**Supplement**

**The SCIR Compartmental Model Fitting Method**

The equations representing attack rate and invasive odds ratio are derived from the stable endemic equilibrium [*S*\*, *C*\*, *I*\*, *R*\*] solutions of the SCIR model given below,

where 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.

Carriage prevalence for a focal serotype is defined by the proportion of carriage cases, *C\**, which we express as a percentage (). Incidence of IPD (Iinc) is calculated as number of new invasive disease cases in 100,000 person-years (Sleeman et al., 2006). The number of new invasive infections is the sum of inflow terms from the invasive disease differential equation,

In our model we track proportions of *S, C, I, and R* and the time unit is weeks. By multiplying the proportions by the population size *n* to get counts and converting the time unit, we generated the following equation for invasive incidence in 100,000 person-years,

Similarly, the incidence of carriage acquisition (Cinc) is calculated as the number of new nasopharyngeal acquisitions divided by the total population at risk in 100 person-years. The description in Table 1 of Sleeman *et al.* (Sleeman et al., 2006) stated carriage acquisition incidence was per 100,000 child-years but we believe this is a typo. New carriage acquisitions are represented by the following equation,

Again, multiplying the proportions by *n* and converting the time unit yields the following equation for carriage acquisition incidence in 100 person-years,

The serotype specific attack rate (AR) from Sleeman *et al.* (Sleeman et al., 2006) is calculated as the ratio of invasive incidence to carriage acquisition incidence per 100000 acquisitions. Simplifying and cancelling terms yields the following equation,

The invasive odds ratio (IOR) is defined as 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*). In the study by Brueggemann *et al.* (Brueggemann et al., 2004) a single reference (serotype 14) was used to calculate IOR. Thus, we use serotype 14’s carriage duration of 14 weeks for the reference type and define IOR as the following equation where is the focal serotype carriage duration,

Because we had carriage prevalence, invasive incidence, carriage acquisition incidence, AR and IOR data for each serotype and the functions describing these epidemiological attributes all used the same parameters, we wanted to estimate the parameters by model fitting all five functions simultaneously. To achieve this, we use the ‘NonlinearModelFit’ function in Mathematica to estimate the *β*, *p*, *d*, and *f* parameters which represent the transmission rate, rapid invasive progression probability, constant invasive progression probability and immunity clearance rate respectively. One other parameter, invasive clearance rate (*h*), is fixed in the analysis with *h* set to 0.5 weeks based on previous work by Baldo *et al.* (Baldo et al., 2015). We fit these equations simultaneously by building an object function comprised of the five equations and by employing the ‘Piecewise’ Mathematica function which uses a dummy variable as a condition in order to associate each equation with its corresponding epidemiological dataset. The code for this model fitting procedure is available as a supplementary file (Mathematica notebook).

**Deriving the Mathematical Relationship between IOR and Carriage Duration**

To understand the first principles of the relationship between IOR and carriage duration using the compartmental model we constructed odds ratio calculations for a hypothetical set of pathogen strains. As described in the methods section, we assume that multiple strains co-circulate independently (thus co-infections do not change focal strain dynamics and need not be tracked explicitly) and reach an endemic equilibrium. Strain comparisons can be made explicit by subscripting strain numbers to infection classes (e.g., *C*0, *C*1) and to strain life-history parameters (e.g., *p0, p1*). The odds that the reference strain 0 causes disease (given carriage of strain 0) is then simply *I*0/*C*0. Similarly, the odds that focal strain 1 causes disease (given carriage of 1) is *I*1/*C*1. Thus, the odds ratio of disease in strain 1 (relative to strain 0) is *I*1*C*0 / *C*1*I*0. We can now ask: How does the odds ratio of disease vary with focal carriage duration? Formulating the invasive odds ratio equation as a function of focal carriage duration yields the following expression,

Increasing the commensal-specific carriage duration τ1 will decrease IOR, as commensal stages will increase their relative representation in the population (*Figure 2A*). Note that with the *d* parameter set to 0, IOR written as a function of focal carriage duration simply becomes the reciprocal equation . In other words, it is just the clearance rate of the focal serotype multiplied by the reference carriage duration. Given that this is a relational expression IOR can therefore vary by orders of magnitude over short carriage durations.

**Invasive OR Analysis**

We begin by testing the basic premise of invasive OR comparisons that increasing the initial or constant disease risk (increasing *p1* or *d1* in our model) will increase the odds ratio of disease. The effect of increasing the constant disease progression in the focal strain (*d1*) on the odds ratio of disease can be readily identified by inspecting the gradient of IOR on *d1*

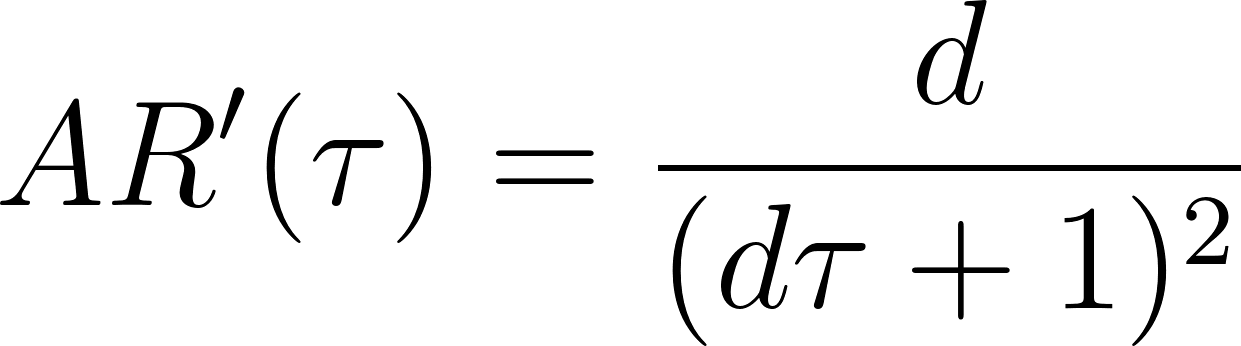
which is positive when . A similar analysis for the effect of the initial carriage risk alone (*p1*) yields an outcome where, assuming , the gradient of OR on *p1*will also be positive. Thus, increasing *p1*, will lead to higher invasive odds ratios only when the reference carriage duration is sufficiently large

To determine what values of *p* and *d* are consistent with the observed negative relationship between IOR and τ we calculate the derivative of IOR with respect to τ

The gradient IOR on τ shows that IOR will maintain a negative association with τ as long as *p* > 0 and *d* is not negative (which is an assumption of the model).

**Deriving the Mathematical Relationship between AR and Carriage Duration**

Similar to our analysis with IOR we can also use the compartmental model to derive an equation for the attack rate and investigate how this same set of focal traits can affect invasiveness. Attack rate is the ratio of invasive disease incidence over carriage acquisition rate. Since both rates are based on the generation of new cases, each measurement only depends on the inflows into their respective compartments as shown in *Figure 1*. Mathematically, attack rate is the sum of inflow terms in the invasive disease differential equation () over the inflow term in the carriage ODE (). The endemic equilibrium solution for the attack rate yielded the following expression

Using the equation for AR(*τ*), we observe that isolating the effect of carriage duration on attack rate reveals a more complicated relationship between *τ*, *d* and AR. Assuming positive parameters, taking the gradient of AR on *τ* yields, . Thus, if the constant progression to invasive disease term *d* is greater than zero, then the gradient of AR on *τ* will always be positive, signifying an increasing proportional relationship between attack rate and carriage duration. However, If the *d* term equals 0, then the attack rate equilibrium expression AR(*τ*) simplifies to which is invariant across carriage duration.

**Attack Rate Analysis**

Using a similar method as the IOR analysis we test the basic premise that increasing the initial or constant disease risk (increasing *p* or *d* in our model) will increase the Spn invasive attack rate. Taking the gradient of AR on *d* reveals that AR’(*d*) is always a positive function and thus AR will always have a positive relationship with the constant invasive risk parameter

This is also true for the initial disease risk scenario. Taking the derivative of the AR with respect to *p* also yields a positive function indicating AR will always increase with a higher initial invasive disease risk

**Two-Stage Carriage Compartmental Model Analysis**

Model design and parameter definitions are the same as described in the main text except the carriage compartment is split into two stages *C1* and *C2* (Figure *S2*). Reconfiguring the Carriage state in this way changes the clearance rate to a sum of 2 independent exponentially distributed random variables with mean 2/ each, resulting in the same average total carriage duration . By introducing a two-step process for the carriage state, we can move away from an exponential distribution of carriage wait times (with modal duration of zero) to an Erlang distribution allowing for a modal duration greater than zero.

The system of ordinary differential equations describing this system is as follows:

Fitting this model to the longitudinal and cross-sectional dataset described in the main text converges to the same solution of parameter estimates as the single carriage stage compartmental model: *p =* 2.9 x10-4(95%CI: 1.4 to 4.6 x 10-4), *d =* 0.0 (CI: 0 to 1.4 x 10-5, *β* = 0.23 (CI: 0.06 to 0.41), f *=* 3.1 x 10-3 (CI: 2.6 x 10-3 to 6.0 x 10-3). Notably, assuming *d* = 0, the IOR and AR expressions simplify to , and *p/*(1-*p*) respectively. Allowing the two carriage clearance rates to vary (*i.e*., setting the duration from *C1* -> *C2* to two weeks and the duration from *C2* -> *R* to *τ* – 2 weeks) does not change the results of the model fit.

**Most Common Metrics Used by Researchers to Measure Pneumococcal Invasiveness**

Several methods can be used for measuring invasiveness across pneumococcal serotypes. They can be categorized into four main approaches: invasive attack rate (IAR), invasive odds ratio (IOR), invasive capacity (IC), and invasive case-to-carrier (Invasive C2C) measurements (see introductions and citations in the main text). The goal of this analysis is to perform a literature survey to estimate how much each metric is used in the field. This survey does present a challenge in that three of the methods (attack rate, odds ratios, and case-to-carrier ratios) are general epidemiological rate measurements and not specific to tracking pneumococcal invasiveness.

Given this constraint, we used the Publish or Perish wrapper tool ([*https://harzing.com/resources/publish-or-perish*](https://harzing.com/resources/publish-or-perish)) to search for selected keywords associated with IAR, IC, IOR and invasive C2C in Google Scholar and sorted the matching papers by year. The search terms used for this analysis are provided below:

Google Scholar Search Terms

Title: ("streptococcus pneumoniae" OR "pneumococcus" OR "pneumococcal")

Keywords: "streptococcus pneumoniae" "invasiveness" ("odds ratio" OR "attack rate" OR ("case-carrier" OR "case-to-carrier") OR "invasive capacity")

When this search was run on 3-12-2023 it yielded 476 papers for review. As mentioned earlier, because attack rates, odds ratios, and case-to-carrier ratios are general epidemiological rates, we must verify the matching papers are referring to Spn invasiveness. Thus, we then read the Methods sections of 83 of the most recent studies (representing studies published since 1-1-2020) to ascertain if the article was using the measurement to track invasive Spn disease. We determined that 36 of the 83 papers made a total of 45 references to Spn invasiveness (some papers referred to more than one reference metric).

The results indicate that invasive metrics derived from cross-sectional data (IOR, IC, and invasive C2C) are more likely to be used than invasive attack rate which is generated using longitudinal data (Table S3). Across metric types, invasive odds ratio was the most common measurement followed by invasive case-to-carrier ratio, invasive capacity, and lastly, invasive attack rate (Table S4).

Table S3

|  |  |
| --- | --- |
| Metric Data | Count |
| Cross-Sectional | 37 |
| Longitudinal | 8 |

Table S4

|  |  |
| --- | --- |
| Metric Type | Count |
| IOR | 18 |
| Invasive C2C | 10 |
| IC | 9 |
| IAR | 8 |

**Fig. S1.**

**Chart

Description automatically generated**

*Figure S1*. Plot of the fitted model of carriage prevalence against invasive OR showing an inverse correlation that matches the negative correlation observed in the cross-sectional data described in Brueggemann *et al.* The parameter is set to 14 weeks.

**Fig. S2.**

**Diagram

Description automatically generated**

*Figure S2*. Schematic diagram of the 2-stage carriage epidemiological model. Boxes represent proportions of hosts in mutually exclusive states: susceptible (*S*), infected asymptomatic carriers (*C1* and *C2*), invasive (*I*) or recovered and immune (*R*). Solid arrows represent flows of individuals between states, and dashed arrows represent factors influencing those flows. 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 *p* and by the rate *d* of invasive disease progression from either of the 2 carriage states.

**Table S1.** Dataset used to fit the epidemiological compartmental model

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **CROSS SECTIONAL DATA** | | **LONGITUDINAL**  **DATA** | | | |
| **Serotype** | **Carriage Prevalence (%)** | **Invasive Odds Ratio** | **Carriage Acquisition Incidence** | **Invasive Incidence** | **Attack Rate** | **Carriage Duration**  **(Weeks)** |
| 5 | 0.11342155 | 6.191531792 | 0 | 0.16 | 75 | 4 |
| 1 | 0.11342155 | 4.613577866 | 0 | 0.74 | 75 | 4 |
| 8 | 0.378071834 | 0.211310528 | 1.78 | 0.43 | 30 | 12.2 |
| 7F | 0.888468809 | 1.674845549 | 1.07 | 0.51 | 36 | 5.5 |
| 4 | 0.964083176 | 1.036985946 | 0.71 | 0.54 | 75 | 4 |
| 38 | 0.699432892 | 0.141202463 | 1.07 | 0.11 | 10 | 12 |
| 18C | 1.890359168 | 0.60801133 | 4.99 | 1.1 | 24 | 13.5 |
| 3 | 2.192816635 | 0.121226123 | 5.35 | 0.4 | 9 | 6.2 |
| 33F | 2.400756144 | 0.220723503 | 2.5 | 0 | 0 | 16.3 |
| 14 | 5.633270321 | 1.003638441 | 9.99 | 5.27 | 53 | 14 |
| 15B/C | 5.633270321 | 0.100729012 | 5.35 | 0.04 | 1 | 6.8 |
| 6A | 12.17391304 | 0.08100603 | 8.93 | 0.76 | 8 | 9.3 |
| 23F | 12.68431002 | 0.130833585 | 16.05 | 1.21 | 8 | 16.7 |
| 6B | 15.29300567 | 0.15074171 | 32.1 | 1.39 | 5 | 19.9 |

*Table S1*. Dataset used to fit the epidemiological compartmental model. This set of observations was created by combining the longitudinal data from Sleeman *et al.* (Sleeman et al., 2006) with the cross-sectional study of Brueggemann *et al.* (Brueggemann et al., 2004). The data were merged using a similar method as that used by Sleeman *et al.* except we included the imputed data points highlighted in yellow. Because the data are derived from separate studies, the column attributes will not perfectly align with one another. For example, serotype 33F registered 0 invasive cases, yet has a non-zero invasive odds ratio. This is because the invasive cases attribute came from the Sleeman study while the IOR data were from Brueggemann *et al.*

**Table S2.** Model comparison of epidemiological models with different invasion timing

|  |  |  |  |
| --- | --- | --- | --- |
| Criterion | Early Progression Model | Cons. Progression Model | Comparison Probability |
| AIC | 568.723 | 599.068 | 2.57e-7 |
| AICc | 569.661 | 599.684 | 3.02e-7 |
| BIC | 579.966 | 608.062 | 7.93e-7 |

*Table S2*: Model comparison of the epidemiological model with the early invasion progression parameter *p* and a constant progression parameter *d* against a model with only the *d* progression parameter. The Akaike information criterion (AIC), corrected Akaike's Information Criterion (AICc) and Bayesian information criterion (BIC) all showed a significantly better fit (lower criterion value) for the model that included the early invasive disease parameter. The comparison probability was generated by using the following formulas:

Δ = constant progression criterion value - early progression criterion value

p = exp(-Δ / 2).