitting conclusion that $d = 0$ (Fig. 4), an information criterion model comparison approach concludes that the initial risk model ($d = 0$) outperforms the constant risk model (Table S2).

2.2.1. Alternate invasiveness metrics—Our model fitting analyses all agree that the risk of invasive disease is front-loaded, with support for $p > 0$ and $d = 0$. Given $d = 0$, we earlier noted that IOR simplifies to τ_0/τ, and therefore bears no relationship to the underlying invasiveness process, *p*. In contrast, the AR metric (derived from longitudinal data) simplifies to $p/(1 - p)$, preserving information on *p* (given $p < < 0.5$). We next ask, are the problems with IOR a common feature of using cross-sectional *versus* longitudinal data? We find that this is not the case – a simple ratio of invasive to carriage cases (I/C) performs better than IOR and captures information on p (Given $d = 0,\ I/C = \frac{p}{h\tau - hp\tau}$). We Note that I/C also suffers from contamination by τ and h, and can be expressed as $I/C = AR/(h\tau)$. For comparisons among strains with well-characterized carriage duration, one path to approach AR more closely from cross-sectional data would be *via* a ‘corrected I/C’ $= h\tau I/C = AR$. Similarly, invasive capacity (IC), defined by Yildirim *et al.* as the ratio of invasive incidence to carriage prevalence, suitably represents p (Given $d = 0,\ \mathrm{IC} = \frac{p}{\tau - p\tau}$) and can also be corrected for bias by multiplying by τ (Yildirim et al., 2010).

3. Discussion

Quantifying serotype invasiveness is centrally important to the public health management of pneumococcal disease. At present, IOR is a commonly used invasiveness metric, as it is easily calculated from cross-sectional epidemiological data (Song et al., 2013). To assess the reliability of IOR (and AR) we developed an epidemiological model of invasive disease progression, separating early colonization and ongoing risk processes. Our mathematical analysis demonstrates that IOR is prone to qualitative failures (*e.g.*, IOR declining with increasing invasiveness parameters, Fig. 2B, C). Fitting this model to both longitudinal and cross-sectional data (Fig. 3), we found that variation in IOR can be, in large part, explained by variation in the life history trait of carriage duration, and thus that, contrary to common assumption, variation in IOR does not necessarily imply an underlying variation in molecular determinants of the ability to cause invasive disease.

Our best fit model supports previous observations that progression to invasive disease occurs at or near the time of carriage acquisition ($p > 0,\ d = 0$). This constraint on the timing of invasion generates an inverse correlation between invasive odds ratio IOR and carriage duration (τ) defined by the reciprocal function $\mathrm{IOR} = \frac{\tau_0}{\tau}$. While a negative association between IOR and τ has been reported, it has been argued that a relatively small 3-fold difference in carriage duration could not account for the 60-fold variation in IOR (Brueggemann et al., 2003; Brueggemann et al., 2004). Our analysis illustrates that small changes in carriage duration τ can, in fact, have large impacts on IOR due to the reciprocal relationship between τ and IOR and may partially explain this variation in IOR. In contrast, the fitted model defined attack rate AR to be independent of carriage duration $\mathrm{AR} = \frac{p}{1-p}$). Together these results agree with the empirical finding of a significant correlation between IOR and carriage duration (Fig. 3A), but no correlation between AR and carriage duration (Fig. 3B, although, in the latter case, there is a negative trend that may be attributable to genuine variation in invasive potential, (Sleeman et al., 2006)).

It is reasonable to ask how much of an effect this potential bias has when using these invasiveness metrics in real-world applications. After all, Sleeman *et al*. found a good correlation between AR and IOR and even our own analysis provides some support for shorter duration serotypes being more invasive. To address this question, we return to our model of IOR, constrained by data (orange line, Fig. 3A). Our model (by assumption) does not allow for any effect of carriage duration (τ) on the underlying drivers of invasiveness (p and τ), so changes in IOR with increasing τ capture the magnitude of bias due to the metric itself. Fig. 3A illustrates that short-carriage ($\tau\ =\ 4 - \mathrm{weeks}$) serotypes generate 5-fold higher IOR values than long-carriage (20-week) serotypes. This work demonstrates that the use of IOR can lead to substantial errors, particularly for short carriage duration serotypes.

Although it is an assumption in our model, we do not contend that all serotypes are truly equally invasive, when assessed by their intrinsic virulence parameters p and d. The positive correlation between innate immune effector interactions and AR invasiveness measurements (Hyams et al., 2013) indicates that serotypes vary, at least somewhat, in their intrinsic invasiveness. And the near significant trend between attack rate and carriage duration (Fig.

3B) may suggest shorter duration serotypes do have a higher propensity to cause invasive disease. Instead, this analysis highlights the fact that IOR values are dependent on carriage duration, so variation in IOR can be due to variation in traits that are not intrinsically related to invasive capacity. The model addresses the timing of infection in a simple manner, separating initial from ongoing risk (governed by p and d respectively). A key future task will be to unpick the within-host dynamics that shape an elevated early risk (potentially linking pneumococcal regulatory dynamics (Shen et al., 2019) with microbiome and host responses).

This investigation has several limitations. First, the data analysis component relies on data obtained from longitudinal and cross-sectional studies from more than 15 years ago, as this offers the most comprehensive data available. Since then, several additional formulations of the pneumococcal conjugate vaccine (PCV) have been released that have dramatically altered the composition of circulating serotypes (Devine et al., 2017). In addition, specific measurements of serotype carriage duration are also impacted by environmental factors that can vary across populations (Lees et al., 2017). Finally, our datasets were pediatric surveillance studies that do not capture unique features of adult invasive disease epidemiology (Alanee et al., 2007). While we acknowledge invasive risk factors vary between adults and children, evidence suggests that children are the major source of community transmission and drive pneumococcal spread which is represented in the model (Althouse et al., 2017). More broadly, we note that the foundation of our investigation is a mathematical analysis based on general principles with the model fitting component added to help ground the model within a real-world context based on the best available data.

Reliable measures of invasiveness are essential for Spn research and infection management, from identifying genomic loci associated with invasive disease (Hyams et al., 2013) to future vaccine development (Løchen et al., 2020). In this paper we have shown that IOR is confounded by carriage duration and may be fundamentally flawed as a result. We further show that the limitations of IOR are not entirely the result of using cross-sectional data, as alternate cross-sectional metrics such as the ratio of invasive to carriage cases can preserve more information about underlying invasive disease processes. While attack rate is more difficult to calculate due to its reliance on longitudinal data for carriage acquisition rates and invasive incidence, our analysis indicates it is a valid invasiveness metric and therefore a more solid platform for basing critical decisions in our public health management of invasive pneumococcal disease.

4. Materials and methods

4.1. Epidemiological model description

To analyze how different invasive disease progression processes can affect measures of invasive disease risk, we construct a compartmental epidemiological model. In this model framework, host individuals are classified as being either susceptible, infected or recovered and immune, with regard to a specific, focal serotype. The infected class is further broken down into a carrier state and an invasive state. The proportions of individuals in a susceptible, carrier, diseased and recovered class for a focal strain are denoted by the variables S, C, I and R respectively (Fig. 1), and their dynamics are given by the following system of four ordinary differential equations,

$$\begin{aligned}\frac{dC}{dt} &= (1-p)\beta CS - \left(\frac{1}{\tau} + d\right)C\\ \frac{dI}{dt} &= p\beta CS - hI + dC\\ \frac{dR}{dt} &= \frac{1}{\tau}C + hI - fR\\ S &= 1 - C - R - I\end{aligned} \tag{1}$$

The model assumes purely frequency-dependent transmission, from carriers only, and no multiple infection. In order to isolate the effect of variable serotype carriage duration on invasiveness measures (*i.e.,* IOR and AR), we assume that intrinsic virulence, the core set of traits that concern adhesion, invasion and proliferation in disease sites, are equivalent for all serotypes. Specifically, we assume that our core set of virulence traits result in a fixed probability p of causing disease directly following initial colonization, and subsequently a fixed rate d of disease progression from the carriage state to the disease state. Both of these invasive disease parameters (p, d) are held constant across serotypes with different carriage durations. Finally, all rates are defined using a time unit of one week. The model variables and parameter definitions are detailed in Table 1.

4.2. Study collection

In order to fit attack rates and invasive odds ratios for this analysis, both longitudinal and cross-sectional data sets are needed. The incidence of IPD and carriage acquisition data are obtained through a series of longitudinal studies outlined in Sleeman *et al.* (Sleeman et al., 2006). The cross-sectional data are described in Brueggemann *et al.* (Brueggemann et al., 2004). Both the invasive odds ratio and carriage prevalence data are extracted from Fig. 3 in Brueggemann *et al.* (Brueggemann et al., 2004) using the WebPlotDigitizer software (Marin et al., 2017). Fitting both datasets poses a challenge since the cross-sectional study included serogroup information while the longitudinal datasets characterized strains down to their serotype. A serogroup is a more general category of antigenically related but distinct serotypes. For example, serotype 12 A, 12B and 12 F are all part of serogroup 12. We address this problem using a method similar to that described by Sleeman *et al.* (Sleeman et al., 2006), where capsular serotypes 19 A, 19 F, 9 A, 9 N and 9 V were removed from the analysis due to the significant variation in attack rates within these serogroups. Unlike Sleeman *et al.* (Sleeman et al., 2006) we included serotypes 1, 4 and 5 in the analysis even though these strains contained missing information, which is highlighted in yellow in Table S1. Here we assume that the carriage durations for these serotypes were too short to be reliably detected given the sampling times of the longitudinal studies, and so a conservative estimate of four weeks (the longest sampling interval used in Sleeman *et al.* (Sleeman et al., 2006)) is used. Because it has been shown that serotypes 1 and 5 are associated with invasive disease (Alanee et al., 2007; Melin et al., 2010), an attack rate of 75 (the attack rate of serotype 4) is given to both these strains, as has been done previously (Hyams et al., 2013). All data analyzed in this paper is presented in Table S1.

4.2.1. Epidemiological model fitting—

All statistical analyses and model fitting are carried out using Mathematica 11.1.1.0. To optimize parameter fitting under multiple constraints we simultaneously fit the cross-sectional and longitudinal datasets. This is accomplished by defining a set of equations that share independent variables as components of a piecewise function (implemented *via* the Mathematica 'Piecewise' function). To associate each expression with its respective dataset we define an index variable that uniquely identifies each dataset and is passed as an additional independent variable. This allows the model fitting function (implemented *via* the Mathematica 'NonlinearModelFit' function) to identify and switch between the respective datasets and expressions. Parameters are estimated by fitting the cross-sectional and longitudinal datasets to a set of equations representing carriage, IOR, AR, carriage acquisition rate and invasive incidence derived from the compartmental model. A more detailed explanation of the compartmental model fitting method with links to code is provided in the Supplement.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Joshua Weitz, Marvin Whiteley, Yuan Li, and members of the Brown lab (GT) for discussion and feedback on this manuscript. We thank the following organizations for funding this work: KWW received a grant from the Wenner-Gren Foundation (https://wennergren.org/). SB is partially funded by NIH awards 1R21AI143296 and 1R21AI156817 (https://www.nih.gov/grants-funding) and CDC contract 75D30120C-09782 (https://www.cdc.gov/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the CDC.

Funding statement

KWW received a grant from the Wenner-Gren Foundation (https://wennergren.org/). SB is partially funded by NIH awards 1R21AI143296 and 1R21AI156817 (https://www.nih.gov/grants-funding) and CDC contract 75D30120C-09782 (https://www.cdc.gov/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Alanee SR, McGee L, Jackson D, Chiou CC, Feldman C, Morris AJ, Ortqvist A, Rello J, Luna CM, Baddour LM, Ip M, Yu VL, Klugman KP, International Pneumococcal Study G, 2007. Association of serotypes of Streptococcus pneumoniae with disease severity and outcome in adults: an international study. Clin. Infect. Dis. 45, 46–51. [PubMed: 17554699]

Allen RC, Popat R, Diggle SP, Brown SP, 2014. Targeting virulence: can we make evolution-proof drugs? Nat. Rev. Microbiol 12, 300–308. [PubMed: 24625893]

Althouse BM, Hammitt LL, Grant L, Wagner BG, Reid R, Larzelere-Hinton F, Weatherholtz R, Klugman KP, Rodgers GL, O'Brien KL, Hu H, 2017. Identifying transmission routes of Streptococcus pneumoniae and sources of acquisitions in high transmission communities. Epidemiol. Infect. 145, 2750–2758. [PubMed: 28847317]

Baldo V, Cocchio S, Lazzari R, Furlan P, Bertoncello C, Russo F, Saia M, Baldovin T, 2015. Estimated hospitalization rate for diseases attributable to Streptococcus pneumoniae in the Veneto region of north-east Italy. Prev. Med Rep. 2, 27–31. [PubMed: 27114894]

Brueggemann AB, Griffiths DT, Meats E, Peto T, Crook DW, Spratt BG, 2003. Clonal relationships between invasive and carriage Streptococcus pneumoniae and serotype- and clone-specific differences in invasive disease potential. J. Infect. Dis. 187, 1424–1432. [PubMed: 12717624]

Brueggemann AB, Peto TE, Crook DW, Butler JC, Kristinsson KG, Spratt BG, 2004. Temporal and geographic stability of the serogroup-specific invasive disease potential of Streptococcus pneumoniae in children. J. Infect. Dis. 190, 1203–1211. [PubMed: 15346329]

Devine VT, Cleary DW, Jefferies JM, Anderson R, Morris DE, Tuck AC, Gladstone RA, O'Doherty G, Kuruparan P, Bentley SD, Faust SN, Clarke SC, 2017. The rise and fall of pneumococcal serotypes carried in the PCV era. Vaccine 35, 1293–1298. [PubMed: 28161425]

Domenech de Cellès M, Opatowski L, Salomon J, Varon E, Carbon C, Boëlle PY, Guillemot D, 2011. Intrinsic epidemicity of Streptococcus pneumoniae depends on strain serotype and antibiotic susceptibility pattern. Antimicrob. Agents Chemother. 55, 5255–5261. [PubMed: 21788454]

Domínguez-Hüttinger E, Boon NJ, Clarke TB, Tanaka RJ, 2017. Mathematical modeling of streptococcus pneumoniae colonization, invasive infection and treatment. Front. Physiol. 8, 115. [PubMed: 28303104]

Dowell SF, Whitney CG, Wright C, Rose CE, Schuchat A, 2003. Seasonal patterns of invasive pneumococcal disease. Emerg. Infect. Dis. 9, 573–579. [PubMed: 12737741]

Gray BM, Converse GM, Dillon HC, 1980. Epidemiologic studies of Streptococcus pneumoniae in infants: acquisition, carriage, and infection during the first 24 months of life. J. Infect. Dis. 142, 923–933. [PubMed: 7462701]

Hyams C, Trzcinski K, Camberlein E, Weinberger DM, Chimalapati S, Noursadeghi M, Lipsitch M, Brown JS, 2013. Streptococcus pneumoniae capsular serotype invasiveness correlates with the degree of factor H binding and opsonization with C3b/iC3b. Infect. Immun. 81, 354–363. [PubMed: 23147038]

Lees JA, Croucher NJ, Goldblatt D, Nosten F, Parkhill J, Turner C, Turner P, Bentley SD, 2017. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. Elife 6.

Lees JA, Ferwerda B, Kremer PHC, Wheeler NE, Serón MV, Croucher NJ, Gladstone RA, Bootsma HJ, Rots NY, Wijmega-Monsuur AJ, Sanders EAM, Trzciński K, Wyllie AL, Zwinderman AH, van den Berg LH, van Rheenen W, Veldink JH, Harboe ZB, Lundbo LF, de Groot LCPG, van Schoor NM, van der Velde N, Ängquist LH, Sørensen TIA, Nohr EA, Mentzer AJ, Mills TC, Knight JC, du Plessis M, Nzenze S, Weiser JN, Parkhill J, Madhi S, Benfield T, von Gottberg A, van der Ende A, Brouwer MC, Barrett JC, Bentley SD, van de Beek D, 2019. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. Nat. Commun. 10, 2176. [PubMed: 31092817]

Løchen A, Croucher NJ, Anderson RM, 2020. Divergent serotype replacement trends and increasing diversity in pneumococcal disease in high income settings reduce the benefit of expanding vaccine valency. Sci. Rep. 10, 18977. [PubMed: 33149149]

Løchen A, Truscott JE, Croucher NJ, 2022. Analysing pneumococcal invasiveness using Bayesian models of pathogen progression rates. PLoS Comput. Biol. 18, e1009389. [PubMed: 35176026]

Marin F, Rohatgi A, Charlot S, 2017. WebPlotDigitizer, a polyvalent and free software to extract spectra from old astronomical publications: application to ultraviolet spectropolarimetry. arXiv: Instrumentation and Methods for Astrophysics.

Melegaro A, Gay NJ, Medley GF, 2004. Estimating the transmission parameters of pneumococcal carriage in households. Epidemiol. Infect. 132, 433–441. [PubMed: 15188713]

Melin M, Trzciq ski K, Meri S, Käyhty H, Vd kevd inen M, 2010. The Capsular Serotype of Streptococcus pneumoniae Is More Important than the Genetic Background for Resistance to Complement. Infect. Immun. 78, 5262–5270. [PubMed: 20855513]

Metcalf BJ, Chochua S, Walker H, Tran T, Li Z, Varghese J, Snippes Vagnone PM, Lynfield R, McGee L, Li Y, Pilishvili T, Beall B, 2021. Invasive pneumococcal strain distributions and isolate clusters associated with persons experiencing homelessness during 2018. Clin. Infect. Dis. 72, e948–e956. [PubMed: 33150366]

O'Brien KL, Wolfson LJ, Watt JP, Henkle E, Deloria-Knoll M, McCall N, Lee E, Mulholland K, Levine OS, Cherian T, Team H.a.P.G.B.o.D.S., 2009. Burden of disease caused by Streptococcus pneumoniae in children younger than 5 years: global estimates. Lancet 374, 893–902. [PubMed: 19748398]

Shen P, Lees JA, Bee GCW, Brown SP, Weiser JN, 2019. Pneumococcal quorum sensing drives an asymmetric owner-intruder competitive strategy during carriage via the competence regulon. Nat. Microbiol 4, 198–208. [PubMed: 30546100]

Sleeman KL, Griffiths D, Shackley F, Diggle L, Gupta S, Maiden MC, Moxon ER, Crook DW, Peto TE, 2006. Capsular serotype-specific attack rates and duration of carriage of Streptococcus pneumoniae in a population of children. J. Infect. Dis. 194, 682–688. [PubMed: 16897668]

Song JY, Nahm MH, Moseley MA, 2013. Clinical implications of pneumococcal serotypes: invasive disease potential, clinical presentations, and antibiotic resistance. J. Korean Med Sci. 28, 4–15. [PubMed: 23341706]

Walter ND, Taylor TH, Dowell SF, Mathis S, Moore MR, Team ABCSS, 2009. Holiday spikes in pneumococcal disease among older adults. N. Engl. J. Med 361, 2584–2585. [PubMed: 20032333]

Yildirim I, Hanage WP, Lipsitch M, Shea KM, Stevenson A, Finkelstein J, Huang SS, Lee GM, Kleinman K, Pelton SI, 2010. Serotype specific invasive capacity and persistent reduction in invasive pneumococcal disease. Vaccine 29, 283–288. [PubMed: 21029807]

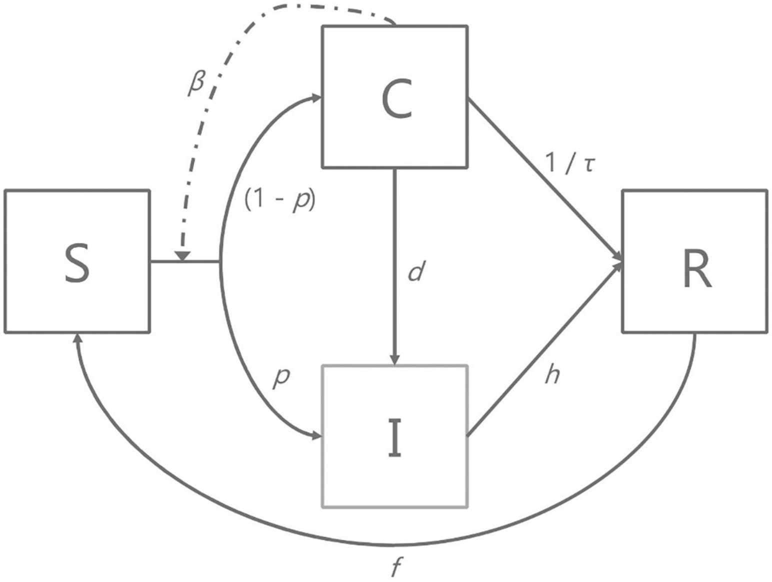

Fig. 1.

Schematic diagram of the epidemiological model. Boxes represent proportions of hosts in mutually exclusive states: susceptible (*S*), infected asymptomatic carriers (*C*), invasive (*I*) or recovered and immune (*R*). Solid arrows represent flows of individuals between states, and dashed arrows represent factors influencing those flows. Equations describing the system are presented in Materials and Methods (methods Eq. 1), along with parameter definitions (Table 1). Note there are two paths from *S* to *I*, a direct path governed by the probability of initial invasion *p*, and an indirect path governed by 1-*p* (probability of initial transition to carriage state) and by the rate *d* of invasive disease progression from a carriage state.

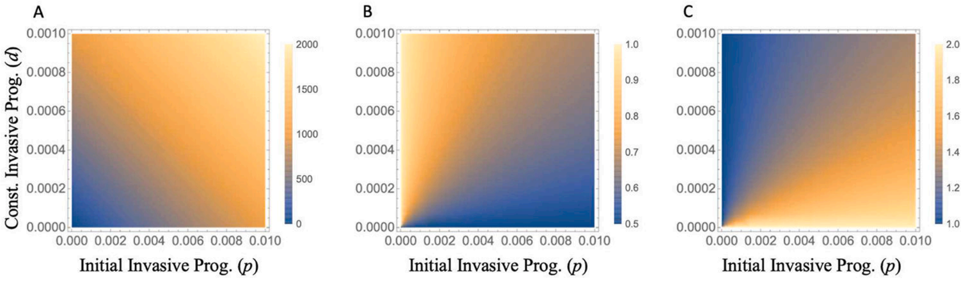

Fig. 2.

Attack rate reliably captures underlying pneumococcal invasiveness parameters while invasive odds ratios fail. (A) Attack rate $\left(\mathrm{AR} = \frac{d\tau + p}{(1 - p)(d\tau + 1)}\right)$ has a positive relationship with both p and d invasive parameters indicating it accurately represents pneumococcal invasiveness. (B), Invasive odds ratios $\mathrm{IOR} = \frac{\tau_0(d\tau + p)}{\tau(d\tau_0 + p)}$ calculated with a low reference carriage duration ($\tau_0 = 5$ fails to capture increasing initial invasive progression (p). (C) Alternatively, IOR fails to capture increasing constant invasive progression (d) when a high carriage duration is used as a reference serotype ($\tau_0 = 20$).

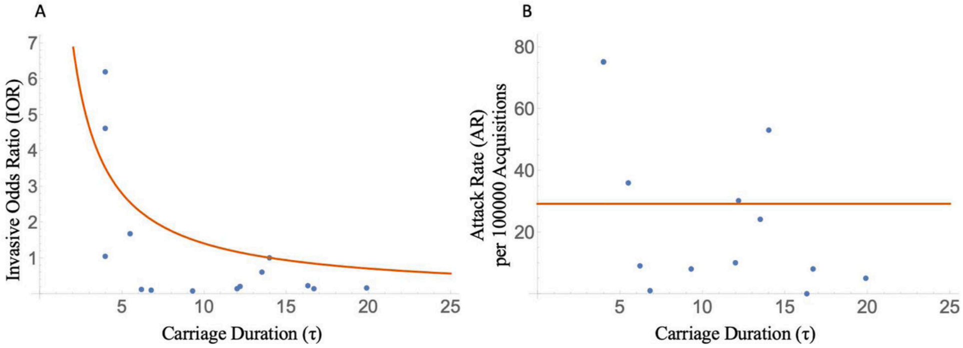

Fig. 3.

Both cross-sectional and longitudinal epidemiological data support the initial risk model and highlight that IOR is confounded by carriage duration. (A) IOR data (blue dots, Brueggemann *et al.* (Brueggemann et al., 2004)) and model fit (orange line, τ_0/τ), against carriage duration (τ). (B) AR data (blue dots, Sleeman *et al.* (Sleeman et al., 2006)) and model fit (orange line, $p(1-p)$), against carriage duration (τ). Simultaneously fitting equations [2] to both datasets (A, B) produced parameter estimates $p = 2.9 \times 10^{-4}$ and $d = 0$ (*i.e.*, invasive disease risk at point of colonization only). Serotype 14 was used as the reference for IOR calculations. IOR and AR data from serotypes 5, 1, 8, 7 F, 4, 38, 18 C, 3, 33 F, 14, 15B/C, 6 A, 23 F, 6B were used in the model fitting (Table S1).

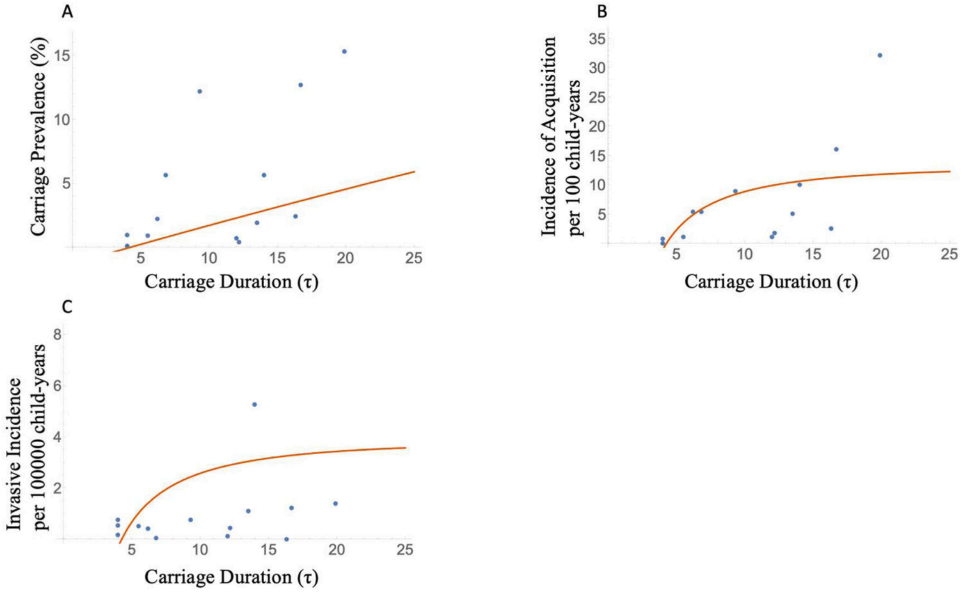

Fig. 4.

Incorporating additional epidemiological data also provides support for the initial risk model. (A), Carriage prevalence data (blue dots, Brueggemann *et al.* (Brueggemann et al., 2004)) and model fit (orange line, $\frac{fh(b(p-1)\tau+1)}{b(h(f(p-1)\tau-1)-fp)}$), against carriage duration (τ). (B) Incidence of acquisition data (blue dots, Sleeman *et al.* (Sleeman et al., 2006)) and model fit (orange line, $\frac{fh(b(p-1)\tau+1)}{b\tau(h(f(p-1)\tau-1)-fp)}$) against carriage duration (τ). (C), Invasive incidence data (blue dots, Sleeman *et al.* (Sleeman et al., 2006)) and model fit (orange line, $\frac{fhp(b(p-1)\tau+1)}{b(p-1)\tau(fh(\tau-p\tau)+fp+h)}$), against carriage duration (τ). Simultaneously fitting endemic equilibrium equations (see SI) to data in Figs. 3A, B and 4A–C produced parameter estimates $p = 2.9 \times 10^{-4}$, $d = 0$, $\beta = 0.24$, and $f = 3.1 \times 10^{-3}$ (*i.e.*, invasive disease risk at point of colonization only). Epidemiological data from serotypes 5, 1, 8, 7 F, 4, 38, 18 C, 3, 33 F, 14, 15B/C, 6 A, 23 F, 6B were used in the model fitting (Table S1).

Table 1

Definitions for the variables and parameters used in the compartmental epidemiological model.

Parameter	Definition
$S(t)$	Proportion of individuals in the susceptible class (at risk of acquiring the focal Spn serotype) at time t
$C(t)$	Proportion of individuals in the carriage class (carrying the focal Spn serotype) at time t
$I(t)$	Proportion of individuals in the invasive class (with an invasive infection caused by the focal serotype) at time t
$R(t)$	Proportion of individuals in the recovered class (individuals who have cleared an infection and whose immunity offers protection from reacquiring the same serotype) at time t.
β	Transmission rate
P	Probability of progressing from carriage to the invasive state at the time of carriage acquisition
τ	The average duration of a carriage for a given Spn strain
d	Rate of progressing from carriage to the invasive state that is constant across the duration of carriage
h	Rate of transition from the invasive to recovered class, due to pathogen clearance.
f	Rate of transition from recovered to susceptible class, due to waning immunity.