

**THE INTERRELATIONSHIP BETWEEN BACTERIAL AND
SOCIETAL FACTORS DRIVES PNEUMOCOCCAL
TRANSMISSION AND DISEASE PROGRESSION**

A Dissertation
Presented to
The Academic Faculty

by

Benjamin Metcalf

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biological Sciences

Georgia Institute of Technology
August 2024

COPYRIGHT © 2024 BY BENJAMIN METCALF

**THE INTERRELATIONSHIP BETWEEN BACTERIAL AND
SOCIETAL FACTORS DRIVES PNEUMOCOCCAL
TRANSMISSION AND DISEASE PROGRESSION**

Approved by:

Dr. Sam Brown, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Joshua Weitz
School of Biological Sciences
Georgia Institute of Technology

Dr. Kristofer Wollein Waldetoft, Co-
Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Joe Lachance
School of Biological Sciences
Georgia Institute of Technology

Dr. Marvin Whiteley
School of Biological Sciences
Georgia Institute of Technology

Dr. Bernard Beall
Division of Bacterial Diseases
*Centers for Disease Control and
Prevention*

Date Approved: May 01, 2024

ACKNOWLEDGEMENTS

I would like to thank my primary advisor, Dr. Sam Brown, and my co-advisor, Dr. Kristofer Wollein Waldetoft, for their invaluable guidance and expertise throughout this journey. I am deeply appreciative of the insights and support provided by the members of my thesis committee: Dr. Bernard Beall, Dr. Joe Lachance, Dr. Joshua Weitz, and Dr. Marvin Whiteley. Special thanks are due to Lisa Redding and Dr. King Jordan for their knowledge and assistance in fulfilling the requirements of the Bioinformatics program.

I am indebted to Dr. Lesley McGee, my supervisor at the CDC, and the WGS team, Dr. Sopia Chochua, Dr. Yuan Li, and Jonathan Gerhart, for their support and collaboration.

Lastly, my heartfelt gratitude goes out to my family – Mom, Dad, Vanessa, Jordan, Jane, and Owen – for their encouragement and support throughout this endeavor.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
LIST OF TABLES	vii
LIST OF FIGURES	ix
LIST OF SYMBOLS AND ABBREVIATIONS	xviii
SUMMARY	xx
CHAPTER 1. Introduction	1
1.1 Streptococcus Genus	1
1.2 <i>Streptococcus pneumoniae</i> Clinical Microbiology and Diagnostics.	3
1.3 <i>S. pneumoniae</i> Life Cycle and epidemiological modeling	4
1.4 <i>S. pneumoniae</i> Virulence Factors	6
1.5 United States Surveillance of IPD	9
1.6 Synopsis	11
1.6.1 Chapter Two	11
1.6.2 Chapter Three	12
1.6.3 Chapter Four	12
1.6.4 Chapter Five	14
CHAPTER 2. VARIATION IN PNEUMOCOCCAL INVASIVENESS METRICS IS DRIVEN BY SEROTYPE CARRIAGE DURATION AND INITIAL RISK OF DISEASE 15	
2.1 Summary	15
2.2 Introduction	15
2.3 Results	18
2.3.1 Analyzing the impact of invasive disease timing on metrics of invasive disease (IOR and AR)	18
2.3.2 Fitting the SCIR Compartmental Model to the Spn Carriage and Invasive Disease Data	23
2.3.3 Alternate invasiveness metrics.	26
2.4 Discussion	27
2.5 Materials and Methods	30
2.5.1 Epidemiological Model Description	30
2.5.2 Study Collection	31
2.5.3 Epidemiological Model Fitting	32
2.6 Acknowledgements	33
CHAPTER 3. HUMAN POPULATION DENSITY DRIVES PNEUMOCOCCAL DISEASE RISK VIA DENSITY-DEPENDENT SEROTYPE TRANSMISSION 34	
3.1 Summary	34
3.2 Introduction	34

3.3	Results	38
3.3.1	Population Density is Correlated with Serotype Carriage Duration and Invasiveness	38
3.3.2	Pneumococcal transmission is both frequency- and density-dependent	41
3.3.3	Sensitivity Analysis	47
3.4	Discussion	49
3.5	Methods	52
3.5.1	Study Collection	52
3.5.2	Epidemiological Model Description	54
3.5.3	Statistical Analysis	61
CHAPTER 4. HUMAN SOCIETAL DETERMINANTS AND BACTERIAL LIFE HISTORY TRAITS SHAPE PNEUMOCOCCAL INVASIVE DISEASE RISK		64
4.1	Summary	64
4.2	Introduction	65
4.3	Results	68
4.3.1	Expected Number of Clustered Isolates Assuming an Independent Outbreak Risk	68
4.3.2	Proportion of cluster-associated cases is substantially greater than our null expectation	68
4.3.3	IPD clustering and serotype.	69
4.3.4	IPD clustering is associated with serotype carriage duration and invasive attack rate	71
4.3.5	Social features associated with health inequality are significant predictors of IPD and clustering	72
4.4	Discussion	78
4.5	Methods	82
4.5.1	CDC ABCs surveillance data (genomic and case history)	82
4.5.2	Census data and bacterial traits	82
4.5.3	Cluster Identification	84
4.5.4	Causal inference methods	85
4.5.5	Statistical modeling	86
4.5.6	Mathematical modeling	87
CHAPTER 5. GENERAL DISCUSSION		90
5.1	Summary of work	90
5.2	Societal and Within-Host Determinants of Invasive Pneumococcal Disease Outbreaks	91
5.2.1	Population Level	93
5.2.2	Within-host Level	94
5.3	Proposed Future Work	99
5.3.1	Identifying genes and loci associated with invasive clustering	99
5.3.2	In vitro and in vivo model systems to assess the impact of Neutrophil delay	101
APPENDIX A. Supplemental Material for Chapter 2		104
APPENDIX B. Supplemental Material for Chapter 3		118

APPENDIX C. Supplemental Material for Chapter 4	125
REFERENCES	130

LIST OF TABLES

Table 2.1	Definitions for the variables and parameters used in the compartmental epidemiological model.	19
Table 3.1	Fitted coefficient values for parameters used in the power law and piecewise contact rate models.	43
Table 3.2	Definitions for the variables and parameters used in the compartmental epidemiological model along with values and reference sources used in the model fitting.	56
Table 3.3	The distribution function and sampling boundaries used in the sensitivity analyses testing the inverse relationship between population density and carriage duration. Parameter ranges are modified from Table 3.1 to use truncated Normal distributions where required to avoid parameter selections outside of biological limits (for parameters r and b). The parameter combinations plots sampled from the normal and truncated normal distributions to generate 500 sets of parameter estimates for both the power and piecewise contact rate models. The GSI sensitivity analysis used the upper and lower boundary constraints representing the 95 th percentiles of the normal and truncated normal distributions to implement the GSI factorial design method.	62
Table 4.1	Definitions of the demographic attributes included as possible IPD case and clustering determinants, extracted from US Census data (both the American Community Survey (ACS) and the Planning Database (PDB)). Units are provided as parentheticals in the Definition column.	73
Table A.1	Dataset used to fit the epidemiological compartmental model. This set of observations was created by combining the longitudinal data from Sleeman <i>et al.</i> (Sleeman <i>et al.</i> , 2006) with the cross-sectional study of Brueggemann <i>et al.</i> (Brueggemann <i>et al.</i> , 2004). The data were merged using a similar method as that used by Sleeman <i>et al.</i> 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 <i>et al.</i>	115
Table A.2	Model comparison of epidemiological models with different invasion timing. Model comparison of the epidemiological model with the early invasion progression parameter p and a constant progression	116

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: $\Delta = \text{constant progression criterion value} - \text{early progression criterion value}$ and $p = \exp(-\Delta / 2)$.

Table B.1	Pneumococcal life-history traits and host population metadata aggregated by serotype. This set of observations was created by combining the longitudinal data from Sleeman et al. with population density data from United States Census Bureau's 2021 Planning Database. Table provides invasive case counts (IPD case count), locations sampled (State Count), census tracts sampled (Census Tract Count), average census tract population density (Avg Pop. Dens.), average census tract population size (Avg Pop. Size), per capita IPD rate (IPD Rate), and carriage duration.	120
Table C.1	Relationship between attribute and dependent variable (cluster cases or total cases) can vary depending on the causal model. Rows represent the variables of interest and columns represent a causal model choice. Cells shaded red indicates either a change in coefficient significance or sign compared to the baseline model, DAG Model 1. Asterisks designate the following significance thresholds: * = <0.05, ** = <0.01, *** = <0.001.	125
Table C.2	Definitions for the variables and parameters used in the compartmental epidemiological model.	128

LIST OF FIGURES

- Figure 1.1 Phylogenetic tree representing the six major species groups of Streptococcus based on a comparison of the 16S rRNA gene sequence (Kilian, 2010, Nobbs et al., 2009). The three major pathogens of the Streptococcus genus are spread across two species groups. *S. agalactiae* (GBS) and *S. pyogenes* (GAS) belong to the pyogenic group (black rectangles) while *S. pneumoniae* (Spn) is a member of the mitis group (red rectangle). Members of the viridans group are underlined in green. 3
- Figure 1.2 Schematic of the human nasal cavity, illustrating the location of Spn carriage in the nasopharynx. 5
- Figure 1.3 Schematic diagram of Spn epidemiological models. 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. Parameters regulating the transition between each compartment are shown next to each arrow (e.g. *p*, *h*, *f*). 6
- Figure 1.4 The Active Bacterial Core surveillance catchment area. The blue areas are counties spread across 7 US states including San Francisco Bay Area, California (CA) (3 counties), Denver, Colorado (CO) (5 counties), Atlanta, Georgia (GA) (20 counties), Portland, Oregon (OR) (3 counties), Rochester and Albany, New York (NY) (16 counties), select urban counties in Tennessee (TN) (20 counties), and Baltimore area, Maryland (MD) (6 counties). The orange areas represent the three U.S. states Connecticut (CT), Minnesota (MN), and New Mexico (NM). 10
- Figure 2.1 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 equations 2.1), along with parameter definitions (table 2.1). 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. 19

- Figure 2.2 Attack rate reliably captures underlying pneumococcal invasiveness parameters while invasive odds ratios fail. (A) Attack rate ($AR = \frac{d\tau+p}{(1-p)(d\tau+1)}$) has a positive relationship with both p and d invasive parameters indicating it accurately represents pneumococcal invasiveness. (B), Invasive odds ratios ($IOR = \frac{\tau_0(d\tau+p)}{\tau(d\tau_0+p)}$) 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 ($\tau_0 = 20$). 22
- Figure 2.3 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.* [5]) and model fit (orange line, τ_0/τ), against carriage duration (τ). (B) AR data (blue dots, Sleeman *et al.* [4]) and model fit (orange line, $p(1-p)$), against carriage duration (τ). Simultaneously fitting equations [2.2] 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, 7F, 4, 38, 18C, 3, 33F, 14, 15B/C, 6A, 23F, 6B were used in the model fitting (Table A.1). 24
- Figure 2.4 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 figures 2.3A, 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, 7F, 4, 38, 18C, 3, 33F, 14, 15B/C, 6A, 23F, 6B were used in the model fitting (Table A.1). 26
- Figure 3.1 Alternate models for the force of infection λ as a function of population density N . The blue line defines a frequency-dependent contact structure ($\lambda = \kappa I$, where I is the frequency of infecteds and κ is a rate constant) that is independent of total population density N . The green line defines a density-dependent contact structure ($\lambda = \kappa NI$, 35

where NI defines the total density of infecteds in the population). The orange line defines a power contact structure ($\lambda = \kappa N^r I$, contact rate = κN^r) and represents a ‘hybrid’ of frequency- and density-dependent limit cases (when $r = 0$ we recover the pure frequency-dependent model, and when $r = 1$ we recover the pure density dependent model).

- Figure 3.2 Average serotype population density varies widely across serotypes. Bar chart of average host population density by serotype where population density is measured in persons per square mile (PPSM). The bar color encodes the number of IPD cases for each serotype with the darker color representing higher case counts. Horizontal black line: average population density across all census tracts with IPD cases regardless of serotype. 40
- Figure 3.3 Average serotype population density has a negative correlation with carriage duration and a positive correlation with invasive attack rate (AR). (A) Serotype carriage duration (blue dots) against log10 average serotype population density (measured using census tracts) yields a significant inverse association (Kendall Tau = -0.348, p-value = 0.00961). (B) Invasive AR (blue dots) against log10 average serotype population density (measured using census tracts) yields a significant positive association (Kendall Tau = 0.328, p-value = 0.011). Two examples of short duration, high invasive serotypes (4 and 7F) are labeled in each plot. 41
- Figure 3.4 Both the fitted power and piecewise contact rate models indicate that Spn transmission is density dependent. The fitted power contact rate model transmission coefficient β (orange line, $\beta(b,r) = bN^r$) where $r = 0.988$ and $b = 7.69 \times 10^{-5}$ against population density N . The fitted piecewise contact rate model transmission coefficient β (green line, $\beta(b_D) = \begin{cases} b_D N, & N < 3722.82 \\ b_D N_{\kappa}, & N \geq 3722.82 \end{cases}$) where $b_D = 6.67 \times 10^{-5}$ against population density N . The population density threshold of 3722.82 is calculated by plugging the fitted $\kappa = 11.07$ value into the function for population density (methods equation 3.8) and using the average per capita invasive incidence of 7.69×10^{-7} cases/week. Both models have a non-constant relationship with population density suggesting that Spn spreads through density-dependent transmission at least in lower population density regions. Model comparisons are made using AICc scores (Burnham et al., 2011). For comparison, the pure density- and frequency-dependent limit cases had AICc scores of 1032.78 and 1222.08 respectively. 44
- Figure 3.5 Fitting population density and invasive incidence using a power contact rate function indicates Spn transmission is density dependent. A) Spn serotype average population density data (blue dots) and the model fit (green-to-white gradient, $N_p(\tau, D^*) =$ 47

$$\begin{cases} N_D P(\tau, D^*), & \tau < \kappa \\ N_D P(\kappa, D^*), & \tau \geq \kappa \end{cases}, \text{ where } N_D P(\tau, D^*) = \frac{hp}{b_D(p-1)\tau(-fhD+h(Dp\tau-D\tau+p)-Dp)}$$
 of population density as a function of carriage duration τ and equilibrium invasive incidence D^* . Model fits support the hypothesis Spn transmission is density-dependent and that shorter-duration serotypes are more likely to cause invasive disease in higher-density areas. The model fitting produced the parameter estimates $b_D = 6.67 \times 10^{-5}$, $\kappa = 11.07$, and $f_P = 229$. B) A 3D plot of population density as a function of carriage duration τ and invasive incidence D^* . The vertical lines connecting each datapoint to the fitted function represent the residual error. Additional fixed parameter values are invasive clearance $h = 0.5$ and probability of invasion $p = 0.00029$.

- Figure 3.6 Sensitivity analysis of the population density functions (equations 3.4 and 3.9) reveals that the model fitting results are robust across a wide range of parameter values. A) and C) Parameter combination plots of equations 3.4 (Figure 3.6A) and 3.9 (Figure 3.6C) (see methods) generated by sampling 500 sets of parameter estimates from the distributions described in Table 3.3 (grey lines) overlaid with Spn serotype average population density data (blue dots). Both graphs show that even with wide variation in parameter values the overall negative association between carriage duration and population density generally holds. B) First order (grey bars) and total (black bars) generalized sensitivity indices (GSI) for the population density function based on the power contact rate model (equation 3.4). Results show that population density is mainly influenced by changes in the power exponent parameter r and immune protection parameter f with a smaller contribution from the transmission rate constant b . D) First order (grey bars) and total (black bars) generalized sensitivity indices (GSI) for the population density function based on the piecewise contact rate model (equation 3.9). The model was most sensitive to changes in the immune protection parameter f_P and the density-dependent transmission constant b_D . 49
- Figure 3.7 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 the main text (methods equations 3.1), along with parameter definitions (Table 3.2). See Chapter Two for initial model development. 54
- Figure 4.1 The amount of invasive clustering varies widely between Spn serotypes and is correlated with disadvantaged populations (PEH or 70

PWID). Invasive clustering has a significant positive relationship with proportion of PEH or PWID (Kendall tau = 0.48, p-value = 0.00016). The average proportion of cluster-associated isolates weighted by invasive case count was 24%. The scatter plot reveals that serotypes vary widely in their propensity to cluster ranging from 0% cluster-associated isolates (serotypes 1, 18C, 21, 23F, 37, 6A, 6B) to 68% cluster-associated (serotype 12F). The top 4 most clustered serotypes (12F, 4, 7F, 20) are labeled in the figure.

- Figure 4.2 IPD clustering proportion is associated with bacterial life-history traits (positively correlated with high invasive attack rates and short carriage durations). (A) Serotype IPD clustering proportion (blue dots) against serotype carriage duration yields a significant inverse association (Kendall tau = -0.32, p-value = 0.018). (B) Serotype IPD clustering proportion (blue dots) against serotype invasive attack rates yields a significant positive correlation (Kendall tau = 0.29, p-value = 0.028). 71
- Figure 4.3 The relationship between both household crowding and bus transit with IPD clustering is confounded by financial instability and population density. The directed acyclic graph (DAG) represents the causal interrelationships between variables (nodes) influencing the development of IPD and IPD clusters. Dark grey nodes are observed variables and light grey node are unobserved (latent) variables. Demographic features of interest (Table 4.1) are highlighted in blue and potential confounders (Table 4.1) are highlighted in green. The IPD cluster outcome variable is highlighted in red. 75
- Figure 4.4 Several demographic features associated with health inequality and crowding are significant predictors of IPD disease and clustering. (A) and (B) The results show that census tracts with lower average household incomes, higher population densities, and a higher use of public bussing are positively associated with both IPD total cases (panel A) and IPD clustering cases (panel B). The standardized coefficient confidence intervals for each variable are derived from a negative binomial multiple regression model that is adjusted for each variable of interest to have a non-biased effect on IPD outcome based on the causal DAG in Figure 4.3. (C) The minimal sufficient adjustment set of each variable of interest as specified by DAG Model 1 and DAG Model 2. 77
- Figure 5.1 Schematic of IPD cluster hypothesis. We hypothesize that PEH (grey oval) suffer high rates of IPD clusters (red oval) via 3 interacting components (green boxes) mediated by two processes (blue boxes): human density, immune dysfunction, and bacterial life-history traits. These components influence IPD cluster formation at both the 92

population (yellow, left half of figure) and within-host levels (grey, right half of figure).

- Figure 5.2 We propose that the progression to IPD is driven by a positive feedback loop between invasive inflammation and pneumococcal infection. The width of the green activation and red inhibition symbols represent the magnitude of the effect. Innate immune cells engage pneumococcus in a coupled dynamical system structured around inhibition of pneumococcal growth by innate immune response (red inhibition symbol) and a potential positive feedback loop (clockwise circuit of arrow symbols). (A) Within the general population, the interaction between innate immunity and Spn is shifted towards inhibition (wider red inhibition symbol) that leads to Spn control and eventual clearance. (B) Within vulnerable populations the delay of innate cell recruitment to the colonization site and diminished inhibition capability (thinner red arrow) leads to uncontrolled growth of Spn. 96
- Figure 5.3 People experiencing homeless (PEH) have suppressed immunological responses compared to housed individuals, as measured by an array of neutrophil functional assays. In a previously published case-control study, Arranz *et al* found significant differences in innate immune response between unhoused individuals and housed persons matched for gender and age. Using human peripheral blood neutrophil samples, they found that Neutrophil adherence (a), neutrophil chemotaxis (b), neutrophil phagocytosis (c, d) and non-stimulated and stimulated superoxide anion levels (e) were all significantly higher in housed versus unhoused individuals. AI = Adherence index; CI = chemotaxis index; PI = phagocytic index; PE = phagocytic efficiency. 99
- Figure 5.4 Population structure of *Streptococcus pneumoniae*. Phylogeny was constructed from 2,416 IPD genomic isolates acquired through the CDC Active Bacterial Core surveillance (ABCs) program during 2015. The deep branching lineages apparent across the tree demonstrate the importance of controlling for population structure when conducting bacterial genome-wide association studies across serotypes or lineages. 101
- Figure A.1 Parametric plot of the fitted model of carriage prevalence against the fitted model of invasive OR with carriage duration as the independent variable. The plot shows an inverse correlation that matches the negative correlation observed in the cross-sectional data described in Brueggemann *et al*. The parameter τ_0 is set to 14 weeks. 114
- Figure A.2 Schematic diagram of the 2-stage carriage epidemiological model. Boxes represent proportions of hosts in mutually exclusive states: 115

susceptible (S), infected asymptomatic carriers (CI 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.

Figure B.1 Average serotype population density has a negative correlation with carriage duration and a positive correlation with invasive attack rate (AR). (A) Serotype carriage duration (blue dots) against log10 average serotype population density (measured using ZCTAs) yields a significant inverse association (Kendall Tau = -0.385, p-value = 0.00415). (B) Invasive AR (blue dots) against log10 average serotype population density (measured using ZCTAs) yields a significant positive association (Kendall Tau = 0.256, p-value = 0.048). These relationships indicate that shorter duration and more highly invasive serotypes are more likely to cause IPD in denser environments. 121

Figure B.2 Fitting invasive incidence using a power contact rate function provides support for Spn density dependent transmission. Spn per capita invasive incidence data in weeks (blue dots) and the model fit 122

$$(\text{blue line, } D_p^*(\tau, N) = \begin{cases} -\frac{p(b_D h N (p-1)\tau + h)}{(b_D N \tau - b_D N p \tau)(fh - hp\tau + h\tau + p)}, & \tau < \kappa \\ \frac{hp(b_F (p-1)\tau + 1)}{b_F (p-1)\tau (fh - hp\tau + h\tau + p)}, & \tau \geq \kappa \end{cases}) \text{ of}$$

invasive incidence as a function of carriage duration τ and population density N . Model fit supports the hypothesis Spn transmission is density-dependent and that shorter-duration serotypes are more likely to cause invasive disease in higher-density areas. Population density N was set to $N_\kappa = 3722.82$ PPSM, making the density- and frequency-dependent transmission coefficients equivalent resulting in a continuous curve across the two piecewise components.

Figure B.3 Fitting population density and invasive incidence using a power contact rate function indicates Spn transmission is density dependent. 123

A) Spn serotype average population density data (blue dots) and the model fit (green-to-white gradient, $N(\tau, D^*) = \left(\frac{b(1-p)\tau(fh - hp\tau + h\tau + p) \left(-\frac{hp^2}{(1-p)(fh - hp\tau + h\tau + p)} + \frac{hp}{(1-p)(fh - hp\tau + h\tau + p)} - D^* \right)}{hp} \right)^{-1/r}$)

of population density as a function of carriage duration τ and equilibrium invasive incidence D^* . B) Spn per capita invasive incidence data in weeks (blue dots) and the model fit (green-to-white gradient, $D^*(\tau, N) = \frac{pN^{-r}(bh(p-1)\tau N^r + h)}{b(p-1)\tau(fh - hp\tau + h\tau + p)}$) of invasive incidence as a function of carriage duration τ and population density N . C) A 3D plot of both population density and invasive incidence as functions of

carriage duration τ . The vertical lines connecting each datapoint to the fitted function represent the residual error. Both model fits support the hypothesis Spn transmission is density-dependent and that shorter-duration serotypes are more likely to cause invasive disease in higher-density areas. The model fitting estimated the power contact rate exponent $r = 0.988$ (i.e., Spn transmission is constrained by a nearly linear contact density function), the immune duration parameter $f = 237$, and the transmission rate constant parameter $b = 0.0000769$. Additional fixed parameter values are invasive clearance $h = 0.5$ and probability of invasion $p = 0.00029$.

- Figure B.4 Additional Sensitivity plots of the population density functions (equations 3.4 and 3.8). A) Dynamics of the sensitivity indices of the power contact rate density function (equation 3.4) from $\tau = 1$ to 35 reveals that the sensitivity of population density to the main effects of r , f , and b remain stable across the range of carriage durations. B) The sensitivity indices of the piecewise contact rate density function (equation 3.9) from $\tau = 1$ to 35 indicates the f_P and b_D parameters remain stable across carriage durations but sensitivity to the threshold parameter κ increases as the model transitions from density- to frequency-dependent transmission. 124
- Figure C.1 The relationship between IPD outcome (case or cluster) with household crowding, bus transit, and household income is confounded. The directed acyclic graph (DAG) represents the causal interrelationships between variables (nodes) influencing the development of IPD and IPD clusters. Dark grey nodes are observed variables and light grey node are unobserved (latent) variables. Demographic features of interest are highlighted in blue and potential confounders are highlighted in green (Table 4.1). The IPD case and cluster outcome variable is highlighted in red. 125
- Figure C.2 The Active Bacterial Core surveillance catchment area. The blue areas are counties spread across 7 US states including San Francisco Bay Area, California (CA) (3 counties), Denver, Colorado (CO) (5 counties), Atlanta, Georgia (GA) (20 counties), Portland, Oregon (OR) (3 counties), Rochester and Albany, New York (NY) (16 counties), select urban counties in Tennessee (TN) (20 counties), and Baltimore area, Maryland (MD) (6 counties). The orange areas represent the three U.S. states Connecticut (CT), Minnesota (MN), and New Mexico (NM). 127
- Figure C.3 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. 128

Equations describing the system are presented in Materials and Methods (methods equations 4.1), along with parameter definitions (Table C.2).

LIST OF SYMBOLS AND ABBREVIATIONS

Spn	Streptococcus pneumoniae (pneumococcus)
IPD	Invasive pneumococcal disease
GAS	Group A Streptococcus
GBS	Group B Streptococcus
PS	Polysaccharide capsule
ABCs	Active bacterial core surveillance
Cps	Capsular polysaccharide biosynthesis
PLY	Pneumolysin
PI-1	Pilus islet 1
PI-2	Pilus islet 2
Cbp	Choline binding proteins
Pht	Histidine triad proteins
FH	Human factor H
CDC	Centers for disease control and prevention
IOR	Invasive odds ratios
AR	Invasive attack rates
PEH	People experiencing homelessness
PWID	People who inject drugs
GSI	Generalized sensitivity indices
GWAS	Genome-wide association study
SCIR	Susceptible-carrier-invasive-recovered compartmental model
PCV	Pneumococcal conjugate vaccine

PPSM Persons per square mile
ZCTAs ZIP code tabulation areas
PDB Planning database
SNPs Single nucleotide polymorphisms
DAG Directed acyclic graph
HPC High-performance computing
LTCF Long term care facility
PRR Pattern recognition receptors
MRC-1 Mannose receptor C type 1
AIC Akaike information criterion
BIC Bayesian information criterion

SUMMARY

Streptococcus pneumoniae (Spn) is a gram-positive opportunistic pathogen and a leading cause of bacterial pneumonia, meningitis, and sepsis worldwide. Spn is typically categorized by serotype and in most cases colonizes the nasopharynx asymptotically (defined as carriage). On rare occasions, it can spread into normally sterile sites and cause severe infections known as invasive pneumococcal disease (IPD). Both the ability of Spn to cause these life-threatening illnesses (*i.e.*, its invasiveness) and the length of time in carriage (*i.e.*, its carriage duration) vary substantially across serotypes. IPD cases can occur in isolation or, occasionally, spread as localized outbreaks, often arising within disadvantaged communities like people experiencing homelessness (PEH) and people who inject drugs (PWID). In this dissertation, we seek to explore how pneumococcal life-history traits, host demographic factors, and the interrelationships between them can impact invasive disease at the population level.

In Chapter Two we show that variation in carriage duration can confound serotype invasiveness measurements and identify the best performing metrics for invasiveness. Our results also support previous work showing that when invasion occurs, it does so at or near the time of carriage acquisition. In Chapter Three we show that Spn transmission is host density-dependent, resulting in a stratification of serotypes by human population density. The fourth chapter provides evidence that the risk of IPD and IPD clustering is in part governed by human social conditions, with increased risks in disadvantaged populations. In the final discussion chapter, we provide a broader discussion of potential future directions of research, building on the work in this thesis.

CHAPTER 1. INTRODUCTION

In this introductory chapter, I will cover broad topics to introduce general concepts underlying the thesis as a whole. Chapter Two, Chapter Three, and Chapter Four all have self-contained introductions that go into more specific detail about their specific topics. A central thread to this thesis is the major human opportunistic pathogen *Streptococcus pneumoniae* (Spn, also referred to as the pneumococcus). In this chapter I will review the basic biology of the *Streptococcus* genus, with a focus on Spn, followed by a synopsis of the thesis chapters that follow. The synopsis will introduce the central questions in this thesis, dealing with the interrelationships between bacterial life-history traits, social determinants of health, and impacts on human disease.

1.1 Streptococcus Genus

The *Streptococcus* genus represents a diverse set of over 100 gram-positive bacterial species that are marked by spherical or ovoid cells and typically arranged in pairs or chains (Hardie and Whiley, 1997, Krzyściak et al., 2013). They primarily colonize the skin and mucosal surfaces (e.g. oral cavity, throat, and gastrointestinal tract) of humans and animals and include both commensals and opportunistic pathogens (Krzyściak et al., 2013). Historically, there have been a number of classification systems employed to define taxa within the genus to aid in the biological and epidemiological study of this diverse set of species (Hardie and Whiley, 1997). Using a taxonomic system based on 16S rRNA, the genus can be subdivided into 6 species groups: pyogenic, mitis, anginosus, bovis, mutans, and salivarius ((Kilian, 2010, Nobbs et al., 2009). Figure 1.1). The five species groups excluding pyogenic have also been included within a larger division called viridans group

streptococci (viridis is Latin for green and many of the species produce a greenish discoloration on blood agar) though this categorization has been previously criticized as being inconsistent and poorly defined (Doern and Carey-Ann, 2010). The pyogenic group is notable in that it contains three major β -hemolytic human streptococcal pathogens: *Streptococcus pyogenes*, *Streptococcus agalactiae*, and *Streptococcus dysgalactiae* subsp. *equisimilis*.

S. pyogenes or group A *Streptococcus* (GAS) is a common cause of bacterial pharyngitis and skin infections and is the etiology of a broad variety of serious invasive clinical manifestations including bacteremia, pneumonia, septic arthritis, necrotizing fasciitis and streptococcal toxic shock syndrome (Walker et al., 2014). Rates of invasive GAS disease (iGAS) have increased over recent years from 3.6 cases per 100,000 population in 2013 to 7.6 cases per 100,000 population in 2019 in the United States (<https://www.cdc.gov/abcs/bact-facts-interactive-dashboard.html>). *S. agalactiae* (or group B *Streptococcus*, GBS) has been recognized as a major cause of illness in neonates and pregnant women with the severity of disease ranging from mild urinary tract infections to more serious conditions like bacteremia, endocarditis, and meningitis (Lannes-Costa et al., 2021). Work by Edwards *et al.* has shown that *S. agalactiae* colonization is significantly associated with black race and younger maternal age (Edwards et al., 2019).

The mitis group is home to the third major pathogen of the streptococcus genus, *Streptococcus pneumoniae*, also known as the pneumococcus. Pneumococcus is the focus of this dissertation and in following sections we will discuss the clinical microbiology, epidemiology, virulence, and public health surveillance of this organism.

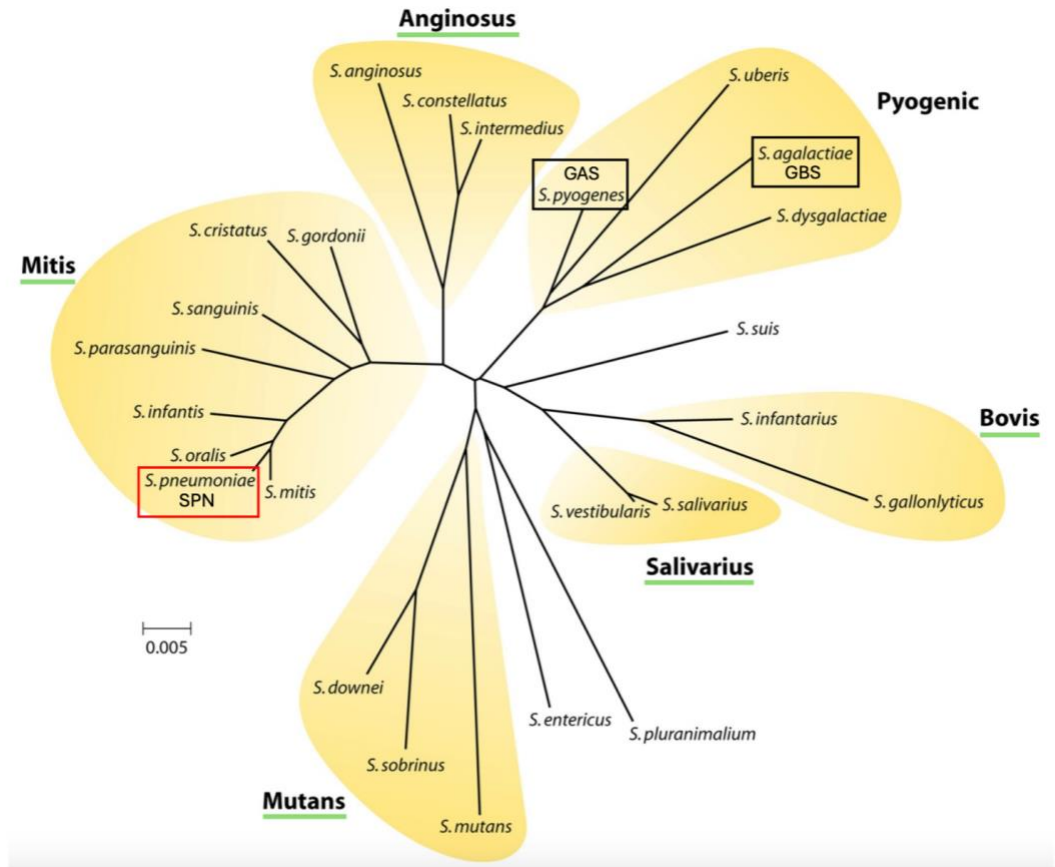


Figure 1.1: Phylogenetic tree representing the six major species groups of *Streptococcus* based on a comparison of the 16S rRNA gene sequence (Kilian, 2010, Nobbs et al., 2009). The three major pathogens of the *Streptococcus* genus are spread across two species groups. *S. agalactiae* (GBS) and *S. pyogenes* (GAS) belong to the pyogenic group (black rectangles) while *S. pneumoniae* (Spn) is a member of the mitis group (red rectangle). Members of the viridans group are underlined in green.

1.2 *Streptococcus pneumoniae* Clinical Microbiology and Diagnostics.

Streptococcus pneumoniae (Spn, or pneumococcus) is a gram positive opportunistic pathogen and a leading cause of bacterial pneumonia, meningitis, and sepsis worldwide (O'Brien et al., 2009, Drijkoningen and Rohde, 2014). Several assays have been traditionally employed (either alone or in combination) for identification of pneumococcus

including colony morphology, hemolysis on sheep blood agar, optochin susceptibility, bile solubility, and reaction with specific antisera (Kellogg et al., 2001). Of the various methods available, the optochin susceptibility test and the bile solubility reaction are the most common (Reller et al., 2008). Optochin, or ethylhydrocuprein hydrochloride, is a derivative of ethylhydrocuprein which was recognized for its inhibitory effects on pneumococcus since it was first synthesized in the early 20th century. Though for solubility and tolerance reasons ethylhydrocuprein was never used as a human therapeutic, its specific antimicrobial effects on Spn did lead to the eventual use of optochin susceptibility as a laboratory diagnostic for *S. pneumoniae* (Burckhardt et al., 2017). The bile solubility test assesses whether a microbial organism will undergo lysis in the presence of the bile salt sodium deoxycholate (Reller et al., 2008). Bile salts like sodium deoxycholate and sodium taurocholate induce lysis in pneumococcus by activating its production of autolysin (Shahsikala, 2008). The Optochin and bile solubility reactions are necessary to distinguish Spn from other viridans group species because *S. pneumoniae* isolates are typically optochin susceptible and are bile soluble, whereas other viridans streptococci are typically optochin resistant and are bile insoluble (Reller et al., 2008).

1.3 *S. pneumoniae* Life Cycle and epidemiological modeling

Although it is considered a major human pathogen, in most cases Spn colonizes the nasopharynx (Figure 1.2) asymptotically (defined as carriage), and is cleared by the immune system without causing harm (Henriques-Normark and Tuomanen, 2013). The length of asymptomatic colonization, referred to as carriage duration, is primarily mediated by a polysaccharide capsule (PS) (Lees et al., 2017). By employing an array of virulence factors including the PS capsule, Spn colonization can sometimes lead to illness

(Henriques-Normark and Tuomanen, 2013). In the event of illness, Spn will usually cause non-invasive diseases like otitis media and non-bacteremic pneumonia, but, on rare occasions, can also spread into normally sterile sites (e.g., blood, cerebrospinal fluid, pleural fluid, peritoneal fluid) and cause severe infections known as invasive pneumococcal disease (IPD) (Henriques-Normark and Tuomanen, 2013).

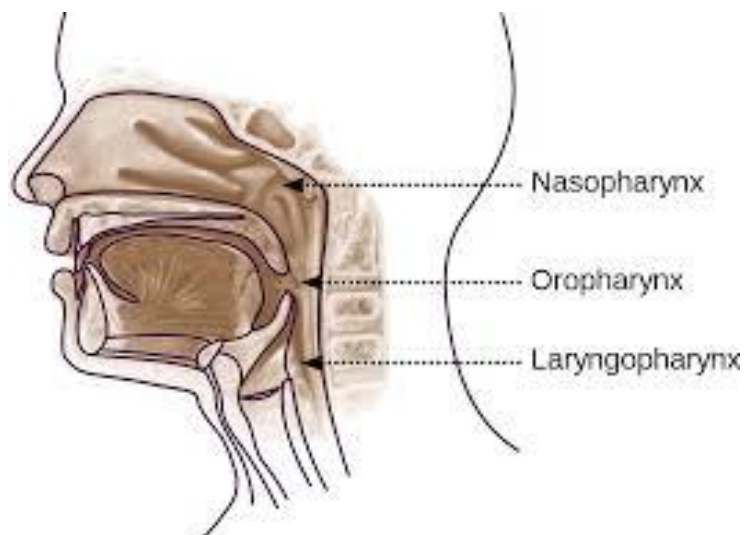


Figure 1.2: Schematic of the human nasal cavity, illustrating the location of Spn carriage in the nasopharynx.

We can employ epidemiological compartmental models to represent pneumococcal carriage and disease within a population. In this model framework, human host individuals are classified as being either susceptible, infected or recovered (immune), with regard to a specific, focal strain. Given the important distinction between carriage and disease states discussed above, the infected class is often further broken down into a carrier state and an invasive disease state. The proportions of individuals in a susceptible, carrier, invasive and recovered class for a focal strain can then be denoted by the variables S , C , I and R

respectively, and solid arrows represent flows of individuals between states (Figure 1.3). In any specific epidemiological model, the functional forms and parameters governing transitions between compartments are more precisely defined by model equations, most often by coupled ordinary differential equations (Anderson and May, 1991).

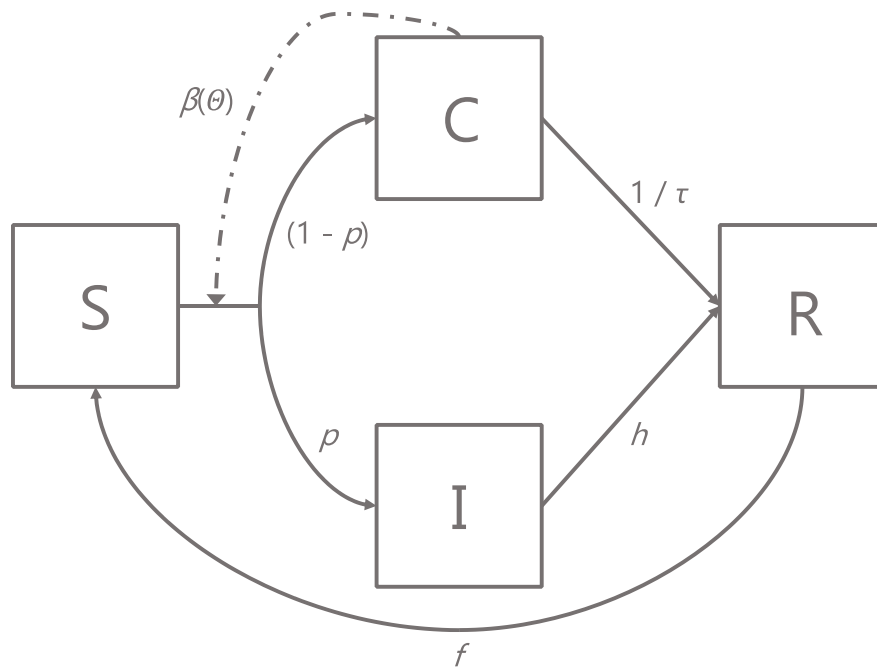


Figure 1.3: Schematic diagram of Spn epidemiological models. Boxes represent proportions of hosts in mutually exclusive states: susceptible (*S*), infected asymptomatic carriers (*C*), invasive (*I*) or recovered and immune (*R*). Solid arrows represent flows of individuals between states, and dashed arrows represent factors influencing those flows. Parameters regulating the transition between each compartment are shown next to each arrow (e.g. *p*, *h*, *f*).

1.4 *S. pneumoniae* Virulence Factors

Perhaps the most important virulence factor within the Spn repertoire is the polysaccharide (PS) capsule. The PS capsule protects the bacterium by limiting its

interaction with the host immune system and avoiding complement-mediated opsonophagocytosis (Lees et al., 2017, Yother, 2011). The chemical structural and serologic characteristics of the PS capsule are determined by both the saccharide repeating units that comprise the PS as well as how they are linked together (Geno et al., 2015, van Selm et al., 2003). Genes that encode the enzymes used to produce the PS capsule reside together in a cluster called the capsular polysaccharide biosynthesis (*cps*) locus. Significant genetic variation within this site, driven by homologous recombination, has enabled *Spn* to generate a wide variety of PS capsules (Mostowy et al., 2017). The antigenic properties of these diverse polysaccharide capsules also provide a way to characterize *S. pneumoniae* through capsular serotyping tests. To date over 90 unique PS capsules, called serotypes, have been identified although only a limited number of them are known to cause serious illness (Hausdorff et al., 2005). Both the ability to cause invasive disease (*i.e.*, invasiveness) and carriage duration have been shown to vary across pneumococcal serotypes (Sleeman et al., 2006).

Another key virulence factor enabling *S. pneumoniae* pathogenicity is a secreted cytotoxin called pneumolysin (PLY). PLY belongs to a family of pore-forming toxins that are expressed during the late log phase of growth (Mitchell and Mitchell, 2010). The toxin first binds to membrane cholesterol on host cells and then oligomerizes as rings. Following a structural change, the pores perforate the host cell membrane, leading to cell lysis (Vögele et al., 2019). Pneumolysin can also manipulate innate immune response by generating both innate pro- and anti-inflammatory effects (Marriott et al., 2008, Subramanian et al., 2019).

They are also several other virulence factors besides capsule and pneumolysin that play important roles in *Spn* pathogenicity. Among the best characterized accessory

virulence factors in the pneumococcal genome are the two pili organelles encoded by pilus islet 1 (PI-1) and pilus islet 2 (PI-2). These long polymeric structures aid in long-distance cell adhesion and play an important role in host-pathogen interaction (Hilleringmann et al., 2015). Another important set of cell wall associated virulence factors are the family of choline binding proteins (cbp) which use a choline-binding module to anchor themselves to the cell wall through non-covalent binding of phosphorylcholine (Maestro and Sanz, 2016). These proteins are components of both the core and accessory Spn genome and participate in several cellular functions including proteolytic activities, immune evasion, and cell division (Vilhena et al., 2023, Donati et al., 2010). Histidine triad proteins (Pht) are another class of surface factors that are found among both the core and accessory genomes of *S. pneumoniae* (Donati et al., 2010). Histidine triad proteins are identified by their histidine triad motif (HxxHxH) that is repeated multiple times within the polypeptide and is predicted to bind divalent metal cations (Bersch et al., 2013). These surface proteins have been associated with several functions within the cell including metal scavenging, complement inhibition and adherence to epithelial cells.

Pneumococcal invasiveness varies across both serotypes and Spn lineages (Sleeman et al., 2006, Brueggemann et al., 2004). Why is it that some strains are more likely to cause invasive disease remains an open question. Much of the research on this question has focused, quite naturally, on variation in virulence factors across Spn serotypes. Investigators have found that invasiveness is positively correlated with the amount of human factor H (FH) bound to the choline binding protein PspC and the histidine triad protein PhtD on the Spn capsular surface (Hyams et al., 2013). Factor H is a host inhibitor of the alternative pathway (a component of the complement system), so by binding FH

pneumococcus can evade complement deposition and neutrophil phagocytosis. Given the key role complement plays in preventing sepsis, FH binding may explain some of the variation in serotype invasiveness. Studies have shown that Spn lineages also carry a significant amount of variation in their accessory genome and that some of these mobile elements contain virulence factors (Mitchell and Mitchell, 2010, Obolski et al., 2018). However, it has been difficult to correlate this pneumococcal variation with the observed disparity in serotype invasiveness (Donkor et al., 2012, Obert et al., 2006). For example, a study of 47 *S. pneumoniae* clinical isolates found very few accessory regions with any notable correlation with high invasiveness serotypes (Blomberg et al., 2009).

1.5 United States Surveillance of IPD

Given the array of virulence factors Spn employs to cause disease and the fact that the risk of IPD varies across serotypes and lineages, it is critical to understand what strains are causing serious infections in a population and what communities are most at risk. To accomplish this task, the CDC is a part of the Active Bacterial Core surveillance (ABCs) program. The ABCs program is an active laboratory- and population-based surveillance system that tracks invasive disease caused by several bacterial pathogens including *S. pneumoniae*. The total population for pneumococcal ABCs surveillance is approximately 34.6 million. The catchment area (Figure 1.4) is spread across 10 US states including San Francisco Bay Area, California (CA) (3 counties), Denver, Colorado (CO) (5 counties), Atlanta, Georgia (GA) (20 counties), Portland, Oregon (OR) (3 counties), Rochester and Albany, New York (NY) (16 counties), select urban counties in Tennessee (TN) (20 counties), Baltimore area, Maryland (MD) (6 counties); and the entire states of Connecticut (CT), Minnesota (MN), and New Mexico (NM)). An IPD case is defined as the isolation

of *S. pneumoniae* from a normally sterile site and is accompanied by a detailed case report form that captures patient demographic and clinical information. When a pneumococcal invasive case is identified, the causal strain is isolated, further cultured, and sent to the StrepLab for strain characterization (e.g., serotyping, multilocus sequence typing, antimicrobial resistance typing). In the past this would entail running a series of lab tests but since 2015 each case isolate is whole genome sequenced and run through our strain typing bioinformatics pipeline (Metcalf et al., 2016b, Metcalf et al., 2016a). Multiple aliquots of each isolate are then sent to a facility in Georgia for storage and 24-hour monitoring. Currently, the ABCs biobank contains 71370 independent Spn isolates spanning 24 years from 1998 to 2022 including 17,167 isolates collected since 2015 that have been whole genome sequenced.

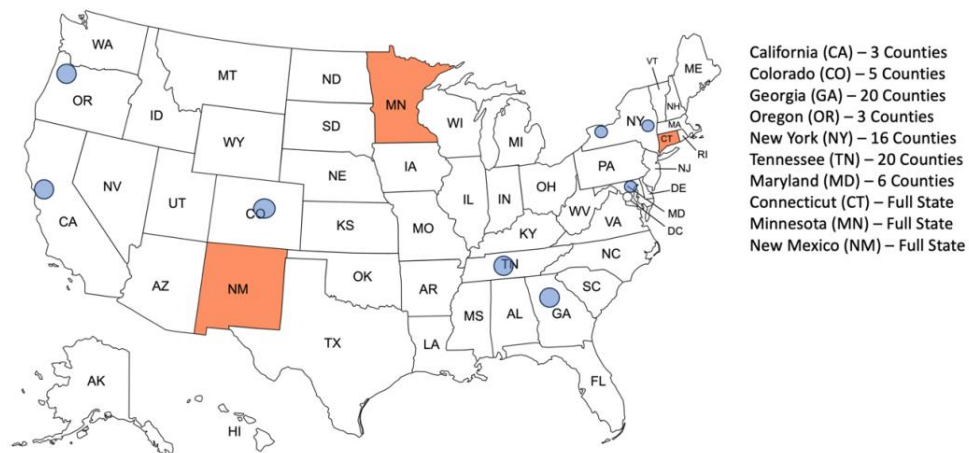


Figure 1.4: The Active Bacterial Core surveillance catchment area. The blue areas are counties spread across 7 US states including San Francisco Bay Area, California

(CA) (3 counties), Denver, Colorado (CO) (5 counties), Atlanta, Georgia (GA) (20 counties), Portland, Oregon (OR) (3 counties), Rochester and Albany, New York (NY) (16 counties), select urban counties in Tennessee (TN) (20 counties), and Baltimore area, Maryland (MD) (6 counties). The orange areas represent the three U.S. states Connecticut (CT), Minnesota (MN), and New Mexico (NM).

1.6 Synopsis

The focus of this thesis is to understand how the interrelationships between bacterial life-history traits and social determinants of health can impact aspects of pneumococcal disease as diverse as how we measure Spn invasiveness (Chapter Two) to how it causes IPD within at-risk populations living in underserved geographical areas (Chapters Three - Five). In the following synopsis I briefly outline the contents of the chapters to come.

1.6.1 Chapter Two

Streptococcus pneumoniae (Spn) is an opportunistic pathogen that, while usually carried asymptotically, 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 in Chapter Two 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.

1.6.2 Chapter Three

Infectious diseases are generally dependent on social interactions among their hosts, with infection requiring susceptible individuals to encounter infected individuals. Intuitively, the per-capita likelihood of infection in a susceptible individual will depend on 1) the density of hosts (more individuals mean more social contacts), and 2) the frequency of infected individuals in the population (higher infection prevalence means more risk of infection per contact). In mathematical models of infectious disease dynamics, these two dependencies have become associated with two limit-case models, termed ‘density-dependent transmission’ (more common in wildlife disease models) and ‘frequency-dependent transmission’ (the norm in human infection studies). In Chapter Three we use extensive nationwide CDC surveillance data to test for density-dependent transmission in the pneumococcus. We show that 1) Spn transmission is dependent on host population density and 2) serotypes with shorter carriage durations and higher invasiveness are less likely to transmit in less populated areas. These results have important implications for pneumococcal epidemiology because carriage duration is a key life-history trait that impacts many other aspects of Spn biology including antimicrobial resistance, recombination, invasiveness, and serotype prevalence.

1.6.3 Chapter Four

Invasive pneumococcal disease (IPD) is a rare outcome of pneumococcal acquisition (on the order of 10 cases of IPD per 100,000 carriage acquisitions), yet for reasons poorly understood, cases of IPD are increasingly identified as forming clusters of localized outbreaks. While prior evidence suggests membership within disadvantaged communities (e.g. people experiencing homelessness) are risk factors for IPD and IPD outbreaks, it is unclear if individual or societal factors are driving these associations. It is possible that individual risk factors (e.g. immune deficits) are mediating the correlations between some at-risk subpopulations and IPD. Here, we consider whether there are broader social determinants that are also driving the relationships between IPD clusters and disadvantaged populations. In light of the Chapter Three finding of density-dependent transmission favoring short carriage duration serotypes in crowded environments, we hypothesize that disadvantaged populations are at increased risk of IPD and IPD clusters via living and working in crowded and inadequately ventilated social surroundings, that enrich the most dangerous short carriage duration serotypes. Using a cluster detection tool to scan genomes of 13,798 *Streptococcus pneumoniae* invasive isolates acquired through ABC surveillance, we find that IPD clusters are substantially more common than expected by chance and are most common in short carriage duration and high attack rate serotypes. Using causal analysis tools, we find evidence for social factors associated with crowding to be causal drivers of both IPD and invasive clustering. These interrelationships between social factors and Spn disease suggest that spending extended amounts of time in crowded and inadequately ventilated environments may drive IPD and invasive clustering by acting as a nexus point where more dangerous Spn strains can spread widely and quickly, especially within disadvantaged communities.

1.6.4 Chapter Five

Throughout the three content chapters of the thesis (chapters Two-Four), we provide evidence that 1) the progression to IPD occurs quickly after carriage acquisition, 2) shorter duration, more invasive serotypes are more likely to spread in higher population densities and more likely to cause invasive clustering, and 3) these clusters are more likely to be comprised of PEH/PWID and tend to occur in underserved geographical regions. In chapter 5 we synthesize the results of this thesis and outline a broad hypothesis to guide future work. In light of our results, we hypothesize that people experiencing homelessness (and marginalized communities overall) are at increased risk of IPD and IPD clusters via 1) epidemiological selection for high transmission, high virulence serotypes, and 2) delays in innate immune cell recruitment and diminished innate response. In this concluding chapter I will provide a broader discussion of potential future directions of research, building on the work in this thesis and our integrative hypothesis for the social and bacterial factors driving invasive pneumococcal disease.

CHAPTER 2. VARIATION IN PNEUMOCOCCAL INVASIVENESS METRICS IS DRIVEN BY SEROTYPE CARRIAGE DURATION AND INITIAL RISK OF DISEASE

Adapted from METCALF, B. J., WALDETOFT, K. W., BEALL, B. W. & BROWN, S. P. 2023. Variation in pneumococcal invasiveness metrics is driven by serotype carriage duration and initial risk of disease. Epidemics, 45, 100731.

2.1 Summary

Streptococcus pneumoniae is an opportunistic pathogen that, while usually carried asymptotically, 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.

2.2 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 asymptotically (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 100000 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 to 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.

2.3 Results

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

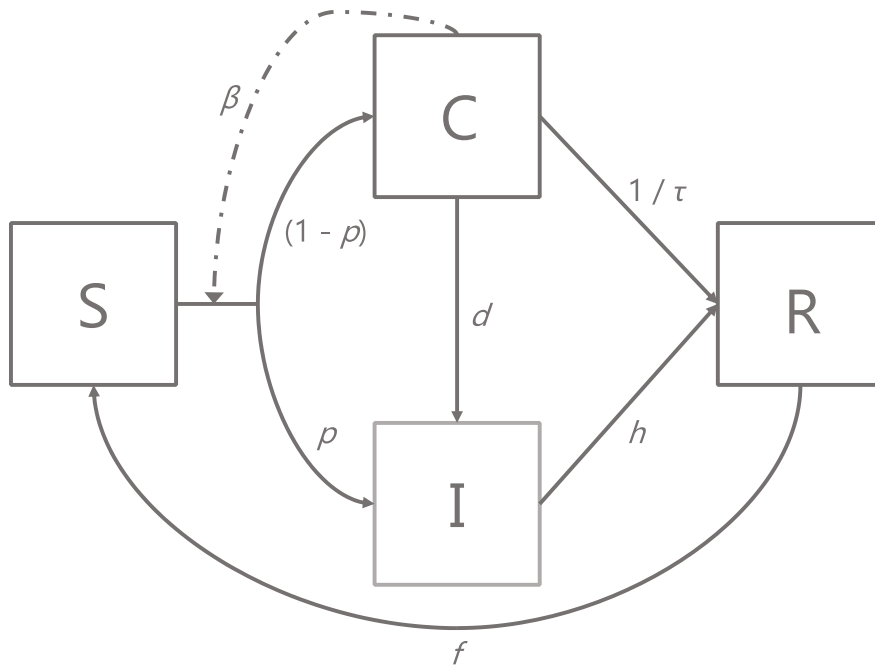


Figure 2.1: Schematic diagram of the epidemiological model. Boxes represent proportions of hosts in mutually exclusive states: susceptible (S), infected asymptomatic carriers (C), invasive (I) or recovered and immune (R). Solid arrows represent flows of individuals between states, and dashed arrows represent factors influencing those flows. Equations describing the system are presented in Materials and Methods (methods equations 2.1), along with parameter definitions (table 2.1). 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.

Table 2.1: 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

Table 2.1 continued

$I(t)$	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.

The susceptible-carrier-invasive-recovered (SCIR) compartmental model used in the following analyses is presented in Figure 2.1 and described in greater detail in Materials and Methods (defined by equations 2.1 and Table 2.1). 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) (equations 2.1) reaches an endemic infection equilibrium, we can derive simple analytical expressions for IOR and AR (see supplementary for derivation), namely that

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

$$\text{AR} = \frac{d\tau+p}{(1-p)(d\tau+1)}. \quad [2.2]$$

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 (equations 2.1), 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)) (Figure 2.2). The AR metric passes this test, as AR is positively associated with both p and d (Figure 2.2A; the gradients $\text{AR}'(p)$ and $\text{AR}'(d)$ 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 (Figure 2.2B; $\text{IOR}'(d) < 0$). In contrast if $\tau_0 < \tau$, then IOR becomes negatively associated with p (Figure 2.2C; $\text{IOR}'(p) < 0$).

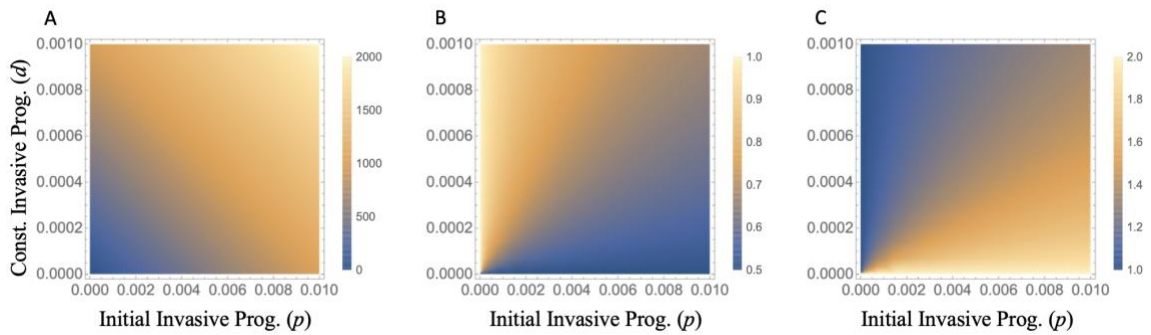


Figure 2.2: Attack rate reliably captures underlying pneumococcal invasiveness parameters while invasive odds ratios fail. (A) Attack rate ($AR = \frac{d\tau+p}{(1-p)(d\tau+1)}$) has a positive relationship with both p and d invasive parameters indicating it accurately represents pneumococcal invasiveness. (B), Invasive odds ratios ($IOR = \frac{\tau_0(d\tau+p)}{\tau(d\tau_0+p)}$) 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 ($\tau_0 = 20$).

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 (Figure 2.3). Indeed, prior studies report a clear negative relationship between serotype IOR and carriage duration τ [5], 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 $IOR'(\tau) < 0$) is possible if and only if $p > 0$ (SI for details). In the case of AR, we find that a negative relationship (*i.e.* $AR'(\tau) < 0$) is not possible for any combination of d and p . The gradient $AR'(\tau)$ is minimized at zero if and only if $d = 0$ (Figure 2.3B, 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 ($\ll 0.5$), $AR = p/(1-p)$ 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 $IOR = \tau_0/\tau$ 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.

2.3.2 *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 (equations 2.2 above) to existing AR and IOR data for serotypes with defined carriage durations (Figure 2.3). 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 to 4.6×10^{-4}) and $d = 0.0$ (CI: 0 to 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 $\text{IOR} = \tau_0/\tau$, and $\text{AR} = p/(1-p)$, see fitted lines in Figure 2.3. While the best fit model found no influence of carriage duration on attack rate (Figure 2.3B) 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.

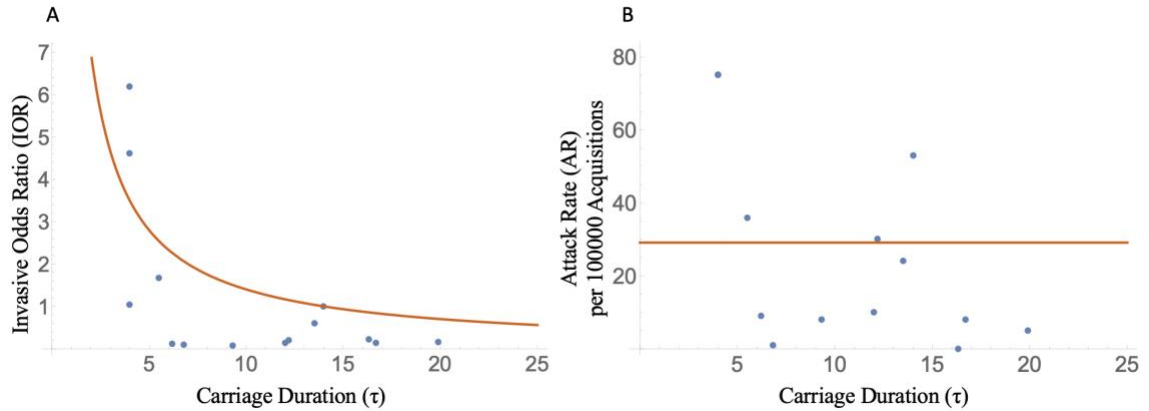


Figure 2.3: Both cross-sectional and longitudinal epidemiological data support the initial risk model and highlight that IOR is confounded by carriage duration. (A) IOR data (blue dots, Brueggemann *et al.* [5]) and model fit (orange line, τ_0/τ), against carriage duration (τ). (B) AR data (blue dots, Sleeman *et al.* [4]) and model fit (orange line, $p(1-p)$), against carriage duration (τ). Simultaneously fitting equations [2.2] 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, 7F, 4, 38, 18C, 3, 33F, 14, 15B/C, 6A, 23F, 6B were used in the model fitting (Table A.1).

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 (Figure 2.1; 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 to 4.5 $\times 10^{-4}$) and $d = 0.0$ (CI: 0 to 1.4 $\times 10^{-5}$). In addition, we now

simultaneously estimate transmission ($\beta = 0.24$ (CI: 0.06 to 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 (Figure A.1).

Expanding the data to incorporate additional epidemiological measures did not change our model fits to AR or IOR data (still captured by Figure 2.3, fitted to AR and IOR data only). In Figure 2.4 we assess the full dataset fitted model against the additional epidemiological data on carriage prevalence, carriage acquisition rate and invasive incidence. Together, Figures 2.3 and 2.4 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 (Figure 2.4C). This is also reflected in the attack rate fit shown in Figure 2.3B.

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).

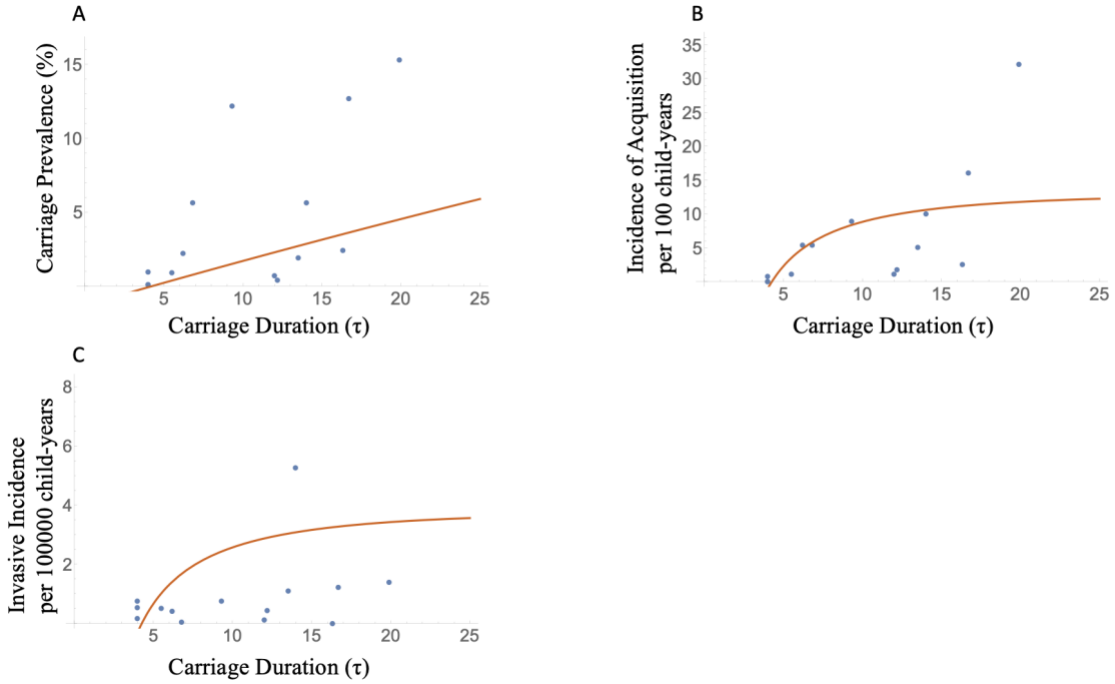


Figure 2.4: Incorporating additional epidemiological data also provides support for the initial risk model. (A), Carriage prevalence data (blue dots, Brueggemann et al. (Brueggemann et al., 2004)) and model fit (orange line, $\frac{fh(b(p-1)\tau+1)}{b(h(f(p-1)\tau-1)-fp)}$), against carriage duration (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 figures 2.3A, 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, 7F, 4, 38, 18C, 3, 33F, 14, 15B/C, 6A, 23F, 6B were used in the model fitting (Table A.1).

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 $d = 0$ (Figure 2.4), an information criterion model comparison approach concludes that the initial risk model ($d = 0$) outperforms the constant risk model (Table A.2).

2.3.3 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 $p \ll 0.5$). 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 $d = 0$, $IC = \frac{p}{\tau - p\tau}$) and can also be corrected for bias by multiplying by τ (Yildirim *et al.*, 2010).

2.4 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, Figure 2.2B, C). Fitting this model to both longitudinal and cross-sectional data (Figure 2.3), 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 $\text{IOR} = \frac{\tau_0}{\tau}$. 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 ($\text{AR} = \frac{p}{1-p}$). Together these results agree with the empirical finding of a significant correlation between IOR and carriage duration (Figure 2.3A), but no correlation between AR and carriage duration (Figure 2.3B, 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, Figure 2.3A). 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. Figure 2.3A illustrates that short-carriage ($\tau = 4$ -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 (Figure 2.3B) 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.

2.5 Materials and Methods

2.5.1 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 (Figure 2.1), and their dynamics are given by the following system of four ordinary differential equations,

$$\frac{dC}{dt} = (1 - p)\beta CS - \left(\frac{1}{\tau} + d\right)C$$

$$\frac{dI}{dt} = p\beta CS - hI + dC$$

$$\frac{dR}{dt} = \frac{1}{\tau}C + hI - fR$$

$$S = 1 - C - R - I \quad [2.1]$$

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 2.1.

2.5.2 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 Figure 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 12A, 12B and 12F 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 19A, 19F, 9A, 9N and 9V 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 A.1. 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 A.1.

2.5.3 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*.

2.6 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.

CHAPTER 3. HUMAN POPULATION DENSITY DRIVES PNEUMOCOCCAL DISEASE RISK VIA DENSITY-DEPENDENT SEROTYPE TRANSMISSION

Adapted from METCALF, B. J., WALDETOFT, K. W., BEALL, B. W. & BROWN, S. P. 2023. Human population density drives pneumococcal disease risk via density-dependent serotype transmission (in prep).

3.1 Summary

In human infectious diseases, the per-capita likelihood of infection in a susceptible individual is generally assumed to be dependent on the frequency of infected individuals in the population (higher infection prevalence means more risk of infection per contact) and not dependent on population density. In Chapter Three we use CDC surveillance data on millions of US citizens to test for density-dependent transmission in the pneumococcus. We show that 1) Spn transmission is dependent on host population density and 2) serotypes with shorter carriage durations and higher invasiveness are less likely to transmit in less populated areas. These results have important implications for pneumococcal epidemiology because carriage duration is a key life-history trait that impacts many other aspects of Spn biology including antimicrobial resistance, recombination, invasiveness, and serotype prevalence.

3.2 Introduction

Infectious diseases are generally dependent on social interactions among their hosts. Contact-mediated infection requires susceptible individuals to encounter infected individuals (Borremans et al., 2017). Intuitively, the per-capita force of infection experienced by a susceptible individual will depend on 1) the density of hosts (more individuals mean more social contacts), and 2) the frequency of infected individuals in the population (higher infection prevalence means more risk of infection per contact). In mathematical models of infectious disease dynamics, these two dependencies have become associated with two limit-case models (Begon et al., 2002), termed ‘density-dependent transmission’ and ‘frequency-dependent transmission’ (Figure 3.1).

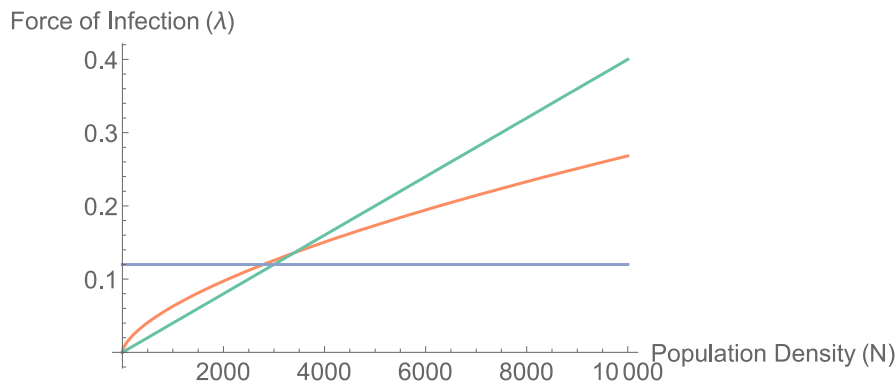


Figure 3.1: Alternate models for the force of infection λ as a function of population density N . The blue line defines a frequency-dependent contact structure ($\lambda = \kappa I$, where I is the frequency of infecteds and κ is a rate constant) that is independent of total population density N . The green line defines a density-dependent contact structure ($\lambda = \kappa NI$, where NI defines the total density of infecteds in the population). The orange line defines a power contact structure ($\lambda = \kappa N^r I$, contact rate = κN^r) and represents a ‘hybrid’ of frequency- and density-dependent limit cases (when $r = 0$ we recover the pure frequency-dependent model, and when $r = 1$ we recover the pure density dependent model).

Under the frequency-dependent limit case, the force of infection λ (the per-capita rate of infection experienced by susceptible individuals) is proportional to the *frequency* of infecteds in the population, I (blue line, Fig. 3.1). For human communicable diseases, it

has been generally assumed that interaction patterns don't scale strongly with density and that a frequency-dependent contact structure is sufficient (Grassly and Fraser, 2008, Ferrari et al., 2011) since whether you live in a small community or a mega-city, class-room sizes and workplaces tend to be on a similar scale. In agreement with this interpretation, an analysis of UK measles time-series data in the pre-vaccination era supported a frequency-dependent model (Bjornstad, 2002).

Under the density-dependent limit case, λ is proportional to the *density* of infected individuals, IN (green line, Figure 3.1), where N is the total population density. Wildlife ecologists have demonstrated the importance of density-dependent transmission effects especially in the context of pathogen persistence in wildlife reservoirs and control of invasive species (Cross et al., 2013, Cross et al., 2010, Berthier et al., 2000). Accounting for host population density has also been an area of interest for other viral respiratory pathogens like influenza and, more recently, SARS-CoV-2 though the results have been mixed. Investigations into the relationship between influenza and population density have yielded inconclusive results (Garrett, 2007, Mills et al., 2004, Nishiura, 2008, Chandra, 2013) while recent investigations into the spread of SARS-CoV-2 have identified an association between transmission and population density (Sy et al., 2021).

The two models described above can be generalized as limit cases of a broader model, for example by introducing a scaling parameter r (orange line, Figure 3.1), allowing the impact of density to plateau at higher densities (Borremans et al., 2017). In support of this model, work by Cross *et al.* found that *Brucella abortus* transmission in elk was dependent on density but that the contact network of elk herds followed a hybrid model that saturated at higher population densities (Cross et al., 2013, Cross et al., 2010).

With respect to our focal bacterial pathogen *Streptococcus pneumoniae*, like other respiratory pathogens, we find that prior epidemiological models overwhelmingly rely on a purely frequency-dependent contact function (Fig 3.1, blue line). In a structured literature review (Supplement), we identified only a single article that assumed density-dependent transmission and that investigation was focused pneumococcal spread within households (Melegaro et al., 2004). Part of the reliance on frequency-dependent transmission may be because parameterizing a density-dependent model is difficult with a bacterial pathogen that has a complex life-history composed of distinct states of carriage and symptomatic infection.

While *Streptococcus pneumoniae* (Spn) can cause illness (O'Brien et al., 2009), it usually colonizes the nasopharynx asymptotically (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 (PS) capsule. Sometimes, especially in older studies, serogroup (a more general category of related but distinct serotypes) is reported instead of serotype. PS capsule is the primary virulence factor of Spn and influences several traits including the duration of carriage, invasiveness (propensity to cause invasive disease), and carriage prevalence in a population (Sleeman et al., 2006, Chaguzza et al., 2020). Over the first half of the 20th century, the majority of Spn adult invasive disease was attributable to serogroups 1, 2, 3, and 5 despite them being rarely identified in carriage (Feikin and Klugman, 2002). This subset is sometimes referred

to as the ‘epidemic’ serogroups due to their propensity to cause outbreaks in crowded environments (Feikin and Klugman, 2002). This association with crowding suggests that some serotypes with low carriage rates may thrive in highly dense and crowded populations. While serogroups 1, 2 and 5 are now largely eradicated thanks to vaccine and antibiotic interventions, there remains substantial variation in carriage duration among currently circulating serotypes (Chapter One, (Sleeman et al., 2006)).

In this investigation, we use extensive CDC surveillance data to test for density-dependent transmission in the pneumococcus. We hypothesize that short carriage duration serotypes can only persist in dense human populations, as they rely on a high frequency of transmission opportunities in their short duration of carriage. We test this hypothesis using IPD data acquired through the CDC Active Bacterial Core surveillance (ABCs) program which conducts active laboratory and population-based surveillance for invasive Spn infections in a population of more than 34.4 million persons. We will show that 1) short carriage duration serotypes are enriched in the densest human populations, and 2) pneumococcal transmission is both density- and frequency-dependent.

3.3 Results

3.3.1 Population Density is Correlated with Serotype Carriage Duration and Invasiveness

To approach our hypothesis that shorter serotype carriage duration is associated with higher human population density, we begin by characterizing the average human population density associated with each Spn serotype, and for Spn as a whole. The data for this analysis was acquired from the Center for Disease Control and Prevention’s ABCs

program. The average host population density per serotype is calculated using either census tract or zip code data associated with each Spn invasive disease case (see Methods for more details). In Fig. 3.2 we see 12-fold variation in serotype-specific human density, ranging from dense-population 7F (average population density of 15,037 people per square mile (PPSM), greater than the density of Washington DC) through to low density 23F (average population density of 1,224 PPSM, less dense than an average suburban neighborhood (Airgood-Obrycki and Rieger, 2019)). The 12-fold variation in host population density was significant when using the ANOVA test to compare serotypes ($F(55, 9978) = 9.375$, $p\text{-value} = 2e-16$). Across all census tracts with an IPD case, the average population density is 4,502 PPSM, which is not significantly different from the mean population density of 4,600 PPSM across the total surveillance population (Kolmogorov-Smirnov test, $p\text{-value} = 0.974$).

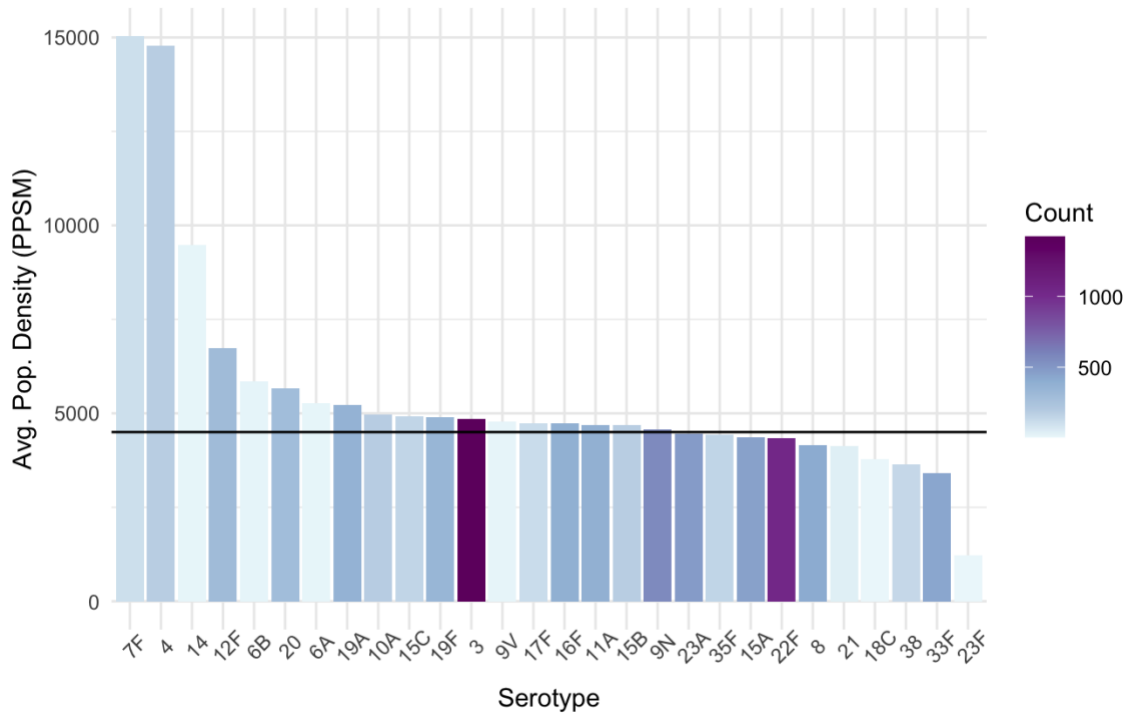


Figure 3.2: Average serotype population density varies widely across serotypes. Bar chart of average host population density by serotype where population density is measured in persons per square mile (PPSM). The bar color encodes the number of IPD cases for each serotype with the darker color representing higher case counts. Horizontal black line: average population density across all census tracts with IPD cases regardless of serotype.

Given the variation in average population density we see across serotypes, we now ask whether there are bacterial attributes associated with this variation. We focus on two life-history traits, carriage duration and invasiveness, as they have been shown to be serotype-dependent (Sleeman et al., 2006). We first look at the association between the average population density of serotypes acquired from invasive cases with their respective carriage durations (Fig. 3.3A). A significant inverse correlation between average population density and carriage duration is evident whether we use census tracts (Kendall Tau = -0.348, p-value = 0.00961; Figure 3.3A) or ZIP codes (Kendall Tau = -0.385, p-value = 0.00415; Fig. B.1A) as geographical units of measure. We also examined the

relationship between population density and serotype invasive attack rate (Fig. 3.3B). Here we found a positive relationship that was significant for both census tracts (Kendall Tau = 0.328, p-value = 0.011; Figure 3.3B) and ZIP codes (Kendall Tau = 0.256, p-value = 0.0476; Figure B.1B) as geographical units of measure. These relationships indicate that shorter duration and more highly invasive serotypes are more likely to cause IPD in more dense environments.

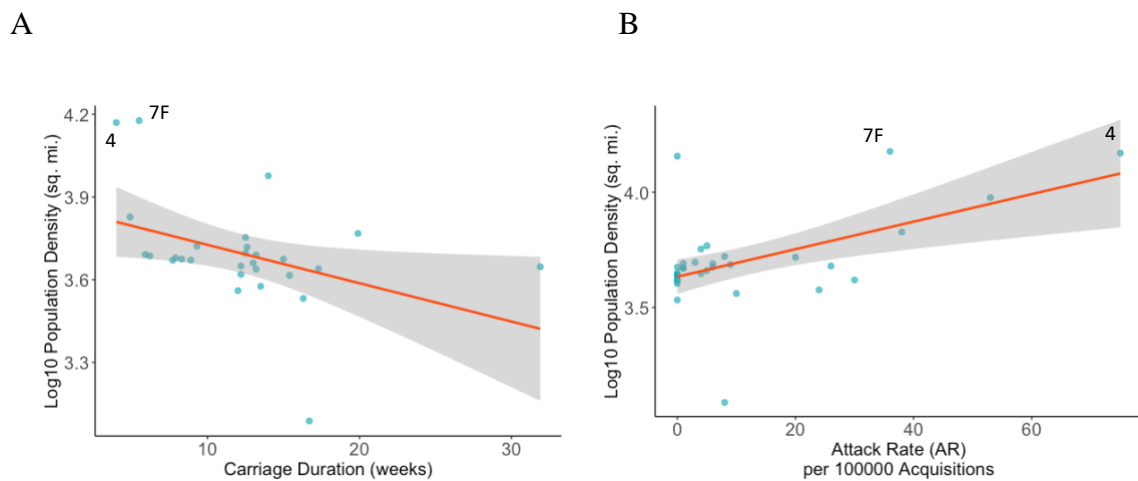


Figure 3.3: Average serotype population density has a negative correlation with carriage duration and a positive correlation with invasive attack rate (AR). (A) Serotype carriage duration (blue dots) against log₁₀ average serotype population density (measured using census tracts) yields a significant inverse association (Kendall Tau = -0.348, p-value = 0.00961). (B) Invasive AR (blue dots) against log₁₀ average serotype population density (measured using census tracts) yields a significant positive association (Kendall Tau = 0.328, p-value = 0.011). Two examples of short duration, high invasive serotypes (4 and 7F) are labeled in each plot.

3.3.2 *Pneumococcal transmission is both frequency- and density-dependent*

The results in Fig 3.3 support our hypothesis that shorter carriage duration strains (e.g. 4 and 7F) are selected for in more dense environments, and more generally indicate the presence of at least partially density-dependent transmission in the pneumococcus. In

order to quantify the extent of density-dependence, we now turn to an explicit model-based approach.

We begin with the compartmental epidemiological model developed in Chapter Two (Metcalf et al., 2023), tracking the prevalence of susceptible, carriage, invasive disease and recovered cases (S , C , I , R , respectively, see methods). On this parameterized baseline, we modify the transmission term to allow for intermediate degrees of density versus frequency dependence. Specifically, we follow two distinct forms of interpolation between the frequency- and density-dependent limit cases described in Figure 3.1. Note that both interpolation models allow either limit case (pure frequency- or density-dependence) to be supported by data. First we assess the established power-law interpolation (transmission coefficient $\beta(b,r) = bN^r$, orange line in Figure 3.1, (Borremans et al., 2017). Second, we assess a piecewise function $\beta(b_D) = \begin{cases} b_D N, & \tau < \kappa \\ b_D N_\kappa, & \tau \geq \kappa \end{cases}$

where τ is serotype carriage duration and κ is a threshold carriage duration that marks the transition point from density- to frequency-dependent transmission. The piecewise cutoff threshold κ would naturally be defined in terms of population density instead of carriage duration but is required to be in terms of carriage duration as a constraint of our model fitting procedure (see methods). In the methods we derive a relationship between carriage duration τ and population density N (methods equation 3.8) that allows us to represent the cutoff threshold κ in terms of population density (parameter N_κ). At population density N_κ , the transmission coefficient β transitions to frequency-dependence where β is fixed at the threshold value ($b_D N_\kappa$) and invariant across higher population densities.

Along with population density, the IPD case data acquired from ABC surveillance also allows us to calculate invasive incidence for each serotype. In the methods we derive a mathematical relationship between carriage duration τ and invasive incidence (methods equations 3.7). Functions for both population density and invasive incidence are then used for model fitting and parameter estimation (see methods).

In addition to fitting the relevant force of infection parameters (b and r for the power law function, and b_D , and κ for the piecewise function), we also fit the duration of immune protection f . All other model parameters are obtained from previous work (see methods). Parameter fits under the two force of infection models are presented in Table 3.1, and the corresponding transmission coefficient β as a function of density is presented in Fig. 3.4.

Table 3.1: Fitted coefficient values for parameters used in the power law and piecewise contact rate models.

Power Law Contact Rate Model		Piecewise Contact Rate Model	
Parameter	Fitted Coefficient Value (95%CI)	Parameter	Fitted Coefficient Value (95%CI)
r	0.988 (0.709 - 1.27)	κ	11.07 (8.52 - 13.62)
b	7.69×10^{-5} (-8.99×10^{-5} - 2.44×10^{-4})	b_D	6.67×10^{-5} (4.11×10^{-5} - 9.23×10^{-5})
f	236.91 (190 - 284)	f_P	229 (189 - 269)

We estimate the piecewise threshold constant κ to be 11.07 (95%CI: 8.52 to 13.62) meaning that Spn transmission is density-dependent for carriage durations less than 11.07 weeks and frequency-dependent otherwise. Plugging this value into our function for

population density (methods equation 3.8) and using the average per capita invasive incidence of $7.69e-7$ cases/week, we find the population density threshold N_κ is 3722.82 (PPSM). Thus, the frequency-dependent transmission coefficient β , calculated by multiplying N_κ by b_D , is 0.248.

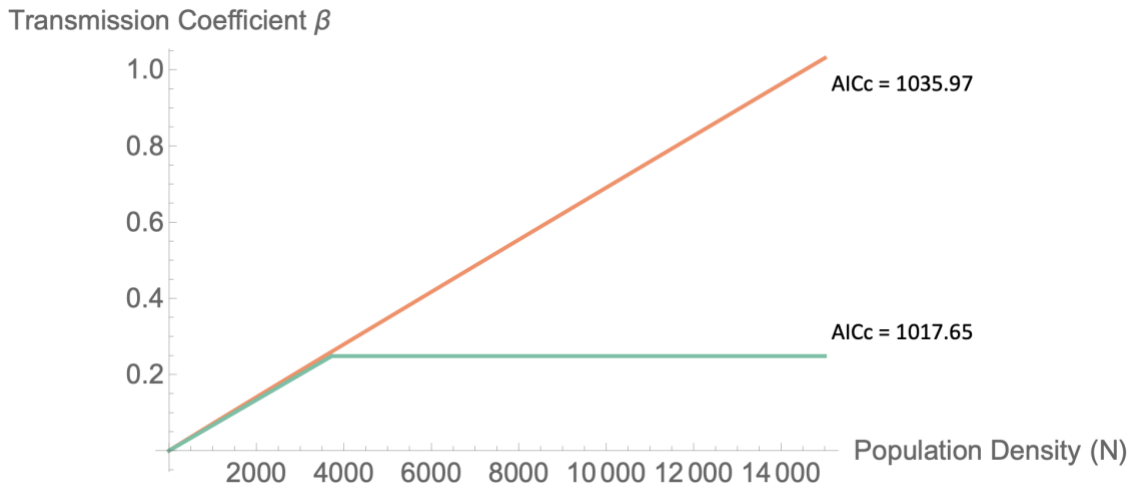


Figure 3.4: Both the fitted power and piecewise contact rate models indicate that Spn transmission is density dependent. The fitted power contact rate model transmission coefficient β (orange line, $\beta(b,r) = bN^r$) where $r = 0.988$ and $b = 7.69 \times 10^{-5}$ against population density N . The fitted piecewise contact rate model transmission coefficient β (green line,

$$\beta(b_D) = \begin{cases} b_D N, & N < 3722.82 \\ b_D N_\kappa, & N \geq 3722.82 \end{cases} \text{ where } b_D = 6.67 \times 10^{-5} \text{ against population density } N.$$

The population density threshold of 3722.82 is calculated by plugging the fitted $\kappa = 11.07$ value into the function for population density (methods equation 3.8) and using the average per capita invasive incidence of $7.69e-7$ cases/week. Both models have a non-constant relationship with population density suggesting that Spn spreads through density-dependent transmission at least in lower population density regions. Model comparisons are made using AICc scores (Burnham et al., 2011). For comparison, the pure density- and frequency-dependent limit cases had AICc scores of 1032.78 and 1222.08 respectively.

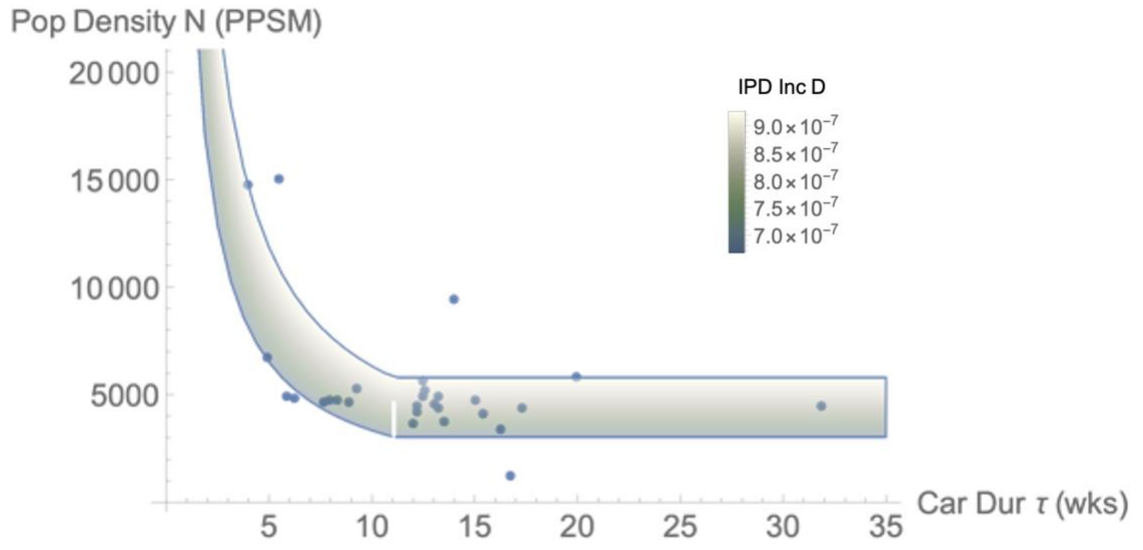
Table 3.1 and Figure 3.4 show that both models support density dependence in Spn.

Among the two model structures, the piecewise model has greater data support (delta AICc

= 18.32, implying a relative likelihood of 9509.06 (Burnham et al., 2011)). For comparison, setting the power law exponent r to 0 and fitting a purely frequency-dependent transmission model, we find substantially weaker support (delta AICc = 1222.08, relative likelihood of $2.46e44$), allowing us to reject the standard epidemiological assumption of no density dependence.

The piecewise model suggests that Spn transmission is dependent on density below a critical density threshold of 3722.82 PPSM, whereas above this threshold a purely frequency-dependent behavior is supported. Plots of the fitted population density and invasive incidence equations for the piecewise contact rate model are provided in Figures 3.5 and B.2 respectively. The plot of invasive incidence shown in Figure B.2 also provides an opportunity to check for the internal consistency of the piecewise model fit. If our model design is correct, then when we set population density N to be the threshold value N_{κ} for the invasive incidence function (equation 3.7), the frequency- and density-dependent transmission terms β will be identical. Thus, the two individual functions should connect at the piecewise threshold point and create a continuous curve as shown in Figure B.2. Bear in mind the curves will connect only when we set population density to N_{κ} . Otherwise, the density-dependent curve of the piecewise model will be either above or below the frequency-dependent function. Plots of the power law model are shown in Figure B.3.

A



B

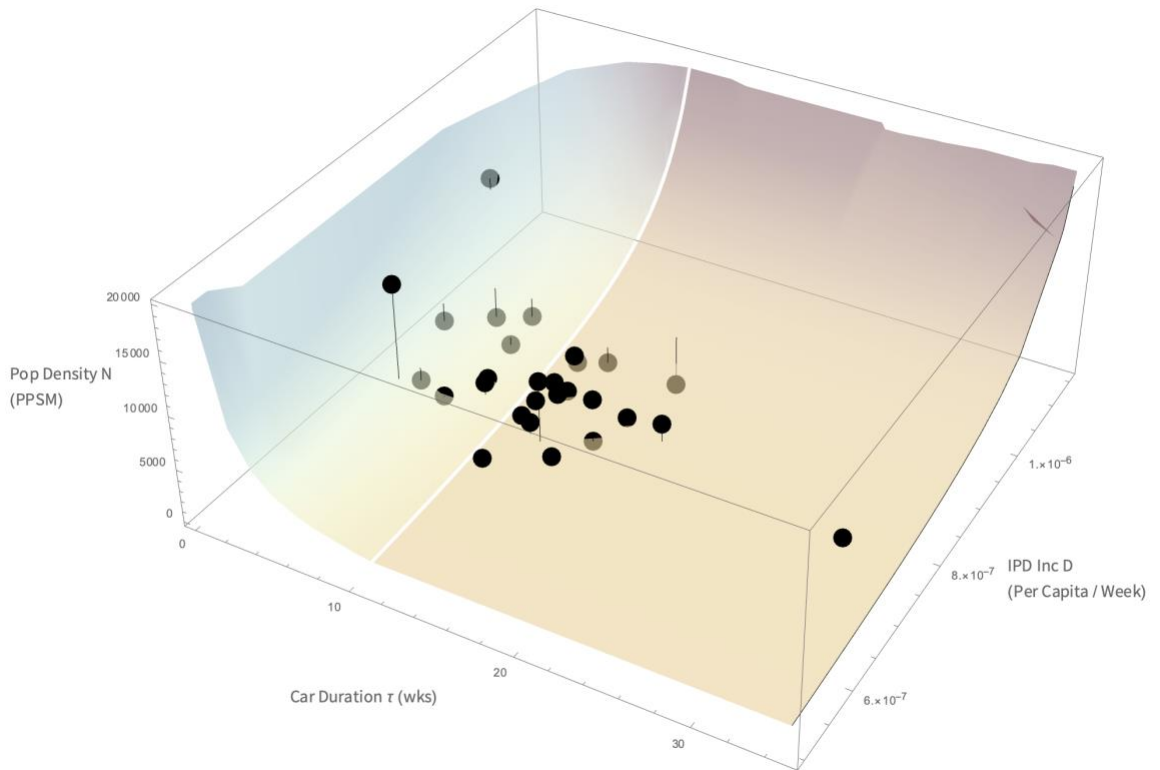


Figure 3.5: Fitting population density and invasive incidence using a power contact rate function indicates Spn transmission is density dependent. A) Spn serotype average population density data (blue dots) and the model fit (green-to-white gradient,

$$N_P(\tau, D^*) = \begin{cases} N_D P(\tau, D^*), & \tau < \kappa \\ N_D P(\kappa, D^*), & \tau \geq \kappa \end{cases} \text{ where } N_D P(\tau, D^*) = \frac{hp}{b_D(p-1)\tau(-fhD+h(Dp\tau-D\tau+p)-Dp)}$$

) of population density as a function of carriage duration τ and equilibrium invasive incidence D^* . Model fits support the hypothesis Spn transmission is density-dependent and that shorter-duration serotypes are more likely to cause invasive disease in higher-density areas. The model fitting produced the parameter estimates $b_D = 6.67 \times 10^{-5}$, $\kappa = 11.07$, and $f_p = 229$. B) A 3D plot of population density as a function of carriage duration τ and invasive incidence D^* . The vertical lines connecting each datapoint to the fitted function represent the residual error. Additional fixed parameter values are invasive clearance $h = 0.5$ and probability of invasion $p = 0.00029$.

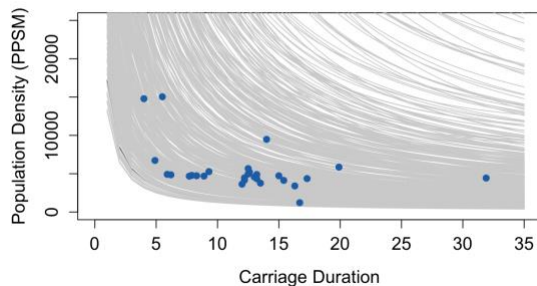
3.3.3 Sensitivity Analysis

While both the power and piecewise contact rate models show that Spn transmission is density-dependent, they differ in the strength of this dependence. The power contact rate model suggests nearly pure ($r = 1$) density-dependence while the piecewise model supports a hybrid transmission function that plateaus at higher densities. Each model has several parameters controlling the level of density-dependence that we fitted in this analysis. We now assess how robust these model results are to varying these parameters by carrying out two sensitivity analyses. First, to get an overall sense of parameter sensitivity we generated a parameter combination plot (Figure 3.6A, C) by sampling 500 datapoints for each estimated parameter from the distributions listed in methods Table 3.3. Each parameter combination was plotted as a grey line and all 500 plots were overlaid with the empirical observations in blue. The monotonically decreasing relationship between carriage duration and population density holds for all 500 curves for

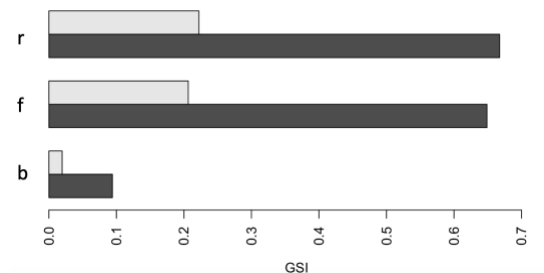
both the power contact rate model (Figure 3.6A) and the piecewise contact rate model (Figure 3.6C).

We next evaluate the relative impact of each parameter and whether their effects are stable across carriage durations. To do this we perform a sequential sensitivity analysis using the ‘multisensi’ R package with carriage duration acting as the sequential variable. The parameter ranges used are the 95% confidence interval ranges listed in methods Table 3.3. For the power contact rate model, Figure 3.6B shows the generalized sensitivity indices (GSI) for each parameter and reveals that the power contact rate function exponent r and the immune duration parameter f had the most impact on the model with the b parameter having relatively less influence. A graph of the variation in sensitivity indices across carriage duration indicates that the main effects are stable across serotype durations (Figure B.4A). The piecewise contact rate model was most sensitive to the f_P and b_D parameters and least impacted by the threshold parameter κ (Figure 3.6D). However, looking at the variation of the indices across carriage duration (Figure B.4B) shows the impact of κ increases when the model transitions from density- to frequency-dependent transmission.

A



B



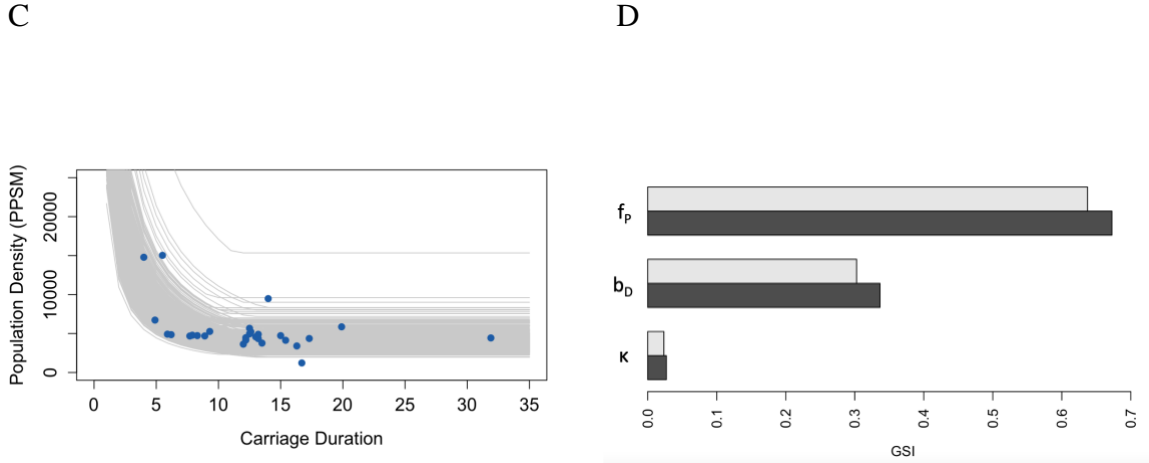


Figure 3.6: Sensitivity analysis of the population density functions (equations 3.4 and 3.9) reveals that the model fitting results are robust across a wide range of parameter values. A) and C) Parameter combination plots of equations 3.4 (Figure 3.6A) and 3.9 (Figure 3.6C) (see methods) generated by sampling 500 sets of parameter estimates from the distributions described in Table 3.3 (grey lines) overlaid with Spn serotype average population density data (blue dots). Both graphs show that even with wide variation in parameter values the overall negative association between carriage duration and population density generally holds. B) First order (grey bars) and total (black bars) generalized sensitivity indices (GSI) for the population density function based on the power contact rate model (equation 3.4). Results show that population density is mainly influenced by changes in the power exponent parameter r and immune protection parameter f with a smaller contribution from the transmission rate constant b . D) First order (grey bars) and total (black bars) generalized sensitivity indices (GSI) for the population density function based on the piecewise contact rate model (equation 3.9). The model was most sensitive to changes in the immune protection parameter f_p and the density-dependent transmission constant b_D .

3.4 Discussion

In this investigation we examined the role of human population density in Spn transmission, and tested the hypothesis that short duration serotypes are attenuated in lower density areas, given they have less time to successfully transmit to a new host as compared to their longer duration counterparts. In agreement with our hypothesis, we show that (i) Spn serotype distributions are structured by host population density (Figure 3.2), (ii) human density correlates with serotype carriage duration and attack rate (Figures 3.3, 3.5), and

(iii) serotype transmission is both frequency- and density-dependent (Figure 3.4). Together these results show that serotypes are differentially impacted by human population density, with short carriage duration serotypes less likely to transmit (and cause invasive disease) in less populated areas.

These results have important and generally overlooked implications for pneumococcal epidemiology, given the substantial variation in human population density around the world, and the connections between carriage duration and other traits. Carriage duration is a key life-history trait that impacts many other aspects of *Spn* biology including antimicrobial resistance, recombination, invasiveness, and serotype prevalence (Chaguza et al., 2016, Kalizang'oma et al., 2021, Sleeman et al., 2006, Li et al., 2013). For example, longer duration serotypes tend to have higher carriage prevalence within a community (Li et al., 2013, Weinberger et al., 2009, Högberg et al., 2007), although these prior studies did not unpack the role of variation in human population density. Our results indicate this advantage can be attenuated in regions with high population densities that allow greater transmission of shorter duration serotypes.

Our results imply that pneumococcal surveillance systems that operate across different distributions of population densities may report different distributions of serotypes even if the underlying serotype diversity is the same. This may play a role in the observed geographical variation of IPD incidence caused by non-vaccine serotypes following the introduction of pneumococcal conjugate vaccine (Lewnard and Hanage, 2019). Also, previous work has identified a positive correlation between beta-lactam resistance and carriage duration, likely due to increased antibiotic exposure in strains with longer human residency times and increased opportunity for recombination with beta-

lactam resistant *Streptococcus mitis* (Lehtinen et al., 2017, Kalizang'oma et al., 2021). Our investigation would therefore suggest that beta-lactam resistance may be more likely to occur in less populated areas. An important area of future work is to determine if beta-lactam resistance is inversely correlated population density and to tease apart how beta-lactam antibiotic prescription rates influences this potential relationship.

More broadly, our results offer a route-map for investigation of density-dependent transmission in other human pathogens with complex carriage versus invasive epidemiological dynamics, as seen in various opportunistic bacterial pathogens (e.g. meningococcus, group A *Streptococcus*). The introduction of density-dependence to bacterial epidemiological models also raises new opportunities to revisit classical epidemiological questions in light of density-dependent effects, including critical community sizes (Finkenstädt et al., 1998, Bartlett, 1960), vaccine thresholds (Anderson and May, 1991), and the evolution of virulence (Alizon et al., 2009).

This study has several limitations. The average census tract population density for the ABC catchment area (4600 PPSM) is considerably higher than the average population density of the United States population (94 PPSM; (Bureau, 2021)), reflecting a surveillance focus on more densely-populated regions of the country. Our results indicate that this surveillance choice may bias the distributions of serotypes (and potentially other pathogens) that are reported. Also, census tract information for our surveillance dataset was missing for 1,506 of 13,798 case-patients (11% of total). For parameter estimation, the quality of the model fits varied widely depending on the algorithms and arguments used. We evaluated several global and local optimization procedures (e.g. NMinimize, Newton, InteriorPoint) using AICc, but our comparisons were not exhaustive and better parameter

estimations may be possible. The Spn carriage duration and invasive attack rate data used in this analysis are the most comprehensive available but were acquired more than 15 years ago as part of a series of longitudinal pediatric studies in the United Kingdom (Sleeman et al., 2006) and may not capture unique features of current adult invasive disease epidemiology (Alanee et al., 2007).

Our analysis shows that more invasive short carriage duration serotypes will spread more easily in communities with higher population densities (Figure 3.3). Importantly, in regions where population or housing density is associated with lower community socioeconomic status, these shorter duration, higher invasive serotypes may be preferentially targeting the very communities that are least equipped to intervene or respond. In the following Chapter Four we examine the implications of density-dependent pneumococcal transmission in the context of the social determinants of health.

3.5 Methods

3.5.1 Study Collection

The CDC Active Bacterial Core surveillance (ABCs) program conducts active laboratory and population-based surveillance for invasive SPN infections (in 10 states (complete states or selected counties), representing more than 34.4 million persons (<https://www.cdc.gov/abcs/reports-findings/survreports/gas18.html>). ABCs defines an invasive SPN case as illness in a surveillance area resident with isolation of SPN from a normally sterile site. Surveillance staff at sites complete a standardized case report form that includes basic demographic characteristics and risk factors of infection on all cases, and coordinate collection and transfer of isolates to CDC's *Streptococcus* lab for

characterization. In total, 11,245 available invasive Spn isolates from cases identified in 2016-2019 were included in the analysis.

We calculated the average population density for serotypes acquired from invasive disease cases that were within the catchment area of the ABC surveillance program from 2016 through 2019. We used data obtained from 2015-2019 5-Year estimates of the American Community Survey (ACS) and U.S. Census Bureau's 2021 Planning Database (PDB) with both census tracts and ZIP Code Tabulation Areas (ZCTAs) as the geographic units of measure. To calculate the average population density for each *S. pneumoniae* serotype acquired through ABC surveillance at the ZCTA level, we used ZCTAs obtained from the 2015- 2019 American Community Survey 5-Year Estimates dataset B01003 to infer population size. Geographic area for each ZCTA was obtained using the 2017 land area data from the Gazetteer file 'ZIP Code Tabulation Areas'. For census tract population density measurements, we used the U.S. Census Bureau's 2021 Planning Database (PDB) which provides both population count and land area data at the census tract level. Population density for each census tract or ZCTA was calculated by dividing the population size by its land area. A population density was associated with each invasive case by linking the ZCTA or census tract population density with the zip code or census tract recorded in the case report form. To ensure the ZCTAs correspond to the correct zip code we validated each pairing using the crosswalk database created by the Census Reporter project (<https://censusreporter.org/>). Serotype invasive incidence was obtained by dividing serotype IPD case counts by total serotype population size. Serotype population size was calculated by identifying the unique set of census tracts where an IPD case was found and summing their individual population sizes.

Carriage duration data was obtained from a study from (Sleeman et al., 2006). We assumed that the carriage duration for serotype 4 was too short to be reliably detected given the sampling times of the longitudinal study, so we imputed a conservative estimate of four weeks (the longest sampling interval used in Sleeman *et al.*) instead (Chapter Two).

3.5.2 Epidemiological Model Description

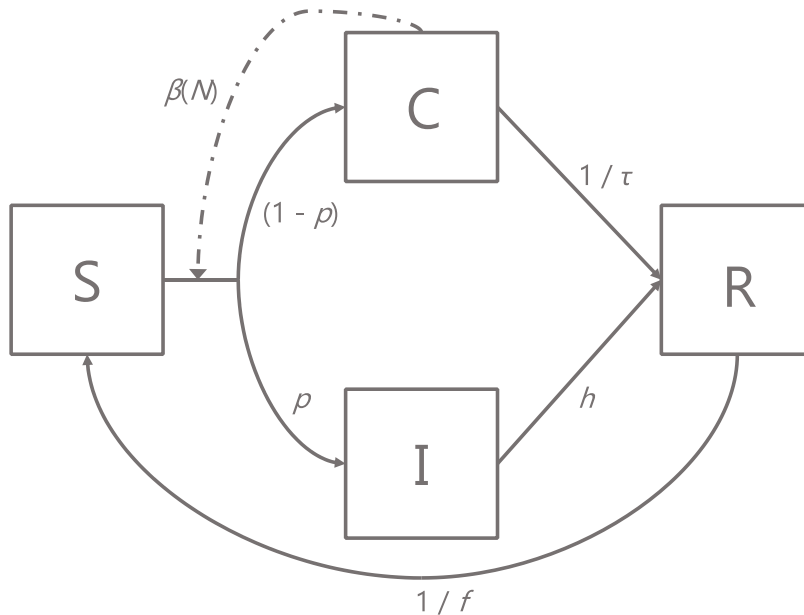


Figure 3.7: Schematic diagram of the epidemiological model. Boxes represent proportions of hosts in mutually exclusive states: susceptible (S), infected asymptomatic carriers (C), invasive (I) or recovered and immune (R). Solid arrows represent flows of individuals between states, and dashed arrows represent factors influencing those flows. Equations describing the system are presented in the main text (methods equations 3.1), along with parameter definitions (Table 3.2). See Chapter Two for initial model development.

3.5.2.1 Power Contact Rate Model

To analyze how different invasive disease progression processes can affect measures of invasive disease risk, we construct a compartmental epidemiological model of pneumococcal transmission based on previous work developed in Chapter Two (Metcalf et al., 2023). 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 (Figure 3.7), and their dynamics are given by the following system of four ordinary differential equations where the force of infection $\lambda = bN^rC$

$$\frac{dC}{dt} = (1 - p)bN^rCS - \frac{C}{\tau}$$

$$\frac{dI}{dt} = pbN^rCS - hI$$

$$\frac{dR}{dt} = \frac{C}{\tau} - \frac{R}{f} + hI$$

$$S = 1 - C - I - R. \quad [3.1]$$

Recall that the force of infection r parameter specifies how population density impacts transmission (Fig 3.1). Depending on the value of r , the power contact density function allows for a purely density-dependent ($r=1$) or frequency-dependent ($r=0$) host contact structure (Borremans et al., 2017). An r value between 0 and 1 generates a hybrid contact density function that has an intermediate saturating shape which is more strongly density-dependent at lower population densities but plateaus at higher densities

(Borremans et al., 2017). The model assumes transmission is from carriers only, there are no multiple infections, and that intrinsic virulence, the core set of traits that concern adhesion, invasion, and proliferation in disease sites, are equivalent for all serotypes. In other words, the invasive disease parameter p is 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 3.2.

Table 3.2: Definitions for the variables and parameters used in the compartmental epidemiological model along with values and reference sources used in the model fitting.

Variables		
	Definition (Units)	
$S(t)$	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	
Fixed Parameters From Prior Literature		
	Definition (Units)	Value (Reference)
p	Probability of progressing to the invasive state at the time of carriage acquisition	0.00029 (Metcalf et al., 2023)
τ	Average duration of carriage for a given Spn strain (weeks)	Serotype dependent (Sleeman et al., 2006)

Table 3.2 continued

h	Rate of transition from the invasive to recovered class, due to pathogen clearance (per week)	0.5 (Baldo et al., 2015)
N	Population density (persons per square mile)	Serotype dependent (Langley et al., 2015)
Estimated Parameters		
Definition (Units)		
f	Power contact rate model duration of immune protection (weeks)	
r	Power contact rate function exponent	
b	Power contact rate model transmission rate constant	
κ	Carriage duration threshold constant for the piecewise contact rate model (weeks)	
b_D	Piecewise contact rate model density-dependent transmission rate constant	
f_P	Piecewise contact rate model duration of immune protection (weeks)	

Solving the system of differential equations representing the SCIR model at a stable endemic equilibrium $[S^*, C^*, I^*, R^*]$ yields the following,

$$S^* = \frac{N^{-r}}{b\tau - bp\tau}$$

$$C^* = -\frac{N^{-r}(bh(p-1)\tau N^r + h)}{b(fh - hp\tau + h\tau + p)}$$

$$I^* = \frac{pN^{-r}(b(p-1)\tau N^r + 1)}{b(p-1)\tau(fh - hp\tau + h\tau + p)}$$

$$R^* = \frac{fhN^{-r}(b(p-1)\tau N^r + 1)}{b(p-1)\tau(fh - hp\tau + h\tau + p)}. \quad [3.2]$$

A key aim for this analysis is to estimate the level of density-dependence within the host contact structure as represented by the power contact rate function exponent r , using serotype-specific epidemiological data on carriage duration τ , population density N , and invasive incidence D . To do this we first derive the steady state invasive incidence D^* from the equilibrium solutions to this system of differential equations by defining a function based on new invasive cases ($D^* = pbN^r C^* S^*$) yielding,

$$D^*(\tau, N) = \frac{pN^{-r}(bh(p-1)\tau N^r + h)}{b(p-1)\tau(fh - hp\tau + h\tau + p)}. \quad [3.3]$$

We then derive an expression for population density as a function of carriage duration and invasive disease abundance by rewriting the equation for invasive incidence at equilibrium (D^* equation 3.3 above) in terms of population density N ,

$$N(\tau, D^*) = \left(\frac{b(1-p)\tau(fh - hp\tau + h\tau + p) \left(-\frac{hp^2}{(1-p)(fh - hp\tau + h\tau + p)} + \frac{hp}{(1-p)(fh - hp\tau + h\tau + p)} - D^* \right)}{hp} \right)^{-1/r}. \quad [3.4]$$

3.5.2.2 Piecewise Contact Rate Model

We also present an alternative epidemiological compartmental model that employs a piecewise transmission term function with a defined density-dependent and frequency-dependent component. Besides the transmission term, the overall structure of the model is unchanged from Figure 3.7. The model dynamics are given by a system of four ordinary differential equations,

$$\frac{dC}{dt} = (1 - p)\beta[b_D]CS - \frac{C}{\tau}$$

$$\frac{dI}{dt} = p\beta[b_D]CS - hI$$

$$\frac{dR}{dt} = \frac{C}{\tau} - \frac{R}{f} + hI$$

$$S = 1 - C - I - R. \quad [3.5]$$

Where the transmission term $\beta[b_D]$ is conditional on the threshold carriage duration parameter κ ,

$$\beta(b_D) = \begin{cases} b_D N, & \tau < \kappa \\ b_D N_\kappa, & \tau \geq \kappa \end{cases}$$

The threshold carriage duration parameter κ marks the transition point from density- to frequency-dependent transmission. This piecewise threshold cutoff would naturally be defined in terms of population density instead of carriage duration. It is required to be written in terms of carriage duration here as a constraint of fitting two functions, population density and invasive incidence, simultaneously. At least one parameter, in this case carriage duration, needs to link the two functions together to use

both equations to fit shared parameters. However, once we estimate κ we can plug that value into our function for population density N (equation 3.8, derived below) to represent the cutoff threshold in terms of N (parameter N_κ). At population density N_κ , the transmission coefficient β transitions to frequency-dependence where β is fixed at the threshold value ($b_D N_\kappa$) and invariant across higher population densities (Figure 3.4). The endemic equilibrium $[S^*, C^*, I^*, R^*]$ solutions for the piecewise contact rate model are provided below,

$$\begin{aligned}
S^* &= -\frac{1}{(p-1)\tau\beta[b_D]} \\
C^* &= \frac{(p-1)\tau\beta[b_D] + 1}{(-f + p(\tau-2) - 1\tau)\beta[b_D]} \\
I^* &= \frac{p(\tau(2-2p)\beta[b_D] - 2)}{(p-1)\tau(-f + p(\tau-2) - \tau)\beta[b_D]} \\
R^* &= \frac{f((p-1)\tau\beta[b_D] + 1)}{(p-1)\tau(f - p\tau + 2p + \tau)\beta[b_D]}. \quad [3.6]
\end{aligned}$$

Similar to the power contact rate model, the piecewise model steady state invasive incidence D_p^* is based on new invasive cases from the density-dependent ($D_D^* = pb_D NC^* S^*$) and frequency-dependent ($D_F^* = pb_D N_\kappa C^* S^*$) components yielding,

$$D_p^*(\tau, N) = \begin{cases} -\frac{p(b_D h N (p-1)\tau + h)}{(b_D N \tau - b_D N p \tau)(f h - h p \tau + h \tau + p)}, & \tau < \kappa \\ \frac{h p (b_D N_\kappa (p-1)\tau + 1)}{b_D N_\kappa (p-1)\tau (f h - h p \tau + h \tau + p)}, & \tau \geq \kappa \end{cases} \quad [3.7]$$

We next derive a piecewise expression of population density as a function of carriage duration τ and invasive incidence D . For the density-dependent component, we rewrite the equation for invasive incidence at equilibrium (D_D^*) in terms of population density N ,

$$N_D P(\tau, D^*) = -\frac{hp}{b_D(p-1)\tau(-fhD+h(Dp\tau-D\tau+p)-Dp)} \quad [3.8]$$

By definition, we cannot rewrite the frequency-dependent invasive incidence equation in terms of N since it is not a function of N . Instead, we specify that past the piecewise threshold cutoff carriage duration κ , serotypes will transition from density- to frequency-dependent transmission, and both the contact and transmission rates will hold constant at that level across all larger population densities. Thus, the piecewise function for population density will be the following,

$$N_P(\tau, D^*) = \begin{cases} N_D P(\tau, D^*), & \tau < \kappa \\ N_D P(\kappa, D^*), & \tau \geq \kappa \end{cases} \quad [3.9]$$

3.5.3 Statistical Analysis

We use the ‘NonlinearModelFit’ function from Mathematica 11.1.1.0 to perform model fitting for both the power and piecewise contact rate models. We simultaneously fit the population density and invasive incidence data captured by ABCs by defining equations for population density and invasive incidence as components of a piecewise function. Parameters are then estimated for both the power contact rate model (equations 3.3 and 3.4) and the piecewise contact rate model (equations 3.7 and 3.9). The data used for model fitting is provided in Table B.1. For the power contact rate model, we use the Automatic method which enables Mathematica to choose a model fitting algorithm based on the data

and model. For the piecewise contact rate model, we use the NMinimize global optimization algorithm employing the Differential Evolution function minimizer with the parameters CrossProbability = 0.1 and ScalingFactor = 0.9. We also specify that local search be performed by the QuasiNewton method using the PostProcess argument. A sensitivity analysis of the model fit employing equations 3.4 and 3.9 is performed using the ‘multisensi’ R package v2.1-1. For the power contact rate model, parameter ranges are calculated using the 95th percentiles of either the Normal distribution (parameter f) or the truncated Normal distribution (parameters r and b). The truncated Normal distribution is required to prevent negative b values and values of r that exceeded the limit case ($r = 1$) (Table 3.3). Ranges for the f , b , and κ parameters of the piecewise contact rate model are all derived from the Normal distribution (Table 3.3). The Normal and truncated Normal mean and standard deviation values come from the best fit model estimates. The truncated normal distributions are generated using the ‘truncnorm’ R package v1.0-9. All additional statistical analyses are carried out using the base package of R v4.3.3 and all plots are created using either the base package or ‘ggplot2’ v3.5.0. Only serotypes with at least 5 invasive cases were included in the analysis.

Table 3.3: The distribution function and sampling boundaries used in the sensitivity analyses testing the inverse relationship between population density and carriage duration. Parameter ranges are modified from Table 3.1 to use truncated Normal distributions where required to avoid parameter selections outside of biological limits (for parameters r and b). The parameter combinations plots sampled from the normal and truncated normal distributions to generate 500 sets of parameter estimates for both the power and piecewise contact rate models. The GSI sensitivity analysis used the upper and lower boundary constraints representing the 95th percentiles of the normal and truncated normal distributions to implement the GSI factorial design method.

Power Law Contact Rate Model				
Parameter	Distribution	Arguments	Lower Bound	Upper Bound
b	Trunc Normal	Mean = 7.69e-5 SD = 8.32e-5	6.35e-6	2.47e-4
f	Normal	Mean = 237 SD = 23.4	191	282
r	Trunc Normal	Mean = 0.988 SD = 0.139	0.680	0.995
Piecewise Contact Rate Model				
Parameter	Distribution	Arguments	Lower Bound	Upper Bound
f_P	Normal	Mean = 229 SD = 19.8	190	268
b_D	Normal	Mean = 6.67e-5 SD = 1.28e-05	4.16e-05	9.18e-05
κ	Normal	Mean = 11.07 SD = 1.27	8.58	13.56

CHAPTER 4. HUMAN SOCIETAL DETERMINANTS AND BACTERIAL LIFE HISTORY TRAITS SHAPE PNEUMOCOCCAL INVASIVE DISEASE RISK

Adapted from METCALF, B. J., WALDETOFT, K. W., BEALL, B. W. & BROWN, S. P. 2023. Human Societal Determinants and Bacterial Life History Traits Shape Pneumococcal Invasive Disease Risk (in prep).

4.1 Summary

Invasive pneumococcal disease (IPD) is a rare outcome of pneumococcal acquisition (on the order of 10 cases of IPD per 100,000 carriage acquisitions), yet for reasons poorly understood, cases of IPD are increasingly identified as forming clusters of localized outbreaks. While prior evidence suggests membership within disadvantaged communities (e.g. people experiencing homelessness) are risk factors for IPD and IPD outbreaks, it is unclear if individual or societal factors are driving these associations. It is possible individual risk factors (e.g. immune deficits) are mediating the correlations between some at-risk subpopulations and IPD. Here, we consider whether there are broader social determinants that are also driving the relationships between IPD clusters and disadvantaged populations. In light of the Chapter Three finding of density-dependent transmission favoring short carriage duration serotypes in crowded environments, we hypothesize that disadvantaged populations are at increased risk of IPD and IPD clusters via living and working in crowded and inadequately ventilated social surroundings, that enrich the most dangerous short carriage duration serotypes. Using a cluster detection tool to scan genomes

of 13,798 *Streptococcus pneumoniae* invasive isolates acquired from a surveillance population of approximately 34.6 million people (the CDC Active Bacterial Core surveillance program), we find that IPD clusters are substantially more common than expected by chance and are most common in short carriage duration and high attack rate serotypes. Using causal analysis tools, we find evidence for social factors associated with crowding to be causal drivers of both IPD and invasive clustering. These interrelationships between social factors and Spn disease suggest that spending extended amounts of time in crowded and inadequately ventilated environments may drive IPD and invasive clustering by acting as a nexus point where more dangerous Spn strains can spread widely and quickly, especially within disadvantaged communities.

4.2 Introduction

Streptococcus pneumoniae (Spn) is a gram positive opportunistic pathogen and a leading cause of bacterial pneumonia, meningitis, and sepsis worldwide (O'Brien et al., 2009, Drijkoningen and Rohde, 2014). Pneumococcus is typically categorized by serotype, with each serotype distinguished by an immunologically distinct polysaccharide (PS) capsule. Although it is considered a major human pathogen, in most cases it colonizes the nasopharynx asymptotically (defined as carriage), and is cleared by the immune system without causing harm (Henriques-Normark and Tuomanen, 2013). However, by employing an array of virulence factors, Spn colonization can sometimes lead to illness (Henriques-Normark and Tuomanen, 2013). Usually, Spn will cause non-invasive diseases like otitis media and non-bacteremic pneumonia, but, on rare occasions, can also spread into normally sterile sites and cause severe infections known as invasive pneumococcal disease (IPD) (Henriques-Normark and Tuomanen, 2013). The ability of Spn to cause these life-

threatening illnesses (*i.e.*, its invasiveness) varies substantially across serotypes (Sleeman et al., 2006, Brueggemann et al., 2004, Metcalf et al., 2023).

Relative to carriage, invasive disease is uncommon with rates ranging from about 1 to 75 cases per 100,000 carriage acquisitions depending on the serotype (Sleeman et al., 2006, Brueggemann et al., 2003). While IPD is a normally rare outcome of Spn, due to circumstances not well understood, it can occasionally also spread as localized outbreaks. Sometimes these outbreaks will occur within disadvantaged communities like people experiencing homelessness (PEH) and people who inject drugs (PWID) and may be linked to populations living in crowded environments that experience intense physical or psychosocial stress (Tomczyk et al., 2016, Romney et al., 2008, Gladstone et al., 2022, Dawood et al., 2011, Ly et al., 2021).

Spn invasive outbreaks have historically been difficult to study because of their rarity (Jefferies et al., 2007). Invasive outbreak identification presents a difficult challenge for traditional epidemiological tools because it requires capturing epidemiologically linked rare events that are temporally and geographically related. An alternative approach is to identify invasive clusters via genomic methods, by identifying sets of IPD cases that have close genomic identity, regardless of any conventional information on epidemiological links. Over the past few years, the CDC streptococcus lab (StrepLab) has been given a unique opportunity to study pneumococcal invasive clusters via genomic means that, until recently, have gone undetected. The StrepLab is a member of the Active Bacterial Core surveillance (ABCs) system which provides active laboratory- and population-based surveillance for key invasive bacterial pathogens. By genomically clustering related strains using a conservative distance threshold of 10 single nucleotide polymorphisms (SNPs)

across the genome, we found that Spn invasive clusters have been significantly underreported (Beall et al., 2022a, Metcalf et al., 2021). Additionally, we have found that Spn serotypes vary in their propensity to form invasive clusters and this variation is correlated with the proportion of cases that come from disadvantaged populations (PEH or PWID) (Beall et al., 2022a, Metcalf et al., 2021). The fact that invasive cluster risk is serotype dependent is important since other life history traits of Spn infection are tied to serotype, including measures of virulence such as attack rate (Chapter Two, (Sleeman et al., 2006)).

While prior evidence suggests membership within disadvantaged communities are risk factors for IPD, IPD outbreaks and IPD genomic clusters, it is unclear if individual or societal factors are driving these associations. It is possible individual risk factors (e.g. immune deficits) are mediating the correlations between some at-risk subpopulations and IPD. Dock workers, for example, may be at higher risk for IPD because they have greater exposure to inorganic dust and metal fumes (Cassir et al., 2021).

Here, we consider whether there are broader social determinants that are driving the relationships between IPD clusters and disadvantaged populations. We propose that disadvantaged populations are at increased risk of IPD and IPD clusters via living and working in crowded and inadequately ventilated social surroundings. In previous work (Chapter Three) we have shown that crowded high population density locations enable the spread of shorter duration serotypes, that tend to be more virulent (higher attack rate). We hypothesize this process will induce a negative correlation between serotype carriage duration and clustering propensity. Using a cluster detection tool to scan genomes of 13,798 *Streptococcus pneumoniae* invasive isolates acquired through ABC surveillance,

we find that IPD clusters are substantially more common than expected by chance, clusters are more common in short carriage duration serotypes, and we find evidence for social factors associated with crowding to be causal drivers of both IPD and invasive clustering.

4.3 Results

4.3.1 *Expected Number of Clustered Isolates Assuming an Independent Outbreak Risk*

While the per-person risk of progressing into invasive disease given pneumococcal acquisition is low ($p = 0.0003$, Chapter Two), given the size of our sampling population (~34M), we would expect some clustering to occur by chance. To estimate how much IPD clustering we should expect to see in our data, assuming independent risk of IPD, we use our previously described math model (Chapter Two, methods, (Metcalf et al., 2023)). Given the assumption that a cluster of size k is the product of k independent invasion events (*i.e.*, p^k), we find that the proportion of cluster-associated cases among all IPD cases is $p/(1-p)$. Using our previous estimate of $p = 0.0003$, this implies that ~ 0.03% of cases ($\sim p$) will be a part of clusters under our null model of independent IPD events (see methods for derivation of $p/(1-p)$ result).

4.3.2 *Proportion of cluster-associated cases is substantially greater than our null expectation*

Using an average invasive incidence of 9.4 cases per 100,000 persons and an average surveillance population of 33,508,337 persons tracked over 5 years, our math analysis above predicts we will find ~ 4.7 cases (0.03%) in clusters over a 5-year period by

chance. To assess this prediction, we used genomic relatedness (10 SNP per genome cluster cutoff, see methods) to define and identify clusters among 13,798 IPD isolates acquired through ABCs during 2015-2019 where a whole genome sequence was available. Using this approach, we found 3,020 cases in clusters, representing a substantial 22% of all IPD cases, far exceeding our 0.03% prediction of 4.7 cases.

The 3,020 clustered cases were allocated to 1,095 IPD clusters across 31 serotypes. Genomic clusters had an average size of 2.76 isolates (ranging from 2 to 40 isolates), with 130 of these clusters (representing 15 serotypes) comprised of at least 4 isolates. The mean duration from first to most recent isolate in a cluster was 440 days (range, 0-1,776 days). The pairwise genomic distance for isolates in a cluster ranged from 0 to 10 SNPs with an average overall pairwise distance of 6.64 SNPs. Cluster-associated isolates were from patients whose average age was 56.5 (range, 0-93) years.

4.3.3 IPD clustering and serotype.

Given we observe a substantial excess of clustered cases compared to our null expectation, we now turn to ask what forces impact the extent of clustering. According to our hypothesis, we expect PEH and PWID to be disproportionately associated with IPD clustering and with shorter duration serotypes. We propose that shorter duration, more invasive serotypes will be more likely to cluster because, based on work presented in Chapter Three, they are more likely to spread and cause invasive disease in highly crowded environments. Fig. 4.1 shows that Spn serotypes do vary in their propensity to form invasive clusters, with serotype-specific proportions of clustered cases ranging from zero (serotype 1, 18C, 21, 23F, 37, 6A, 6B) to 68% (serotype 12F). Figure 4.1 further shows

that this variation is positively correlated with the proportion of PEH or PWID (Kendall tau = 0.48, p-value = 0.00016).

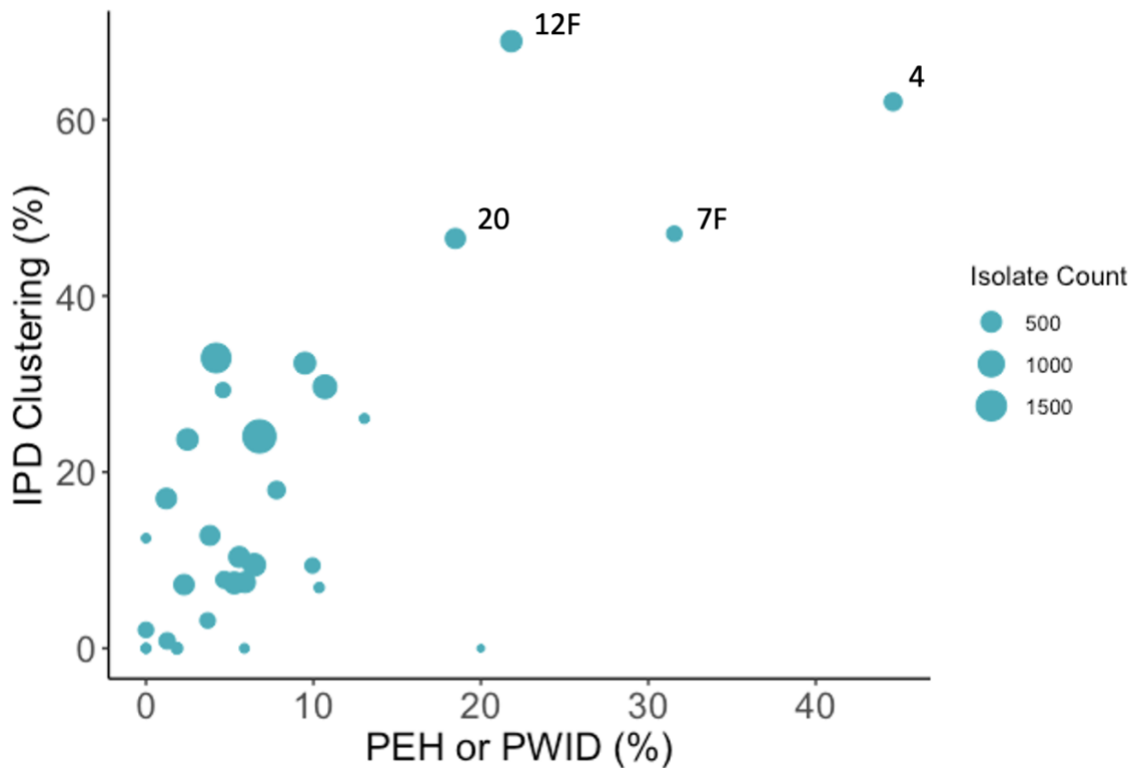


Figure 4.1: The amount of invasive clustering varies widely between Spn serotypes and is correlated with disadvantaged populations (PEH or PWID). Invasive clustering has a significant positive relationship with proportion of PEH or PWID (Kendall tau = 0.48, p-value = 0.00016). The average proportion of cluster-associated isolates weighted by invasive case count was 24%. The scatter plot reveals that serotypes vary widely in their propensity to cluster ranging from 0% cluster-associated isolates (serotypes 1, 18C, 21, 23F, 37, 6A, 6B) to 68% cluster-associated (serotype 12F). The top 4 most clustered serotypes (12F, 4, 7F, 20) are labeled in the figure.

4.3.4 IPD clustering is associated with serotype carriage duration and invasive attack rate

The disparity in IPD cluster proportion seen across serotypes is notable since several Spn life history traits are tied to serotype (Sleeman et al., 2006). Two of the most important with respect to transmission and invasive disease progression are Spn carriage duration and invasive attack rate. Using data obtained from Sleeman *et al.* (Sleeman et al., 2006), we show in Figure 4.2 that % IPD cases in clusters is negatively associated with serotype carriage duration (Kendall tau = -0.32, p-value = 0.018) and positively associated with serotype attack rate (Kendall tau = 0.29, p-value = 0.028), in agreement with our hypothesis.

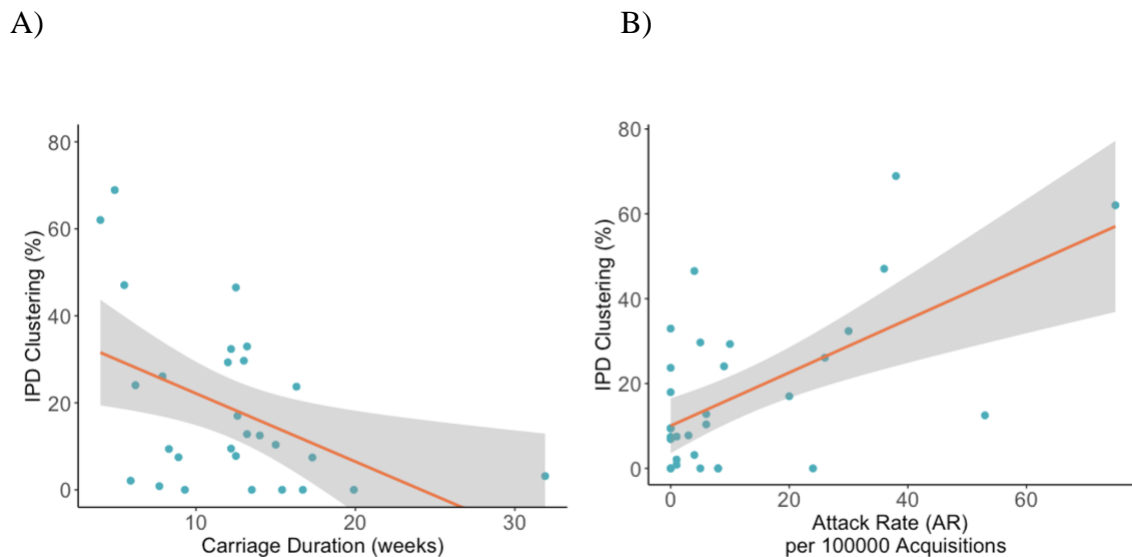


Figure 4.2: IPD clustering proportion is associated with bacterial life-history traits (positively correlated with high invasive attack rates and short carriage durations). (A) Serotype IPD clustering proportion (blue dots) against serotype carriage duration yields a significant inverse association (Kendall tau = -0.32, p-value = 0.018). (B) Serotype IPD clustering proportion (blue dots) against serotype invasive attack rates yields a significant positive correlation (Kendall tau = 0.29, p-value = 0.028).

4.3.5 Social features associated with health inequality are significant predictors of IPD and clustering

Given that certain communities like PEH and PWID are at increased risk for IPD and more likely to form invasive clusters (Ly et al., 2021, Mosites et al., 2019, Beall et al., 2022b), it is possible there are societal factors at play in pneumococcal invasion in addition to host and pathogen effects. The ABCs dataset offers a unique vantage point to study these social determinants since the patient case report form has census tract and zip codes that can be linked with demographic census data. To capture social factors that potentially play a causal role in determining IPD cases and clusters, we extracted a range of potentially relevant demographic variables from US census data (see methods for more details).

Our broad hypothesis is that high density environments will promote IPD clusters, as dense environments promote pneumococcal transmission and select for more invasive serotypes (Chapter Three). To assess the role of human population density, we focus on four demographic parameters connected to aspects of human density that we specifically hypothesize will promote IPD clusters: population density, household crowding, use of bus transit, and low income (Table 4.1). Using a univariate analytical approach, we find that IPD and invasive clustering are significantly positively associated with population density, bussing mass transit, and household crowding and negatively associated with household income (Table C.1). These univariate results support our hypothesis, but what of potential

confounding effects due to other indirect causal connections with a range of potentially confounding demographic and bacterial factors (Table 4.1)?

Table 4.1: Definitions of the demographic attributes included as possible IPD case and clustering determinants, extracted from US Census data (both the American Community Survey (ACS) and the Planning Database (PDB)). Units are provided as parentheticals in the Definition column.

Demographic Features of Interest

<i>Variable</i>	DAG Model	Definition	Data Source
<i>Household Income</i>	Fig. 4.3 Fig. C.1	Average household income of a census tract (United States dollar)	U.S. Census PDB
<i>Population Density</i>	Fig. 4.3 Fig. C.1	Number of people in a square mile of land area within a census tract (persons per square mile)	U.S. Census PDB
<i>% Bussing Transit</i>	Fig. 4.3 Fig. C.1	Households commuting by bus within census tract (percentage)	U.S. Census ACS B08301
<i>% Household Crowding</i>	Fig. 4.3 Fig. C.1	Households living in crowded housing units within census tract (percentage)	U.S. Census PDB

Potential Confounders

<i>Variable</i>	DAG Model	Definition	Data Source
<i>Housing Per Income</i>	Fig. 4.3 Fig. C.1	Proportion of household income devoted to housing costs within census tract	U.S. Census ACS B25071, B25092, B25008
<i>Fed Pov Lev</i>	Fig. 4.3 Fig. C.1	Persons living below the federal poverty level (percentage)	U.S. Census PDB
<i>Child<5</i>	Fig. C.1	Proportion of children under 5 within a census tract	U.S. Census ACS B01001

Table 4.1 continued

<i>Sero Car Dur</i>	Fig. 4.3 Fig. C.1	Average duration of carriage per Spn serotype (weeks)	Sleeman <i>et al.</i> (Sleeman <i>et al.</i> , 2006)
<i>Sero Inv AR</i>	Fig. 4.3 Fig. C.1	The ratio of IPD incidence to carriage acquisition rate per Spn serotype (IPD cases per 100,000 carriage acquisitions)	Sleeman <i>et al.</i> (Sleeman <i>et al.</i> , 2006)

To address potential confounding effects, we formalize causal connections among our variables as a directed acyclic graph (DAG) (Pearl and Mackenzie, 2018). Capturing causal relationships as a DAG allows use of established causal inference logic to guide how to conduct appropriate statistical controls (Pearl and Mackenzie, 2018). Causal inference allows us to avoid logical pitfalls associated with both univariate analyses (e.g. overlooking confounding variables) and all-variable multivariable analyses (e.g. opening collider biases, collinearity (Gregorich et al., 2021, Schisterman et al., 2017, Tönnies et al., 2022)), but we recognize that causal inference tools are still limited by assumptions embedded in the specific defined DAG. In the current analysis, our favored DAG is illustrated in Figure 4.3, with an alternate DAG explored in the supplementary materials (Figure C.1). Using these DAG models, we obtain unbiased effect estimates by generating the minimal sufficient adjustment set for each feature of interest which are the minimal set of covariates that require adjustment (Figure 4.4C).

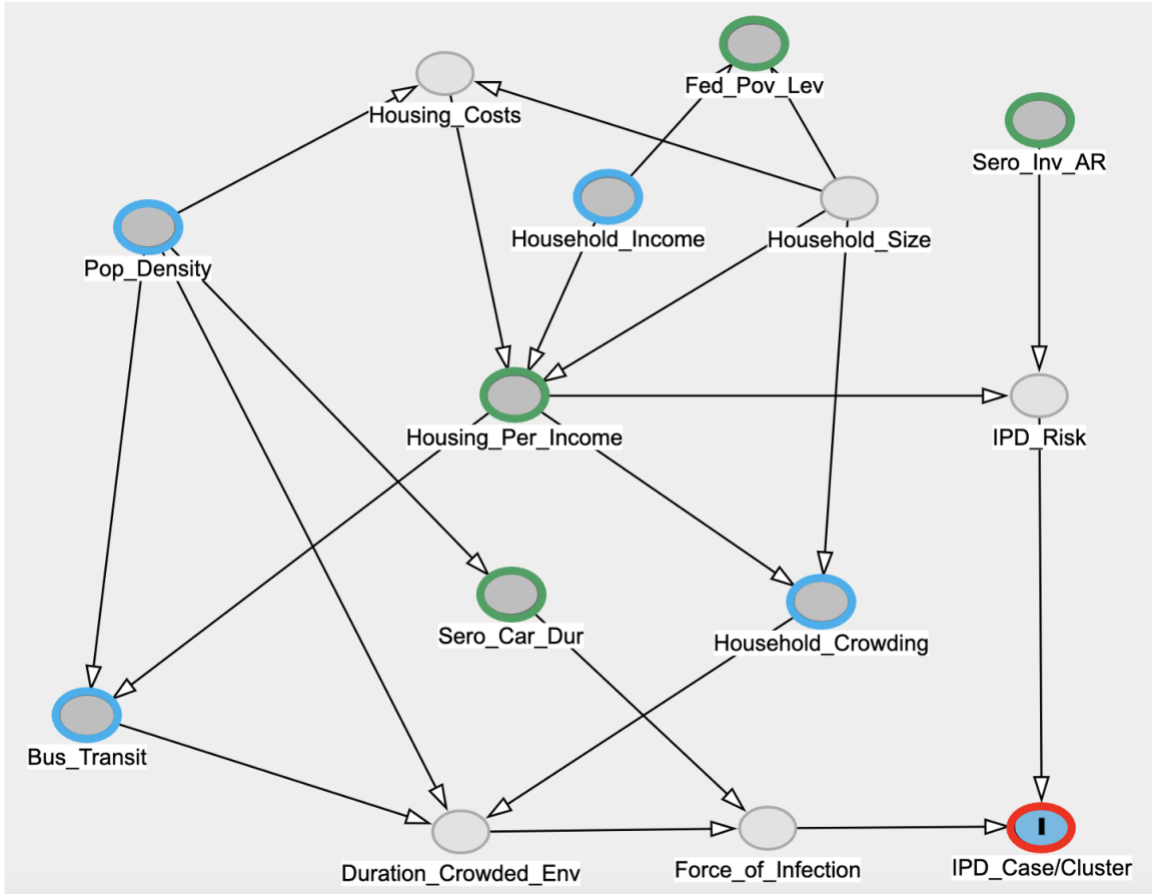
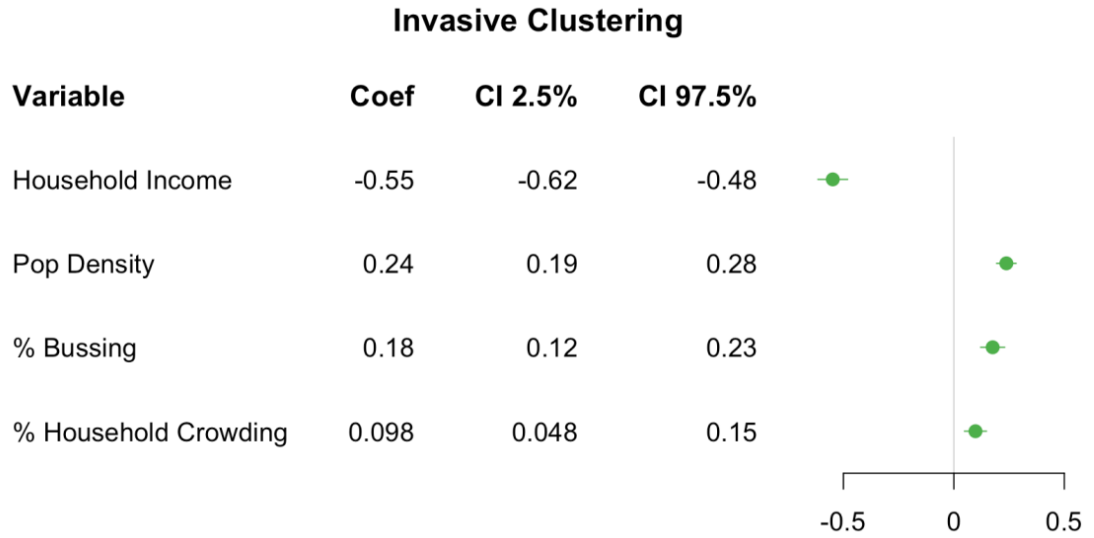


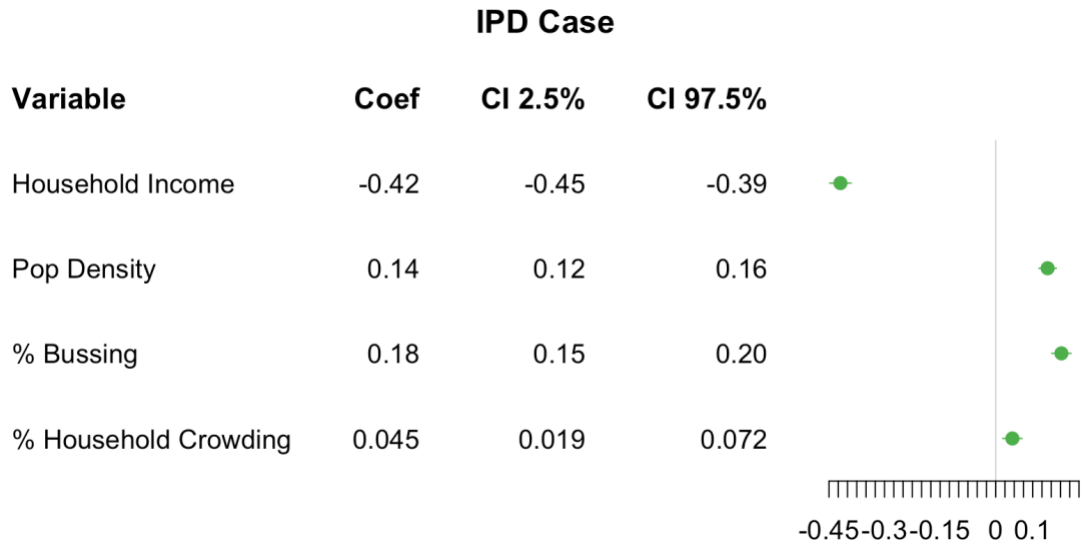
Figure 4.3: The relationship between both household crowding and bus transit with IPD clustering is confounded by financial instability and population density. The directed acyclic graph (DAG) represents the causal interrelationships between variables (nodes) influencing the development of IPD and IPD clusters. Dark grey nodes are observed variables and light grey node are unobserved (latent) variables. Demographic features of interest (Table 4.1) are highlighted in blue and potential confounders (Table 4.1) are highlighted in green. The IPD cluster outcome variable is highlighted in red.

Using the causal DAG in Figure 4.3 to identify non-biased statistical adjustments, we identify several attributes related to financial instability and crowding that are associated with invasive pneumococcal cases and clusters (Figure 4.4).

A)



B)



C)

DAG Model 1 (Figure 4.3)	
Variable	Minimal Sufficient Adjustment Set
Population Density	No adjustment needed
Household Income	No adjustment needed
Mass Transit – Bussing	Housing_Per_Income, Pop_Density
Household Crowding	Housing_Per_Income, Pop_Density
DAG Model 2 (Figure C.1)	
Variable	Minimal Sufficient Adjustment Set
Population Density	No adjustment needed
Household Income	Pop_Density
Mass Transit - Bussing	Housing_Per_Income, Pop_Density
Household Crowding	Child<5, Housing_Per_Income, Pop_Density

Figure 4.4: Several demographic features associated with health inequality and crowding are significant predictors of IPD disease and clustering. (A) and (B) The results show that census tracts with lower average household incomes, higher population densities, and a higher use of public bussing are positively associated with both IPD total cases (panel A) and IPD clustering cases (panel B). The standardized coefficient confidence intervals for each variable are derived from a negative binomial multiple regression model that is adjusted for each variable of interest to have a non-biased effect on IPD outcome based on the causal DAG in Figure 4.3. (C) The minimal sufficient adjustment set of each variable of interest as specified by DAG Model 1 and DAG Model 2.

In Figure 4.4 we find that percent of households commuting by bus within a census tract, census tract population density, census tract average household income, and households living in crowded housing units within census tract are all significantly predictive of IPD and Spn invasive clustering. Given that population density and average household income are on different scales and several features are in different units, we use standardized coefficients to compare the magnitude of effect between the explanatory variables. We see that for both IPD and cluster-associated cases, average household income has the largest magnitude followed by population density and percent of households commuting by bus which had comparable effects. Percentage of overcrowded households within a census tract was predictive for IPD and cluster-associated cases, though its overall impact was the weakest among the four variables studied.

4.4 Discussion

While Spn continues to be the leading cause of community-acquired bacterial pneumonia, its impact on invasive disease has been curtailed by the introduction of the pneumococcal conjugate vaccine (Pilishvili et al., 2010, Peyrani et al., 2019). Adult invasive disease, however, still remains a challenge and especially within at-risk communities like PEH and PWID where IPD outbreaks have been known to occur (Ly et al., 2021, Vanderkooi et al., 2011, Kellner et al., 2021). The goal of this analysis was to leverage ABC surveillance data to understand the forces driving the persistence of IPD and invasive clustering within the US population. We show that Spn invasive clusters are significantly more likely to be comprised of individuals from at-risk populations like PEH and PWID (Figure 4.1) and that the serotypes generating these clusters are more likely to have shorter carriage durations and higher attack rates (Figure 4.2). Our analysis also

supports a causal association of several social determinants of health linked with crowding and income to both IPD and invasive clustering. Using standardized coefficients, we find that household income had the largest effect size, followed by population density and bus transit with comparable effects, and finally household crowding with a marginal impact on IPD and clustering (Figure 4.4).

The lineages with the highest proportions of clustered isolates (serotypes 7F, 12F, 4, and 20) also had the highest proportions of PEH or PWID and were distinctly separated from other serotypes in Figure 4.1. This agrees with previous work associating these serotypes with invasive disease outbreaks and genomic clusters (Beall et al., 2022a, Zulz et al., 2013, Robinson et al., 1999, Gladstone et al., 2022). Our finding that serotypes with shorter carriage durations and higher levels of invasiveness have an increased propensity to cluster is consistent with our hypothesis that these serotypes are more likely to cause IPD clusters because they rely on high population density environments to spread to new hosts (see also Chapter Three). We expect this would lead to an association between invasive clusters and disadvantaged communities (i.e. PEH and PWID) as shown in our investigation because these communities tend to live in crowded and inadequately ventilated conditions that promote disease transmission (Moffa et al., 2019, Badiaga et al., 2008). This also agrees with previous work showing that PEH have higher rates of Spn carriage are more likely to develop invasive disease compared to the general population (Lemay et al., 2019, Mosites et al., 2019, Ly et al., 2021). Broadening out to the population level, we would expect census tracts with lower socioeconomic levels and higher crowding indicators to have a higher propensity to cause IPD and invasive clusters. In fact, this is what we observe in our analysis (Figure 4.4).

This study has several limitations. The census tract information in our ABC surveillance data was missing for 1,506 of 13,798 case-patients (11% of total) and, as might be expected, was disproportionately higher in PEH (609 of 781 cases, 78%) compared to non-PEH (897 of 13,017 cases, 6.9%). Because of a lack of stable housing, there may also be measurement error in the location metadata that was available for people experiencing homelessness. Thus, we are likely underrepresenting the impact of host demographic factors on IPD within PEH communities. In addition, the Spn carriage duration and invasive attack rate data used in this analysis are the most comprehensive available but were acquired more than 15 years ago as part of a series of longitudinal pediatric studies in the United Kingdom (Sleeman et al., 2006). Over this time, several formulations of the pneumococcal conjugate vaccine (PCV) have been released that have altered the composition of circulating serotypes (Devine et al., 2017). Also, these pediatric surveillance studies may not capture unique features of adult invasive disease epidemiology (Alanee et al., 2007).

It's important to underline that the causal DAG presented in Figure 4.3 to guide covariate adjustments represents our assumptions on causal interactions among variables, and clearly other assumptions (in particular the incorporation of additional causal factors) can be made. The results we gain from it should therefore not be considered immutable. Whether we control for biases and, if so, what methods we use, can meaningfully change the predicted effect our selected attributes have on IPD and invasive clustering. We illustrate this by comparing attribute coefficient values from our baseline DAG Model 1 (Figure 4.4) against three alternate models: univariate, multivariate, and an alternate DAG Model 2 (Figure C.1). In the alternate DAG Model 2, we include the proportion of children

under 5 in a census tract as an additional node in the graph which incorporates the effect children have on Spn transmission (Althouse et al., 2017). The coefficients and significance values for each variable of interest across the four models are provided in Table C.2. We see the univariate method generally agrees with our baseline DAG Model 1 and the alternate DAG Model 2, but the multivariate approach yields some substantial differences. These discrepant results may be due to how multivariable regression can distort the effects of independent variables when not guided by a causal framework. Multiple regression may include highly correlated or collinear features that can increase the standard errors and widen confidence intervals (Schisterman et al., 2017, Gregorich et al., 2021). Additionally, if the set of independent variables includes a collider, the regression analysis may induce artificial, and sometimes paradoxical, associations that have no causal basis (Tönnies et al., 2022).

The interrelationships between community factors and Spn disease described in our analysis suggests that spending extended amounts of time in crowded and inadequately ventilated environments may drive IPD and invasive clustering by acting as a nexus point where Spn strains can spread more widely and quickly, especially within disadvantaged communities. These results have important implications for future research and potentially for public health management. We underline at this point that our conclusions are not to guide against public transportation which provides substantial benefits, particularly to disadvantaged populations. Any restrictions to public transit could compound chronic stress effects on immune health and increase IPD potential in at-risk communities (i.e. the causal arrow from 'Housing_Per_Income' to 'IPD Risk' in Figure 4.3). In Chapter Five we address how future research can build on the findings of this and other chapters in this

thesis to further resolve the interacting impacts of bacterial, individual, and social determinants of invasive pneumococcal disease.

4.5 Methods

4.5.1 CDC ABCs surveillance data (genomic and case history)

The Active Bacterial Core surveillance program is an active laboratory- and population-based surveillance system that tracks several invasive bacterial pathogens including *Streptococcus pneumoniae*. The total ABCs surveillance population is approximately 34.6 million people, distributed across 10 US states (Figure C.2). The catchment area includes 3 full US states (Connecticut (CT), Minnesota (MN), and New Mexico (NM)) and selected counties within 7 other states (California (CA), Colorado (CO), Georgia (GA), Oregon (OR), New York (NY), Tennessee (TN), Maryland (MD)). An IPD case is defined as the isolation of *S. pneumoniae* from a normally sterile site (e.g., blood, cerebrospinal fluid, pleural fluid) and is accompanied by a detailed case report form that captures patient demographic and clinical information. We use the CDC's Active Bacterial Core surveillance to identify cases of IPD from January 1, 2015, through December 31, 2019. Laboratory audits are performed at least once per year to ensure all cases of disease under surveillance are reported. Surveillance staff at sites complete a standardized case report form that includes basic demographic characteristics and risk factors of infection on all cases, and coordinate collection and transfer of isolates to CDC's *Streptococcus* lab for characterization. We include all available invasive Spn isolates from cases identified in 2015-2019 where a whole genome sequence was available (13,798 of 15,705 total cases).

4.5.2 Census data and bacterial traits

To identify demographic features associated with IPD and clustering, we use data obtained from 2015-2019 5-Year estimates of the U.S. Census Bureau's American Community Survey (ACS) and 2021 Planning Database (PDB) with census tract as the geographic unit of measure (Table 4.1). For population density measurements, we use the U.S. Census Bureau's 2021 Planning Database (PDB) which provides both population count and land area data at the census tract level. Population density for each census tract is calculated by dividing the population size by its land area. A population density is associated with each IPD case by linking the census tract population density with the census tract recorded in the case report form. Commuting data is acquired from the B08301 data table titled 'Means of Transportation to Work' using 2015-2019 5-Year estimates of the ACS and the proportion of children under 5 in each census tract is obtained from the B01001 data table titled 'Sex By Age' also using 2015-2019 5-Year estimates of the ACS.

An important consideration for this analysis is how to quantify financial instability. Some measures like the percentage of households under federal poverty and average household income do not consider the variation in cost of living across our surveillance sites. To measure the overall financial health of a census tract, we constructed an attribute that measures the proportion of household income devoted to housing costs (Housing_Per_Income variable in Figure 4.3 and C.1) from 3 tables sourced from the 2015-2019 ACS 5-Year Estimates. Rental monthly housing costs as a proportion of income were obtained from table B25071, owner monthly housing costs as proportion of income were obtained from table B25092, and proportion of renters and owners in each census tract was obtained from table B25008. The total monthly housing costs as proportion of income for each census tract was derived from the weighted average of owner and renter housing costs.

We treat this metric as a proxy of non-discretionary income since housing costs are generally the largest non-discretionary expense (Kirkpatrick and Tarasuk, 2011, Stone, 2010). This measure has the advantage over other financial metrics in that it responds to changes in cost-of-living expenses. One issue with using the ‘Housing_Per_Income’ variable as a measure of the overall financial health is that the data is missing for 566 of the 7,608 census tracts used in our analysis. To assess the impact of this level of missingness, we set missing census tracts to both the minimal and maximal ‘Housing_Per_Income’ values from our dataset and found that either imputation did not materially change our results. For our analysis, we set all missing census tracts to the more conservative minimal value.

Both Carriage duration data and invasive attack rates were obtained from a study from Sleeman *et al.* (Sleeman *et al.*, 2006). Carriage duration for serotype 4 was too short to be reliably detected given the sampling times of the longitudinal study, so we imputed a conservative estimate of four weeks (the longest sampling interval used in Sleeman *et al.*) instead.

4.5.3 Cluster Identification

The cluster analysis workflow begins by completing an all-vs-all pairwise core-genome SNP distance matrix generated using Nucmer from the MUMmer package v3.9.4 (Kurtz *et al.*, 2004). To improve computational efficiency, this step of the program is run on a high-performance computing (HPC) system where the pairwise comparisons for each isolate are run as separate jobs. The distances are reformatted into an adjacency list and then fed into the hclust clustering tool in the R “stats” package v3.6.2 which defines clusters

using a SNP cutoff of 10 (R Core Team, 2020). Finally, each isolate belonging to a cluster of at least 2 cases is annotated with the following epidemiological metadata for each patient: patient age, ABC surveillance site, culture date, long term care facility (LTCF) residence status (yes/no), PEH status (yes/no), and past or current injection drug use status (yes/no). Due to the strong correlation between PEH and injection drug use the union of these two attributes (defined as either PEH or PWID) was used in this analysis and was referenced as PEH/PWID status (yes/no).

4.5.4 *Causal inference methods*

We formalize our causal analysis of host demographic features associated with IPD and clustering using a directed acyclic graph (DAG) methodology. Employing a causal DAG framework allows the use of established causal inference methods to properly guide covariate adjustment in our statistical models (Pearl and Mackenzie, 2018). In our focal DAG (Figure 4.3), we explicitly capture causal assumptions linking our variables of interest (*e.g.* a causal link from household income to living at or below the federal poverty level). To account for an increase in IPD potential in disadvantaged communities due to chronic psychosocial stress, we include an arrow from the ‘Housing_Per_Income’ to ‘IPD Risk’ nodes. For our analysis, we specify a direct causal path from host population density and bussing public transit to invasive disease and IPD clusters through duration in crowded environments (a latent variable). We also include a causal path from household crowding to IPD and IPD clustering through duration in crowded environments.

The graph reveals that housing as proportion of income (node ‘Housing_Per_Income’) and population density (node ‘Pop_Density’) act as confounders in the relationships between

bus transit use and household crowding with IPD and IPD cluster formation. However, when they are adjusted for, there is a direct causal path for both bus transit use and household crowding exposures.

In order to assess the robustness of the causal associations delineated in Figure 4.3, we also present an alternative causal DAG that includes a more comprehensive set of variable interactions and looks at mass transit more broadly (Figure C.1). In this formulation 1) Housing_Per_Income and Pop_Density act as confounders for mass transit use, 2) Housing_Per_Income, Pop_Density, and children under 5 confound household crowding, and 3) Pop_Density acts as a confounder for Household_Income.

4.5.5 Statistical modeling

The ABCs case report form provides the census tract for each invasive case. Using this geocoded identifier, we connect each patient case with an area-level measurement of key social determinants of health. These features can then be used as independent variables in a negative binomial regression random effects model to identify features associated with Spn invasive disease and clustering. We treat the U.S. State of each case location as a random effect, the health equity metrics as fixed effects, and either total invasive case or total cluster-associated case counts for each census tract as dependent variables. To account for census tracts varying in population size, we model the counts as proportions by using the offset option with census tract population size as the parameter. All random-effects modeling is performed using the R ‘lme4’ package v1.1.31 (R Core Team, 2020). For estimation of the causal effect between each health equity metric and IPD total and cluster-associated cases, we employ the causal DAGs shown in Figure 4.3 and C.1. We define

each health equity metric as an exposure and use the `ggdag` v0.2.10 and `dagitty` v0.3-1 package in R to identify the adjustments required to isolate a non-biased path between the variable and census tract case counts (either total IPD cases or cluster-associated cases). We then adjust for these biases by adding them as covariates in the negative binomial model described above. Finally, to compare the relative magnitude of the effects of different explanatory variables, we standardized the independent variables using the ‘scale’ function in R and created standardized coefficients (a mean of 0 and a standard deviation of 1).

4.5.6 *Mathematical modeling*

Though the probability of invasion per carriage event is low (previously estimated in Chapter Two to be 0.0003), the large number of Spn carriage acquisitions that occur in the US ensures that thousands of invasive cases will be reported within the CDC Active Bacterial Core surveillance (ABCs) system. If enough invasive cases occur, we expect some number of them to be caused by the same strain simply by chance. Also, we would expect serotypes with higher invasive attack rates (AR) to be more likely to produce outbreaks simply because they will cause a higher number of invasive cases relative to a less invasive serotype with a similar carriage duration. This will confound any relationship between attack rate, carriage duration and clustering propensity. We can quantify how much of the observed clustering signal is attributable to random chance by modifying the Spn compartmental model that we developed previously in Chapter Two (Metcalf et al., 2023). The proportions of individuals in a susceptible, carrier, invasive diseased and recovered class for a focal strain are denoted by the variables S , C , I and R respectively in

Figure C.3, and their dynamics are given by the following system of four ordinary differential equations,

$$\frac{dC}{dt} = (1 - p)\beta CS - \left(\frac{1}{\tau} + d\right)C$$

$$\frac{dI}{dt} = p\beta CS - hI + dC$$

$$\frac{dR}{dt} = \frac{1}{\tau}C + hI - fR$$

$$S = 1 - C - R - I \quad [4.1]$$

The model variables and parameter definitions are presented in Table C.2 and detailed in Metcalf *et al.* (Metcalf et al., 2023). We first derive a function representing the number of cluster-associated isolates by additionally assuming that a cluster of size k is the product of k independent invasion events (*i.e.*, p^k). To calculate the number of cases from clusters of size k , we multiply the cluster probability by the number of new infections (βCS) and sum across all possible clusters greater or equal to size 2 yielding the following infinite sum, $\sum_{k=2}^{\infty}(p^k)\beta CS$.

We then define a function L_P which represents cases from clusters of size k as a proportion of all new invasive cases by dividing the sum of all clusters greater or equal to size 2 by the invasive incidence ($p\beta CS$),

$$L_P = \frac{\sum_{k=2}^{\infty}(p^k)\beta CS}{p\beta CS} = \frac{p}{1-p} \quad [4.2].$$

The proportion of cluster-associated cases from new invasive cases will be approximately p if p is small. To calculate the expected number clusters present within our dataset, we use the 1-year average invasive incidence (9.4 cases per 100,000 persons) and average surveillance population (33,508,337 persons) obtained from the 2015 – 2019 ABC surveillance reports (<https://www.cdc.gov/abcs/reports-findings/surv-reports.html>). Multiplying 1-year incidence, population size, and 5 yields total estimated cases over a 5-year period and multiplying that value by p provides the expected total cluster count.

CHAPTER 5. GENERAL DISCUSSION

5.1 Summary of work

Given the dangerousness of invasive *Streptococcus pneumoniae* infections, understanding the factors and interactions that drive invasive disease and IPD clustering within disadvantaged groups is critical for guiding public health policy. In this dissertation, we sought to explore how pneumococcal life-history traits, host demographic factors, and the interrelationships between them can impact invasive disease at the population level. We began in Chapter Two with an analysis of how variation in carriage duration can confound serotype invasiveness measurements. The modeling from that investigation also supported previous work showing that when invasion occurs, it does so at or near the time of carriage acquisition. In Chapter Three we discussed how host population density interacts with serotype carriage duration and can modulate the distribution of serotypes circulating within communities of different population densities. The fourth chapter focused on how social determinants of health and variation of invasiveness among Spn serotypes can increase the likelihood of genomic IPD clusters occurring.

Each content chapter includes a self-contained discussion, so in this final discussion chapter I will provide a broader discussion of potential future directions of research, building on the work in this thesis. In this concluding chapter we motivate future work by presenting a hypothesis that people experiencing homelessness are at increased risk of IPD and IPD clusters via 1) epidemiological selection for high transmission, high virulence serotypes, and 2) delays in innate immune cell recruitment and diminished innate response. The proposed work presented here will assess how social determinants of health can not

only shape an individual's risk of developing invasive disease but also shape the virulence and transmissibility of the pathogen itself.

5.2 Societal and Within-Host Determinants of Invasive Pneumococcal Disease Outbreaks

Invasive pneumococcal disease (IPD) is a serious and normally rare outcome of *Streptococcus pneumoniae* (Spn) colonization that, under circumstances not understood, can sometimes also spread as localized outbreaks especially among disadvantaged populations. As discussed in Chapter Four, both the risk of IPD and involvement with IPD outbreaks is heightened in disadvantaged communities like people experiencing homelessness (PEH) and people who inject drugs (PWID). In Chapter Four we find evidence for social factors associated with crowding to be causal drivers of both IPD and invasive clustering, but we did not directly investigate the potential role of individual factors such as immune deficits in shaping individual disease risk.

While there is evidence that PEH suffer reduced innate immunity (discussed below), the effect of innate immune cell behavior on IPD risk is also not fully known. Changes in innate immune response in at-risk populations may be important given the evidence detailed in Chapter Two showing that when invasion does occur, it does so very quickly after Spn acquisition. This naturally leads us to ask what conditions following acquisition trigger IPD and what features of Spn strains, host demography and immunity make IPD outbreaks more likely. While this dissertation outlines bacterial and host demographic factors impacting IPD clustering, we do not know how these patterns are shaped by the interplay of homelessness, host immunity and serotype.

Building on the results presented in Chapters Two, Three and Four, we propose that IPD clusters emerge from the confluence of three major components: innate immune response, microbial life-history traits, and host population and housing density (Figure 5.1, green boxes). In this chapter we will discuss how these components influence IPD cluster formation at both the population and within-host levels and lay the groundwork for future work to test the hypotheses laid out in Figure 5.1.

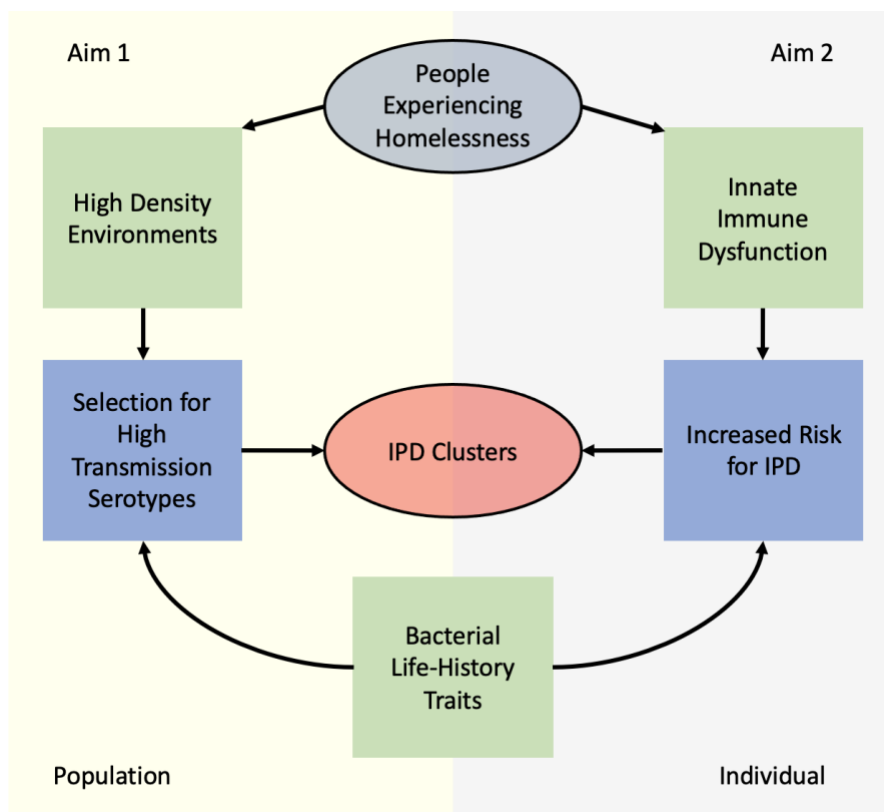


Figure 5.1: Schematic of IPD cluster hypothesis. We hypothesize that PEH (grey oval) suffer high rates of IPD clusters (red oval) via 3 interacting components (green boxes) mediated by two processes (blue boxes): human density, immune dysfunction, and bacterial life-history traits. These components influence IPD cluster formation at

both the population (yellow, left half of figure) and within-host levels (grey, right half of figure).

5.2.1 Population Level

As discussed in Chapter Four, we find that Spn invasive clusters are significantly more likely to be comprised of individuals from at-risk populations (i.e., PEH or PWID) (Figure 4.1). This agrees with previous work showing that PEH have higher rates of Spn carriage and are more likely to develop invasive disease compared to the general population (Lemay et al., 2019, Mosites et al., 2019, Ly et al., 2021). It also dovetails with research demonstrating that PEH and PWID are more likely to live in overcrowded housing environments that promote disease transmission (Moffa et al., 2019). At the population level, we hypothesize that shorter duration; higher attack rate serotypes are more likely to cause IPD clusters because they rely on high population density environments to spread to new hosts (described in Chapter Three). Thus, clusters of invasive disease are more likely to occur within disadvantaged communities because these communities tend to live in crowded and inadequately ventilated conditions (Moffa et al., 2019, Badiaga et al., 2008).

Support for this hypothesis comprises most of the work presented in this dissertation. In Chapter Four we described how Spn serotypes vary widely in their propensity to cause invasive clusters (Figure 4.1) but that serotypes with shorter carriage durations and higher attack rates are positively associated with clustering proportion (Figure 4.2A and 4.2B respectively). Our analysis also supports a causal association of several social determinants of health linked with crowding and income to invasive clustering. In Figure 4.4 we show that percent of households commuting by bus within a census tract, census tract population density, census tract average household income, and

households living in crowded housing units within census tract are all significantly predictive of IPD clustering. The associations between public transit and household crowding with invasive clustering (Chapter Four, Figures 4.4A, 4.4B) provide support that spending extended amounts of time in crowded and inadequately ventilated environments play a role in invasive cluster transmission by acting as a nexus point where Spn strains can spread more widely within disadvantaged communities.

Here we underline that these crowded, high-density environments may also differentially select for transmission of shorter duration and more highly invasive Spn serotypes. The relationship between household crowding and public transit with clustering propensity also connects to the results of Chapter Three demonstrating density-dependent transmission in Spn (Figures 3.4). In Chapter Three we find that shorter duration serotypes are more likely to cause invasive disease in higher population density areas (Figure 3.5). This may be because shorter duration serotypes do not spread as well in lower density areas due to having less time to successfully transmit to a new host as compared to their longer duration counterparts. A similar mechanism may be driving the relationship between IPD clustering and health inequality measures. Spending extended amounts of time in crowded unventilated environments may provide opportunities for shorter duration and more highly invasive serotypes to successfully transmit within disadvantaged populations. This mechanism would explain the correlations we observe between both serotype carriage duration and attack rate with IPD clustering percentage (Figure 4.2).

5.2.2 *Within-host Level*

At the within-host level, we hypothesize that delays in innate immune response and diminished phagocytic efficiency play a critical role in determining whether a carriage acquisition progresses into invasive disease or not. In a normal course of events, host innate immunity will begin mounting a response soon after *Streptococcus pneumoniae* colonization (Figure 5.2). Prompt and efficient response from Neutrophils and Macrophages will blunt Spn growth and epithelial damage caused by inflammation will be limited (Weight et al., 2019, Jochems et al., 2017, Zivich et al., 2018, Weiser et al., 2018). This leads to Spn control and eventual clearance from the host (Figure 5.2A). However, in communities experiencing disadvantage (e.g., people experiencing homelessness) where neutrophil delay and diminished effectiveness have been documented, we hypothesize epithelial damage caused by an overshoot in inflammation response can lead to invasion. Without a prompt innate response, Spn can grow unimpeded generating higher bacterial densities which have been previously associated with IPD (Wolter et al., 2014, Carrol et al., 2007). As the innate immunity cells begin to lyse Spn and release pneumococcal toxins the system shifts towards the positive feedback loop (wider green arrows) where the epithelial cell damage caused by toxin release and innate inflammation promotes more innate cell recruitment leading to further epithelial damage and so on eventually leading to invasion (Figure 5.2B).

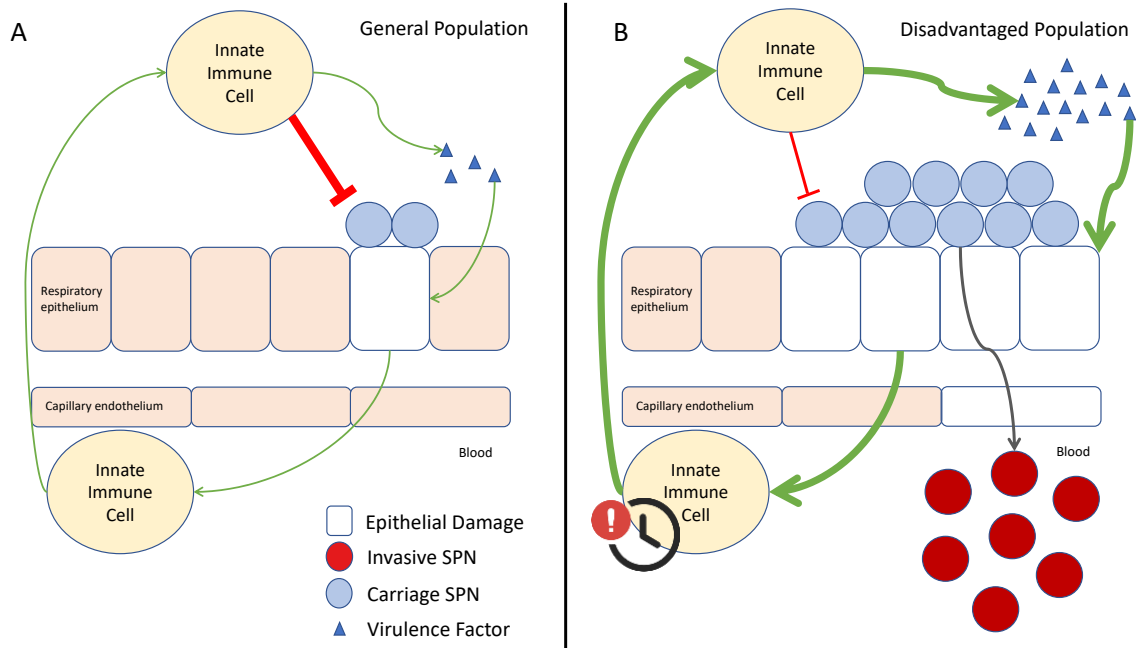


Figure 5.2: We propose that the progression to IPD is driven by a positive feedback loop between invasive inflammation and pneumococcal infection. The width of the green activation and red inhibition symbols represent the magnitude of the effect. Innate immune cells engage pneumococcus in a coupled dynamical system structured around inhibition of pneumococcal growth by innate immune response (red inhibition symbol) and a potential positive feedback loop (clockwise circuit of arrow symbols). (A) Within the general population, the interaction between innate immunity and Spn is shifted towards inhibition (wider red inhibition symbol) that leads to Spn control and eventual clearance. (B) Within vulnerable populations the delay of innate cell recruitment to the colonization site and diminished inhibition capability (thinner red arrow) leads to uncontrolled growth of Spn.

To understand why disadvantaged populations are more likely to be associated with invasive clusters we need to know how the invasion process differs between PEH and PWID versus the general population. Previous work discussed in Chapter Two has shown that when invasion occurs, it tends to occur quickly after Spn acquisition and that serotypes vary in their propensity to cause invasive disease (Sleeman et al., 2006, Domínguez-

Hüttinger et al., 2017). The timing of invasion suggests that the innate immune response can play a critical role in modulating the progression into invasive disease depending on how it initially counters Spn colonization.

The innate branch of the immune system recognizes pneumococcus through pattern recognition receptors (PRR) which primarily activate inflammatory mediators and cytokines (Koppe et al., 2012). The PRR-regulated cytokines then stimulate recruitment of neutrophils and macrophages. It has been shown that pneumococcus can manipulate innate immune response for its own purposes (Marriott et al., 2008). For example, pneumolysin (PLY), a key pneumococcal cytolysin, generates both innate pro- and anti-inflammatory effects. Previous work has revealed that PLY promotes inflammation by means of neutrophil recruitment but can also lower inflammatory signals in through its binding with mannose receptor C type 1 (MRC-1) (Subramanian et al., 2019, Weiser et al., 2018). Work by Zafar *et al* found that promoting inflammation can aid Spn in transmitting to a new host by increasing pneumococcal shedding (Zafar et al., 2017).

There has been some work interrogating how the innate immune responses differ between disadvantaged groups and the general population. A case control study by Arranz *et al*, found that there was a strongly suppressed immune response in the homeless group, with decreased adherence, chemotaxis, phagocytosis, superoxide levels, lymphoproliferation and NK activity (Arranz et al., 2009). IL-2 and plasma antioxidant levels were also impaired (Figure 5.3) (Arranz et al., 2009). Notably, they found that neutrophil chemotaxis appears to be deteriorated in the homeless population as well as a strong functional decline in phagocytic capacity.

The effect of delayed immune response on Spn disease progression was analyzed in previous work that showed perturbing the recruitment of both neutrophils and macrophages was associated with pneumococcal outgrowth and mortality (Paudel et al., 2019, Santos and Vera, 2020). By inhibiting CXCR2, a regulator of inflammatory neutrophil and exudate macrophage recruitment, Herbold *et al* found that reducing neutrophil recruitment 10 to 25% resulted in increased mortality of *S. pneumoniae* infected mice (Herbold et al., 2010). The previous studies by Arranz *et al* and Herbold *et al* supports the hypothesis in that a delay in neutrophil response is more likely to occur in PEH and may induce the positive feedback loop that damages the epithelial layer and eventually leads to Spn invasion.

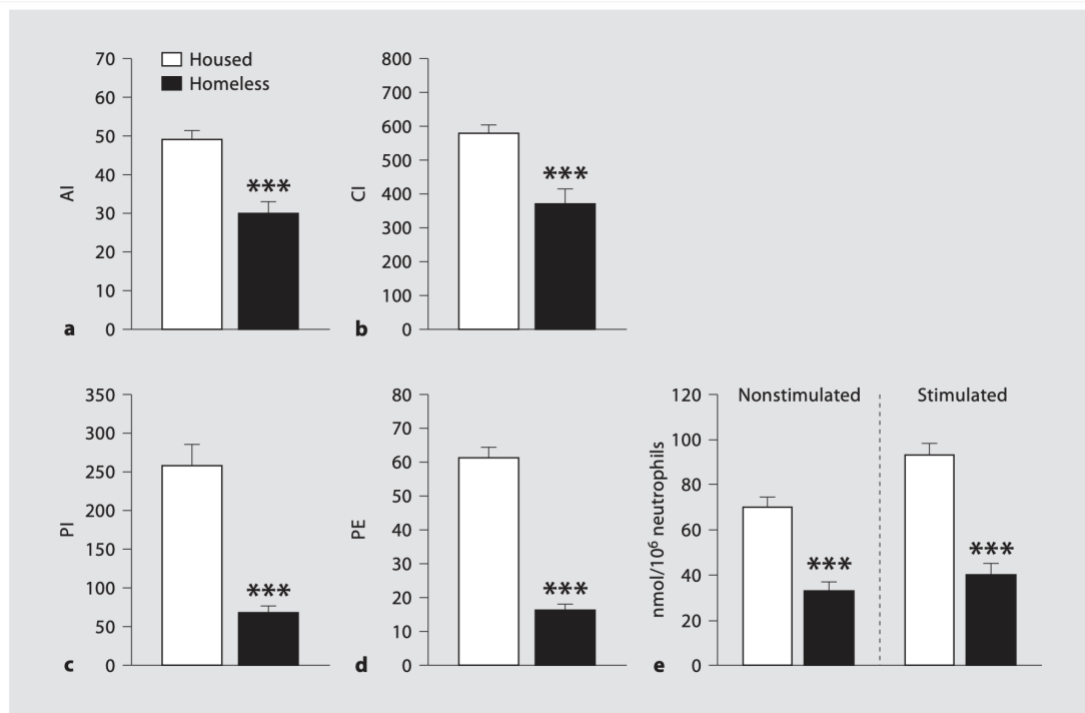


Figure 5.3: People experiencing homeless (PEH) have suppressed immunological responses compared to housed individuals, as measured by an array of neutrophil functional assays. In a previously published case-control study, Arranz *et al* found significant differences in innate immune response between unhoused individuals and housed persons matched for gender and age. Using human peripheral blood neutrophil samples, they found that Neutrophil adherence (a), neutrophil chemotaxis (b), neutrophil phagocytosis (c, d) and non-stimulated and stimulated superoxide anion levels (e) were all significantly higher in housed versus unhoused individuals. AI = Adherence index; CI = chemotaxis index; PI = phagocytic index; PE = phagocytic efficiency.

5.3 Proposed Future Work

To test our overarching hypotheses, I propose to (1) classify the genomic determinants associated with invasive clustering, and (2) quantify the impact of delayed innate immune response on pneumococcal invasion across multiple serotypes sourced from the CDC Streptococcus strain bank. Figure 5.1 provides an overview schematic of how the two aims interact.

5.3.1 Identifying genes and loci associated with invasive clustering

To identify genes and loci associated with invasive clustering we will employ two tools, Scoary and Pyseer, that will perform bacterial genome-wide association studies. To find cluster-associated genes we will first construct a pan-genome of our Spn isolate collection using the Roary v3.11.2 software and then employ Scoary v1.6.16 to carry out a series of pairwise comparisons using an algorithm that will control for bacterial population structure (Figure 5.4) (Brynildsrud et al., 2016, Page et al., 2015). To control for the multiple comparisons executed at this step, we will use the Benjamini-Hochberg correction. As a way to complement the Scoary analysis and to gain additional resolution

beyond the gene level, we will look for genomic loci associated with clustering via the Pyseer v1.3.3 tool (Lees et al., 2018). Pyseer will use unitigs (sequence elements built from kmers) representing the genomic variation across the pneumococcal isolate set to identify cluster-associated loci (Jaillard et al., 2018). Pyseer provides a number of association models that control for lineage effects by either treating them as fixed effects in a generalized linear model or as random effects within a linear mixed model. Our hypothesis anticipates that genes or genomic loci associated with clustering would likely be involved in the adherence, colonization, and inflammatory processes of pneumococcal infection. We believe the presence/absence of virulence factors or allelic variation within these virulence genes would be especially likely to correlate with IPD clustering.

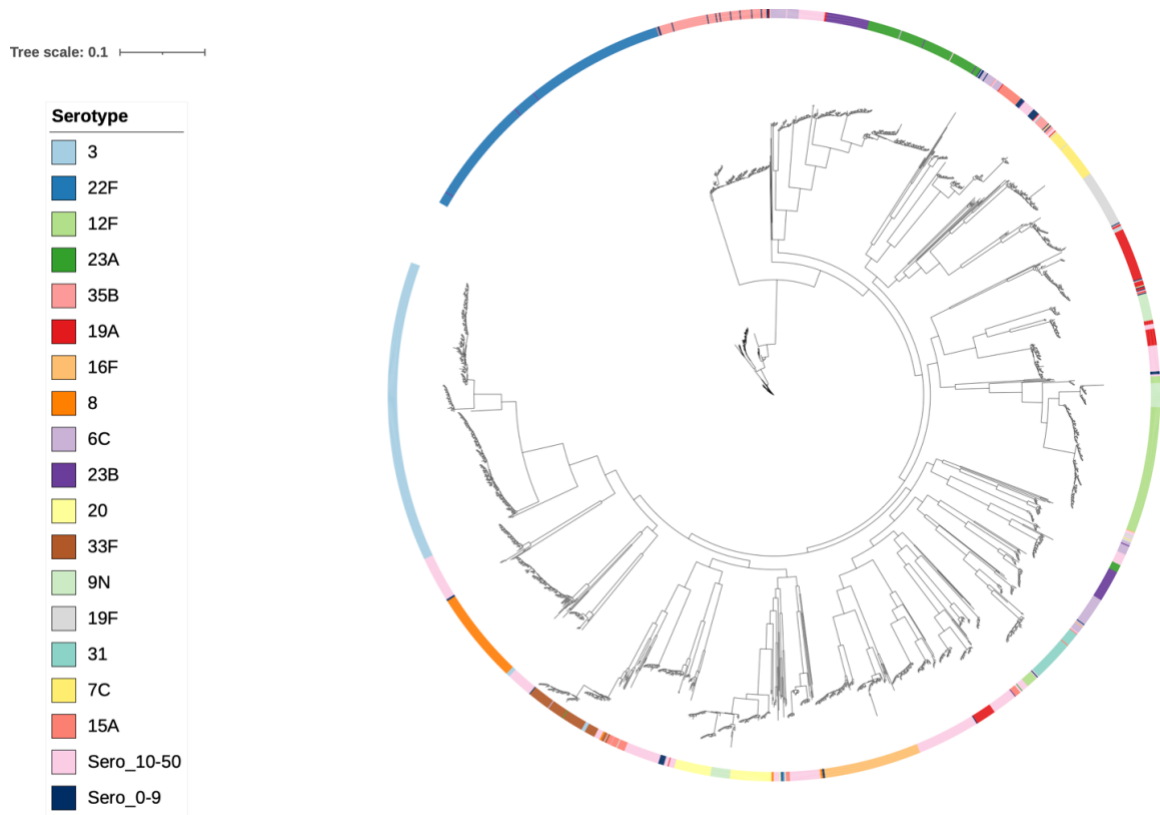


Figure 5.4: Population structure of *Streptococcus pneumoniae*. Phylogeny was constructed from 2,416 IPD genomic isolates acquired through the CDC Active Bacterial Core surveillance (ABCs) program during 2015. The deep branching lineages apparent across the tree demonstrate the importance of controlling for population structure when conducting bacterial genome-wide association studies across serotypes or lineages.

5.3.2 *In vitro and in vivo model systems to assess the impact of Neutrophil delay*

Our goal is to use both in vitro and in vivo model systems to assess the impact that delaying Neutrophil recruitment (a key process of innate immune response) has on Spn invasion. We will test a range of pneumococcal serotypes and lineages obtained from the CDC Active Bacterial Core surveillance (ABCs) biobank (a collection of over 71000 Spn isolates sourced from ABCs) to measure how the impact of this delay varies across strains.

The in vitro model will adapt previously established protocols based on a transwell plate design that approximates neutrophil recruitment and epithelial migration towards the site of infection (Deng et al., 2018, Brissac et al., 2021, Moreland and Bailey, 2006). To measure invasion and transcytosis of pneumococcus we will first seed A549 cells (1×10^5) on the underside (apical side) of a 24 well PET transwell (8 μm pore membrane) and incubate to allow the cells to form a confluent monolayer. Cells will then be rinsed with PBS and infected with 500 μl of a 10^6 CFU/ml suspension of selected Spn strains, or mock infected (DMEM only). At 0, 3, and 6 hours after the 2 h infection incubation time, transwells will be placed in a new ultra-low binding 24 well plate (Corning, Costar) and 5×10^5 neutrophils will be added to the top (basolateral side) of each individual transwell membrane insert. The plates will then be incubated for 3 h. To measure the amount of invading intracellular bacteria, A549 cells will be lysed by addition of cold water and 15

min of incubation at 4°C before plating and CFU counting. Transcytosis will be quantified by plating, and CFU counting growth media collected from the lower chamber. Each neutrophil delay (0, 3, and 6 hours) will be replicated in triplicate. To test the effect of neutrophil delay on different Spn strains we will repeat this experiment using 10 different Spn serotypes that vary in both invasiveness, carriage duration and clustering propensity.

The in vivo mouse model will modify a methodology previously described by Zafar *et al.* and Herbold *et al.* to determine how slowing innate recruitment affects invasion for both high-propensity and low-propensity clustering serotypes (Zafar *et al.*, 2017, Herbold *et al.*, 2010). A cohort of 4-day-old pups will be divided into two experimental groups of 10 pups each. The neutrophil delay group will receive intraperitoneal applications of the CXCR2 antagonist SB-225002 (15 ug/g of body weight) at 12 and 18 hours prior to infection with *S. pneumoniae*. The second group will act as a control and will receive a mock application of 50 ul of Todd-Hewitt broth per mouse. Both groups will be intranasally inoculated with 2,000 CFUs of *S. pneumoniae* suspended in 3 mL of PBS. Bacterial shedding will be measured by tapping the nares of the pups over a TS agar plate supplemented with streptomycin (200 mg/mL) and the pups will be monitored for 10 days post infection to measure the rate of invasive disease onset within each group. This experiment will be run once with a high cluster-propensity serotype and again with a low cluster-propensity serotype. We will select high- and low-clustering representative serotypes from both the bioinformatics analysis and the transwell in-vitro experiment.

Given our hypothesis, we predict that neutrophil delay will increase the risk of invasion for all serotypes and that the magnitude of risk will correlate with the length of delay. The

invasion risk will also be differentially pronounced among short carriage duration serotypes that tend to be highly invasive and highly clustered. Together, the in vitro and in vivo modeling approaches will assess whether, and to what degree, a delayed innate immune response can explain the higher IPD risk observed in PEH (and marginalized communities overall) and what tipping points exist in the host-pathogen interaction that will lead to excessive epithelial damage and invasion.

APPENDIX A.SUPPLEMENTAL MATERIAL FOR CHAPTER 2

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,

$$S^* = \frac{d\tau + 1}{\beta\tau - \beta p\tau}$$

$$C^* = -\frac{fh(\beta(p-1)\tau + d\tau + 1)}{\beta(h(d\tau - fp\tau + f\tau + 1) + f(d\tau + p))}$$

$$I^* = \frac{f(d\tau + p)(\beta(p-1)\tau + d\tau + 1)}{\beta(p-1)\tau(h(d\tau - fp\tau + f\tau + 1) + f(d\tau + p))}$$

$$R^* = \frac{h(d\tau + 1)(\beta(p-1)\tau + d\tau + 1)}{\beta(p-1)\tau(h(d\tau - fp\tau + f\tau + 1) + f(d\tau + p))}$$

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 ($C^* * 100$). Incidence of IPD (I_{inc}) 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,

$$I_{new}^* = p\beta C^* S^* + dC^*$$

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,

$$I_{inc}^* = \frac{p\beta(C^*n) \frac{(S^*n)}{n} + d(C^*n)}{n} 52 * 100000 = (p\beta C^* S^* + dC^*) 5200000$$

Similarly, the incidence of carriage acquisition (C_{inc}) 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,

$$C_{new}^* = (1 - p)\beta C^* S^*$$

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

$$C_{inc}^* = \frac{(1 - p)\beta(C^*n) \frac{(S^*n)}{n}}{n} 52 * 100 = ((1 - p)\beta C^* S^*) 5200$$

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,

$$AR^* = \frac{p\beta C^* S^* + dC^*}{(1-p)\beta C^* S^*} 100000 = \frac{d\tau + p}{(1-p)(d\tau + 1)} 100000$$

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 τ_f is the focal serotype carriage duration,

$$IOR^* = \frac{I^*(\tau = \tau_f)C^*(\tau = 14)}{C^*(\tau = \tau_f)I^*(\tau = 14)}$$

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., p_0 , p_1). 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_1C_0 / C_1I_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,

$$IOR(\tau_1) = \frac{\tau_0(d\tau_1 + p)}{\tau_1(d\tau_0 + p)}$$

Increasing the commensal-specific carriage duration τ_1 will decrease IOR, as commensal stages will increase their relative representation in the population (Figure 2.3A). Note that with the d parameter set to 0, IOR written as a function of focal carriage duration simply becomes the reciprocal equation $IOR(\tau) = \frac{\tau_0}{\tau}$. 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 p_I or d_I in our model) will increase the odds ratio of disease. The effect of increasing the constant disease progression in the focal strain (d_I) on the odds ratio of disease can be readily identified by inspecting the gradient of IOR on d_I

$$IOR'(d_1) = \frac{p\tau_0(\tau - \tau_0)}{\tau(d\tau_0 + p)^2}$$

which is positive when $\tau_0 < \tau$. A similar analysis for the effect of the initial carriage risk alone (p_I) yields an outcome where, assuming $\tau_0 > \tau$, the gradient of OR on p_I will also be positive. Thus, increasing p_I , will lead to higher invasive odds ratios only when the reference carriage duration is sufficiently large

$$IOR'(p_1) = \frac{d\tau_0(\tau_0 - \tau)}{\tau(d\tau_0 + p)^2}$$

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

$$IOR'(\tau) = -\frac{p\tau_0}{\tau^2(d\tau_0 + p)}$$

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 2.1. Mathematically, attack rate is the sum of inflow terms in the invasive disease differential equation ($p\beta CS + dC$) over the inflow term in the carriage ODE ($(1 - p)\beta CS$). The endemic equilibrium solution for the attack rate yielded the following expression

$$AR(\tau) = -\frac{d\tau + p}{(p - 1)(d\tau + 1)}$$

Using the equation for $AR(\tau)$, 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, $AR'(\tau) = \frac{d}{(d\tau + 1)^2}$. 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 $AR(\tau)$ simplifies to $\frac{p}{1-p}$ 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

$$AR'(d) = \frac{\tau}{(d\tau + 1)^2}$$

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

$$AR'(p) = \frac{1}{(p - 1)^2}$$

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 A.2). Reconfiguring the Carriage state in this way changes the clearance rate to a sum of 2 independent exponentially distributed random variables with mean $2/\tau$ each, resulting in the same average total carriage duration τ . 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:

$$\frac{dC1}{dt} = (1 - p)\beta(C1 + C2)S - \left(\frac{2}{\tau} + d\right)C1$$

$$\frac{dC2}{dt} = \frac{2}{\tau}C1 - \left(d + \frac{2}{\tau}\right)C2$$

$$\frac{dI}{dt} = p\beta(C1 + C2)S + d(C1 + C2) - hI$$

$$\frac{dR}{dt} = \frac{2}{\tau}C2 - fR + hI$$

$$S = 1 - C1 - C2 - R - I$$

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: $p = 2.9 \times 10^{-4}$ (95%CI: 1.4 to 4.6×10^{-4}), $d = 0.0$ (CI: 0 to 1.4×10^{-5}), $\beta = 0.23$ (CI: 0.06 to 0.41), $f = 3.1 \times 10^{-3}$ (CI: 2.6×10^{-3} to 6.0×10^{-3}). Notably, assuming $d = 0$, the IOR and AR expressions simplify to τ_0/τ , and $p/(1-p)$ respectively. Allowing the two carriage clearance rates to vary (*i.e.*, setting the duration from $C1 \rightarrow C2$ to two weeks and the duration from $C2 \rightarrow R$ to $\tau - 2$ 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>) 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 representing 37 references) are more likely to be used than invasive attack rate (8 references) which is generated using longitudinal data. Across metric types, invasive odds ratio (18 references) was the most common measurement followed by invasive case-to-carrier ratio (10 references), invasive capacity (9 references), and lastly, invasive attack rate (8 references).

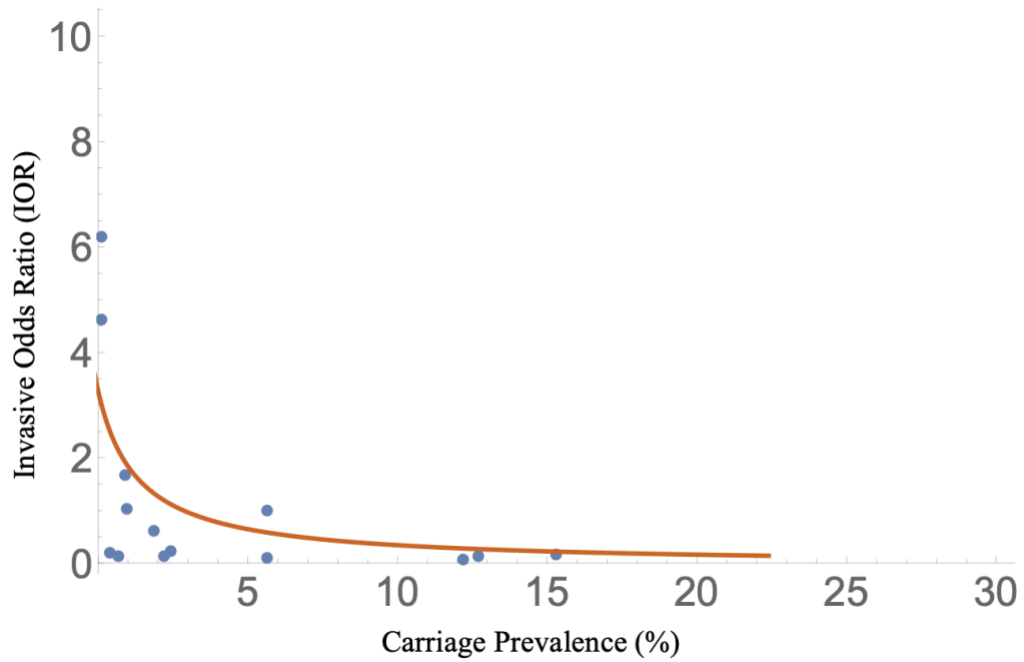


Figure A.1: Parametric plot of the fitted model of carriage prevalence against the fitted model of invasive OR with carriage duration as the independent variable. The plot shows an inverse correlation that matches the negative correlation observed in the cross-sectional data described in Brueggemann *et al.* The parameter τ_0 is set to 14 weeks.

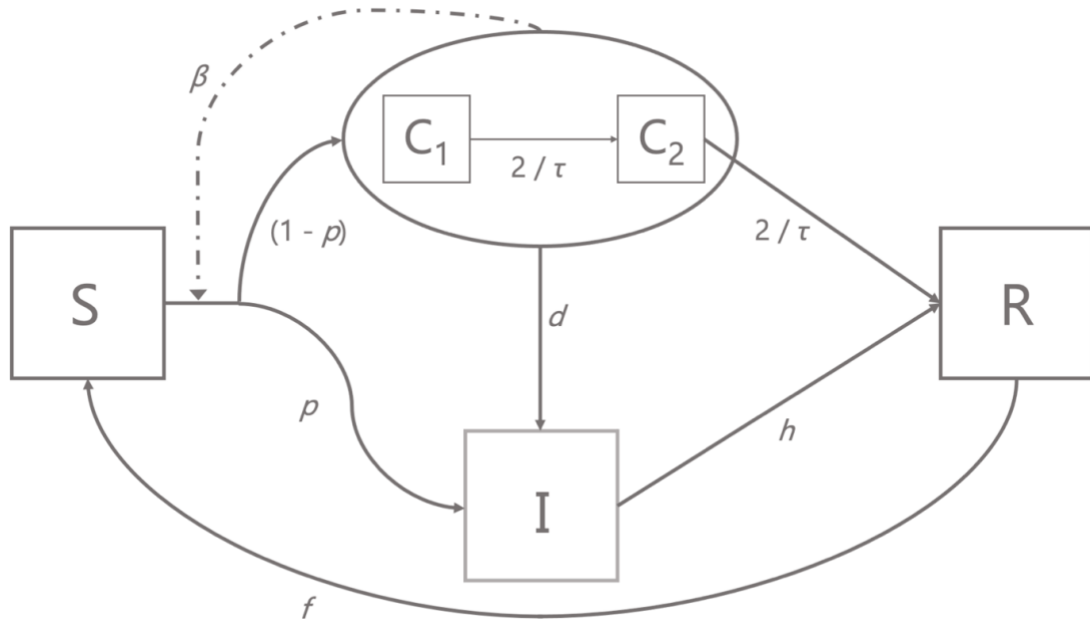


Figure A.2: Schematic diagram of the 2-stage carriage epidemiological model. Boxes represent proportions of hosts in mutually exclusive states: susceptible (S), infected asymptomatic carriers ($C1$ and $C2$), invasive (I) or recovered and immune (R). Solid arrows represent flows of individuals between 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 A.1: 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.*

Serotype	CROSS SECTIONAL DATA		LONGITUDINAL DATA			
	Carriage Prevalence (%)	Invasive Odds Ratio	Carriage Acquisition Incidence	Invasive Incidence	Attack Rate	Carriage Duration (Weeks)
5	0.113	6.19	0	0.16	75	4
1	0.113	4.61	0	0.74	75	4
8	0.378	0.211	1.78	0.43	30	12.2
7F	0.889	1.68	1.07	0.51	36	5.5
4	0.964	1.04	0.71	0.54	75	4
38	0.699	0.141	1.07	0.11	10	12
18C	1.89	0.608	4.99	1.1	24	13.5
3	2.19	0.121	5.35	0.4	9	6.2
33F	2.40	0.221	2.5	0	0	16.3
14	5.63	1.00	9.99	5.27	53	14
15B/C	5.63	0.101	5.35	0.04	1	6.8
6A	12.17	0.081	8.93	0.76	8	9.3
23F	12.68	0.131	16.05	1.21	8	16.7
6B	15.29	0.151	32.1	1.39	5	19.9

Table A.2: Model comparison of epidemiological models with different invasion timing. 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: $\Delta = \text{constant progression criterion value} - \text{early progression criterion value}$ and $p = \exp(-\Delta / 2)$.

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

APPENDIX B. SUPPLEMENTAL MATERIAL FOR CHAPTER 3

Epidemiological compartmental models require that a contact structure be specified. These contact structures are usually frequency-dependent or density-dependent. We performed a literature survey to estimate how much each contact structure is used in the field.

The first step in this analysis is to identify the papers that researchers use when citing density-dependent contact structures for their work. To accomplish this, we used the Publish or Perish tool (<https://harzing.com/resources/publish-or-perish>) to search for selected keywords associated with density-dependent compartmental models in Google Scholar and sort the matching papers by year. We generated a collection of 148 pneumococcal articles using the following keyword search terms:

```
"density dependence"|"density dependent" -"cell density" transmission susceptible  
"streptococcus pneumoniae"|pneumococcus|pneumococcal "compartmental  
model"|"epidemiological model"|"dynamic model"|"mathematical model"|"transmission  
model"
```

Using the R statistical programming language, we then filtered out all papers that didn't contain pneumoniae, pneumococcal, pneumococcus or PCV in the title:

```
setwd('/Users/ben_metcalf/Desktop/SPN-All_popDens_Analysis_6-18-  
2020/Spn_contact-struc_litReview_1-1-2023/')
```

```
df1 <- read.delim("Spn_contact-dens_Rev_PoP_qry-GQ6_1-5-2023.csv",header=TRUE,sep=',', fill=TRUE)
```

```
df2 <- df1[grepl("pneumoniae|pneumococcal|pneumococcus|pcv", tolower(df1$Title)),]
```

This left 5 remaining articles:

[1] "Pneumococcal capsule blocks protection by immunization with conserved surface proteins"

[2] "Frequency-dependent selection in vaccine-associated pneumococcal population dynamics"

[3] "Periodicity of epidemics of invasive disease due to infection with *Streptococcus pneumoniae* in the United States"

[4] "Estimating the transmission parameters of pneumococcal carriage in households"

[5] "Estimating the transmission parameters of pneumococcal"

Papers 1-3 didn't refer to density dependence within the context of epidemiological modeling and were removed. Papers 4 and 5 refer to the same study and is the only article to incorporate density dependence transmission although it was not at the population level. Their work focused on how household size impacted Spn spread. Finally, to validate the google scholar results, we performed a more comprehensive search of *Streptococcus pneumoniae* modeling papers using the PubMed citation database. We generated a collection of 100 pneumococcal articles using the following search terms:

((("streptococcus pneumoniae"[Title/Abstract]|pneumococcus[Title/Abstract]|pneumococcal[Title/Abstract]) AND ("compartmental model"[Title/Abstract]|"epidemiological model"[Title/Abstract]|"dynamic model"[Title/Abstract]|"mathematical model"[Title/Abstract]|"transmission model"[Title/Abstract]))

We then searched the 50 most recent papers for “frequency dependence|dependent” or “density dependence|dependent” keywords. We found no additional pneumococcal transmission papers that were employing a density dependent contact function.

Table B.1: Pneumococcal life-history traits and host population metadata aggregated by serotype. This set of observations was created by combining the longitudinal data from Sleeman et al. with population density data from United States Census Bureau's 2021 Planning Database. Table provides invasive case counts (IPD case count), locations sampled (State Count), census tracts sampled (Census Tract Count), average census tract population density (Avg Pop. Dens.), average census tract population size (Avg Pop. Size), per capita IPD rate (IPD Rate), and carriage duration.

Spn Serotype	IPD Case Count	State Count	Census Tract Count	Avg Pop. Dens. (sq. mi.)	Avg Pop. Size	IPD Rate (per capita)	Carriage Duration (weeks)
4	183	9	146	14790.50	5296.22	2.41E-04	4
12F	300	10	253	6725.08	5540.99	2.09E-04	4.9
7F	108	9	97	15037.27	6079.33	1.83E-04	5.5
15C	151	10	146	4918.19	5370.71	1.90E-04	5.9
3	1421	10	1060	4855.85	5233.52	2.25E-04	6.2
15B	182	10	173	4681.43	5477.11	1.88E-04	7.7
9V	15	7	15	4784.47	4333.27	2.31E-04	7.9
17F	120	10	116	4732.12	4776.48	2.16E-04	8.3
11A	383	10	355	4689.81	5236.60	2.00E-04	8.9
6A	13	6	13	5265.90	5106.15	1.96E-04	9.3
38	138	10	130	3635.66	5894.14	1.74E-04	12
23A	487	10	437	4465.81	5239.81	1.99E-04	12.2
8	416	10	339	4162.81	5081.04	2.20E-04	12.2
10A	189	10	181	4966.21	5396.66	1.89E-04	12.5

Table B.1 continued

20	295	10	257	5671.61	4994.63	2.17E-04	12.5
19A	367	10	341	5227.81	5140.32	2.05E-04	12.6
9N	565	10	499	4571.55	5292.31	2.04E-04	13
19F	346	10	325	4898.12	5583.79	1.88E-04	13.2
22F	1035	10	854	4347.90	5326.07	2.07E-04	13.2
18C	7	4	7	3771.06	4423.00	2.26E-04	13.5
14	13	6	13	9481.60	5355.38	1.87E-04	14
16F	385	10	359	4729.83	5056.40	2.06E-04	15
21	40	10	40	4131.21	5513.20	1.81E-04	15.4
33F	436	10	390	3406.75	5442.27	1.95E-04	16.3
23F	6	4	6	1224.26	4223.00	2.37E-04	16.7
15A	462	10	423	4362.58	5097.45	2.05E-04	17.3
6B	19	8	19	5860.13	4980.89	2.01E-04	19.9
35F	152	10	149	4434.52	5364.64	1.88E-04	31.9

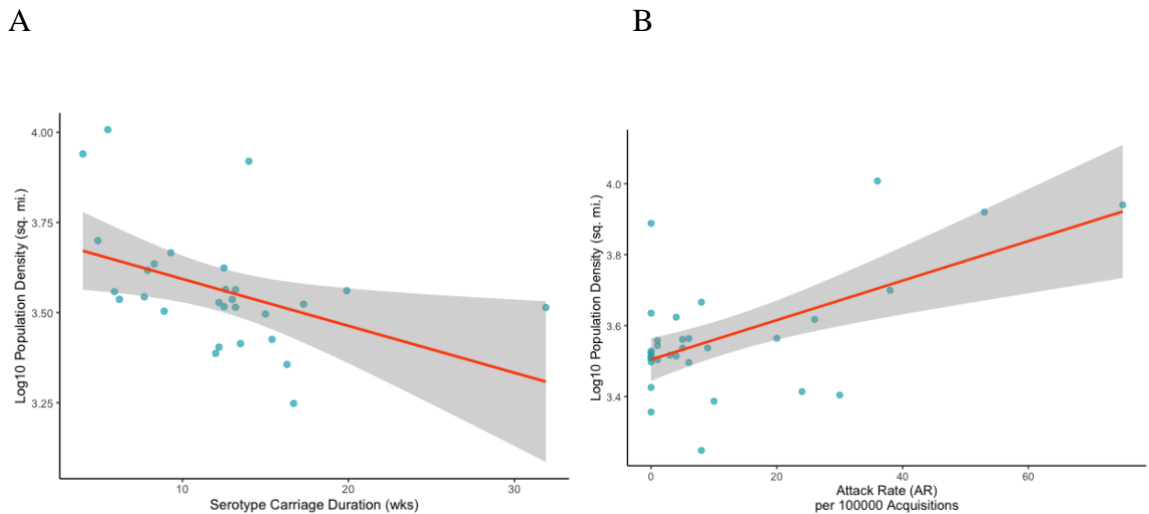


Figure B.1: Average serotype population density has a negative correlation with carriage duration and a positive correlation with invasive attack rate (AR). (A) Serotype carriage duration (blue dots) against log₁₀ average serotype population density (measured using ZCTAs) yields a significant inverse association (Kendall Tau = -0.385, p-value = 0.00415). (B) Invasive AR (blue dots) against log₁₀ average serotype population density (measured using ZCTAs) yields a significant positive association (Kendall Tau = 0.256, p-value = 0.048). These relationships indicate that shorter duration and more highly invasive serotypes are more likely to cause IPD in denser environments.

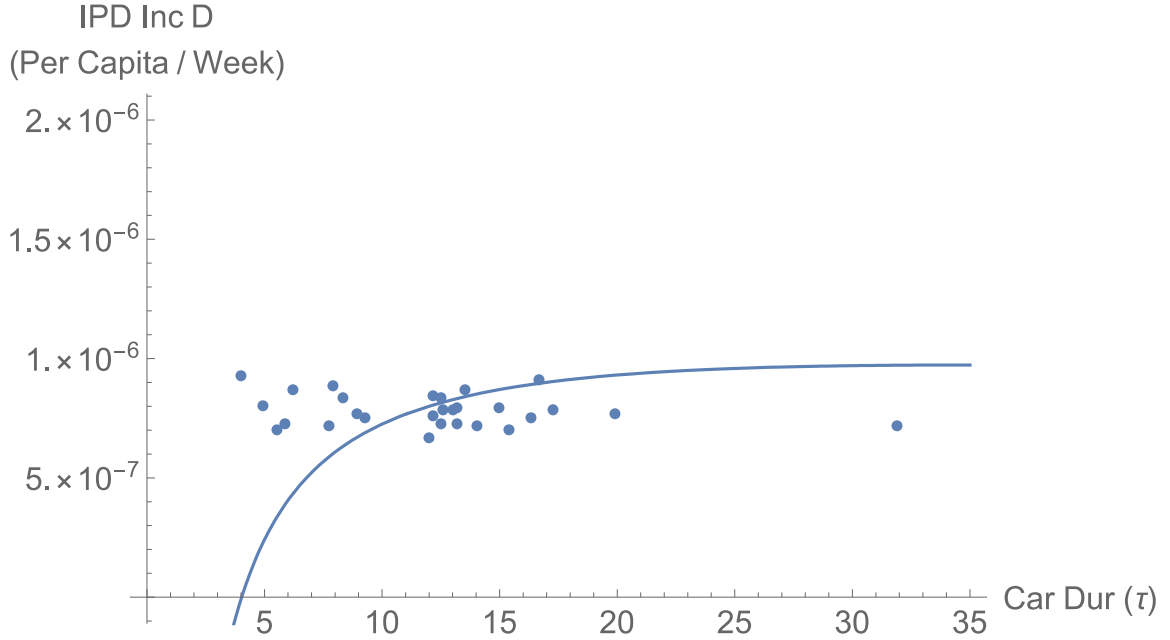
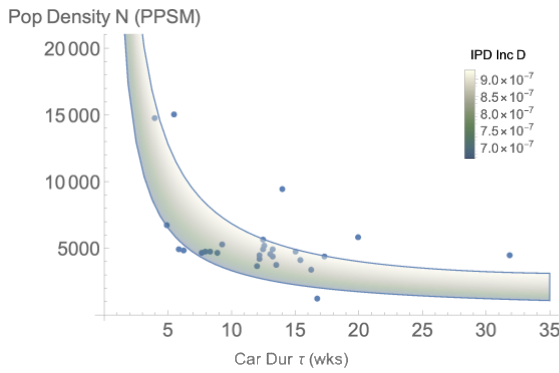


Figure B.2: Fitting invasive incidence using a power contact rate function provides support for Spn density dependent transmission. Spn per capita invasive incidence data in weeks (blue dots) and the model fit (blue line, $D_p^*(\tau, N) =$

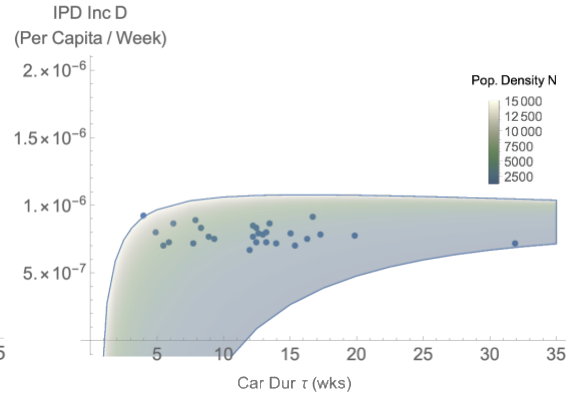
$$\begin{cases} -\frac{p(b_D h N (p-1)\tau + h)}{(b_D N \tau - b_D N p \tau)(f h - h p \tau + h \tau + p)}, & \tau < \kappa \\ \frac{h p (b_F (p-1)\tau + 1)}{b_F (p-1)\tau (f h - h p \tau + h \tau + p)}, & \tau \geq \kappa \end{cases}$$

) of invasive incidence as a function of carriage duration τ and population density N . Model fit supports the hypothesis Spn transmission is density-dependent and that shorter-duration serotypes are more likely to cause invasive disease in higher-density areas. Population density N was set to $N_\kappa = 3722.82$ PPSM, making the density- and frequency-dependent transmission coefficients equivalent resulting in a continuous curve across the two piecewise components.

A



B



C

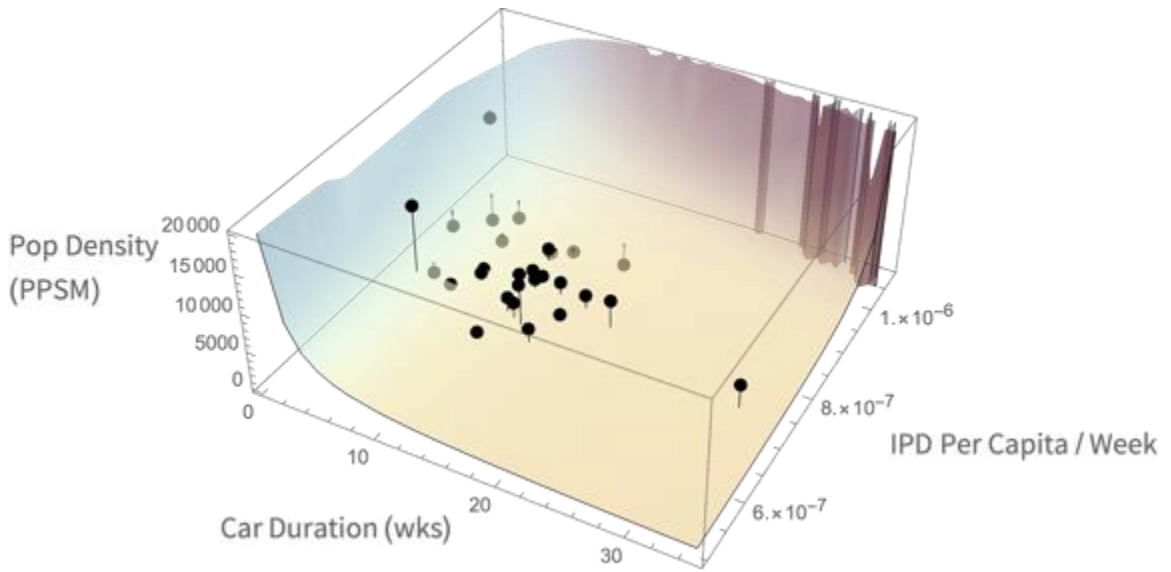


Figure B.3: Fitting population density and invasive incidence using a power contact rate function indicates Spn transmission is density dependent. A) Spn serotype average population density data (blue dots) and the model fit (green-to-white gradient,

$$N(\tau, D^*) = \left(\frac{b(1-p)\tau(fh-hp\tau+h\tau+p) \left(\frac{hp^2}{(1-p)(fh-hp\tau+h\tau+p)} + \frac{hp}{(1-p)(fh-hp\tau+h\tau+p)} - D^* \right)}{hp} \right)^{-1/r}$$

density as a function of carriage duration τ and equilibrium invasive incidence D^* . B)

Spn per capita invasive incidence data in weeks (blue dots) and the model fit (green-to-white gradient, $D^*(\tau, N) = \frac{pN^{-r}(bh(p-1)\tau N^r+h)}{b(p-1)\tau(fh-hp\tau+h\tau+p)}$) of invasive incidence as a function

of carriage duration τ and population density N . C) A 3D plot of both population density and invasive incidence as functions of carriage duration τ . The vertical lines connecting each datapoint to the fitted function represent the residual error. Both model fits support the hypothesis Spn transmission is density-dependent and that shorter-duration serotypes are more likely to cause invasive disease in higher-density areas. The model fitting estimated the power contact rate exponent $r = 0.988$ (i.e., Spn transmission is constrained by a nearly linear contact density function), the immune duration parameter $f = 237$, and the transmission rate constant parameter $b = 0.0000769$. Additional fixed parameter values are invasive clearance $h = 0.5$ and probability of invasion $p = 0.00029$.

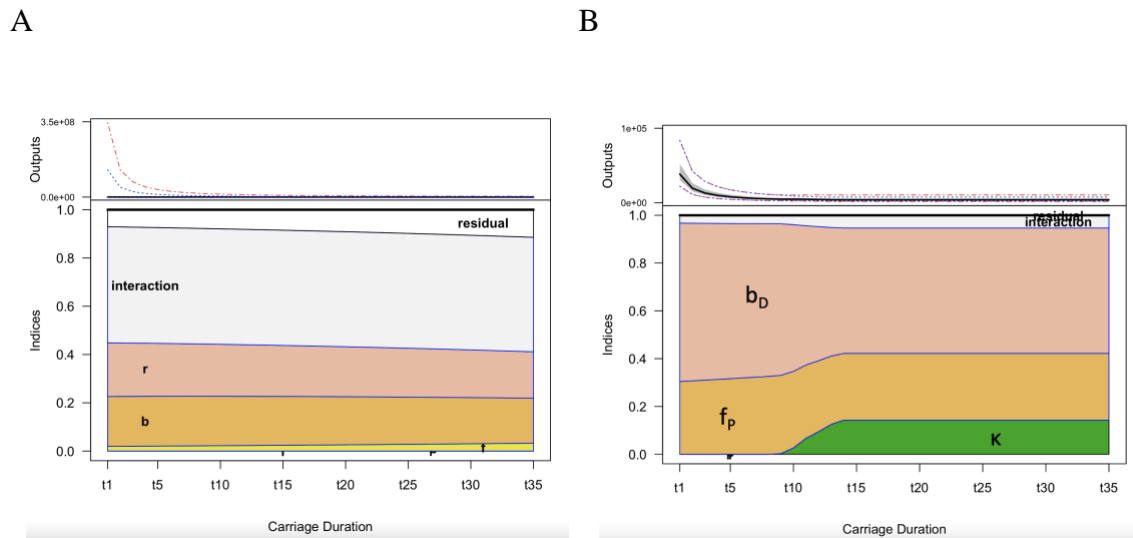


Figure B.4: Additional Sensitivity plots of the population density functions (equations 3.4 and 3.8). A) Dynamics of the sensitivity indices of the power contact rate density function (equation 3.4) from $\tau = 1$ to 35 reveals that the sensitivity of population density to the main effects of r , f , and b remain stable across the range of carriage durations. B) The sensitivity indices of the piecewise contact rate density function (equation 3.9) from $\tau = 1$ to 35 indicates the f_P and b_D parameters remain stable across carriage durations but sensitivity to the threshold parameter κ increases as the model transitions from density- to frequency-dependent transmission.

APPENDIX C.SUPPLEMENTAL MATERIAL FOR CHAPTER 4

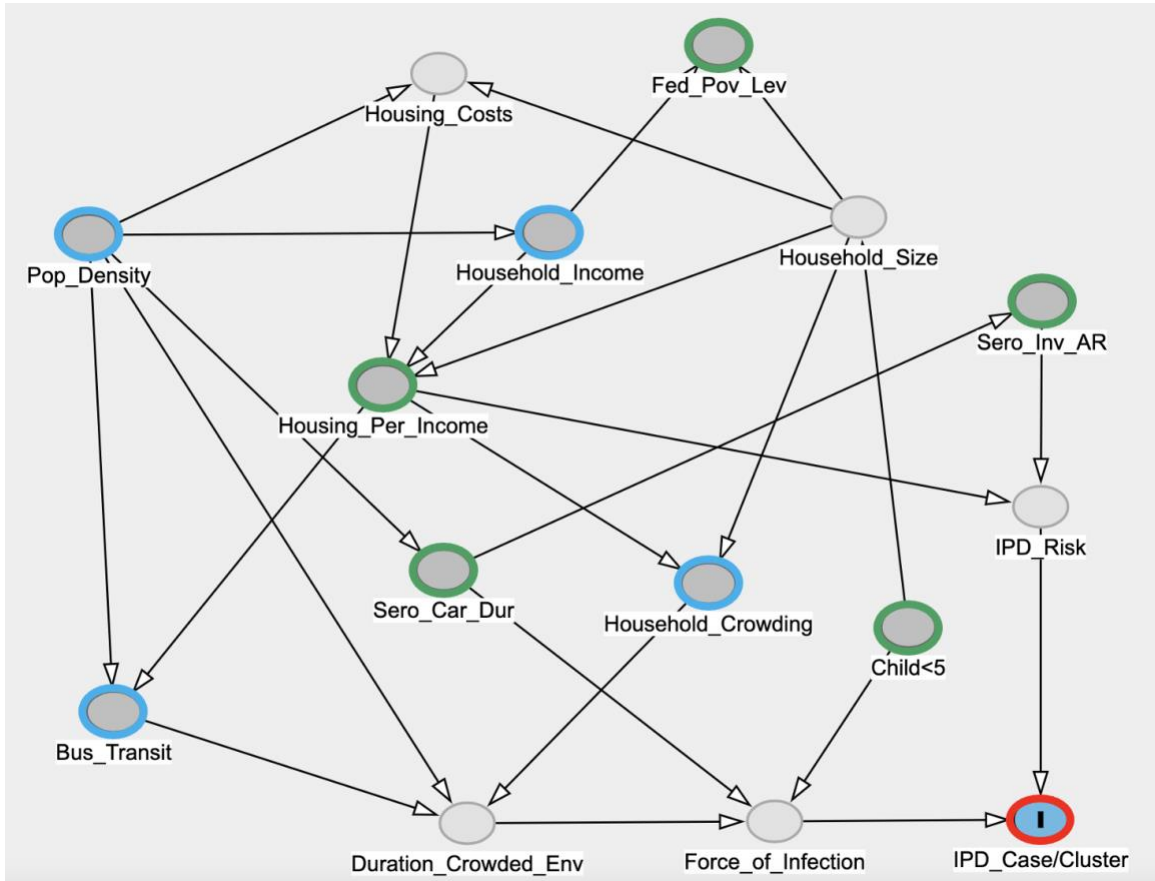
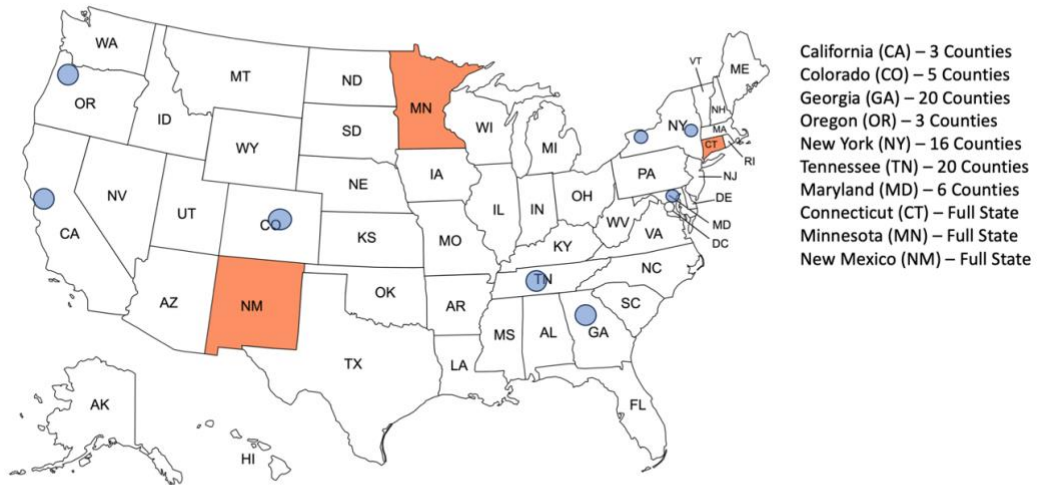


Figure C.1: The relationship between IPD outcome (case or cluster) with household crowding, bus transit, and household income is confounded. The directed acyclic graph (DAG) represents the causal interrelationships between variables (nodes) influencing the development of IPD and IPD clusters. Dark grey nodes are observed variables and light grey nodes are unobserved (latent) variables. Demographic features of interest are highlighted in blue and potential confounders are highlighted in green (Table 4.1). The IPD case and cluster outcome variable is highlighted in red.

Table C.1: Relationship between attribute and dependent variable (cluster cases or total cases) can vary depending on the causal model. Rows represent the variables of interest and columns represent a causal model choice. Cells shaded red indicate either a change in coefficient significance or sign compared to the baseline model,

DAG Model 1. Asterisks designate the following significance thresholds: * = <0.05, ** = <0.01, * = <0.001.**

Cluster Cases				
Variable	DAG Model 1 (Figure 4.3)	Univariate	Multivariate	DAG Model 2 (Figure C.1)
Population Density	0.238***	0.238***	0.113***	0.238***
Household Income	-0.548***	-0.548***	-0.446***	-0.488***
Mass Transit - Bussing	0.176***	0.298***	0.126***	0.176***
Household Crowding	0.0978***	0.207***	0.00307	0.119***
Total Cases				
Variable	DAG Model 1 (Figure 4.3)	Univariate	Multivariate	DAG Model 2 (Figure C.1)
Population Density	0.140***	0.140***	0.0143	0.140***
Household Income	-0.419***	-0.419***	-0.376***	-0.399***
Mass Transit - Bussing	0.178***	0.228***	0.134***	0.178***
Household Crowding	0.0452***	0.127***	-0.0367**	0.0463***



Created with mapchart.net

Figure C.2: The Active Bacterial Core surveillance catchment area. The blue areas are counties spread across 7 US states including San Francisco Bay Area, California (CA) (3 counties), Denver, Colorado (CO) (5 counties), Atlanta, Georgia (GA) (20 counties), Portland, Oregon (OR) (3 counties), Rochester and Albany, New York (NY) (16 counties), select urban counties in Tennessee (TN) (20 counties), and Baltimore area, Maryland (MD) (6 counties). The orange areas represent the three U.S. states Connecticut (CT), Minnesota (MN), and New Mexico (NM).

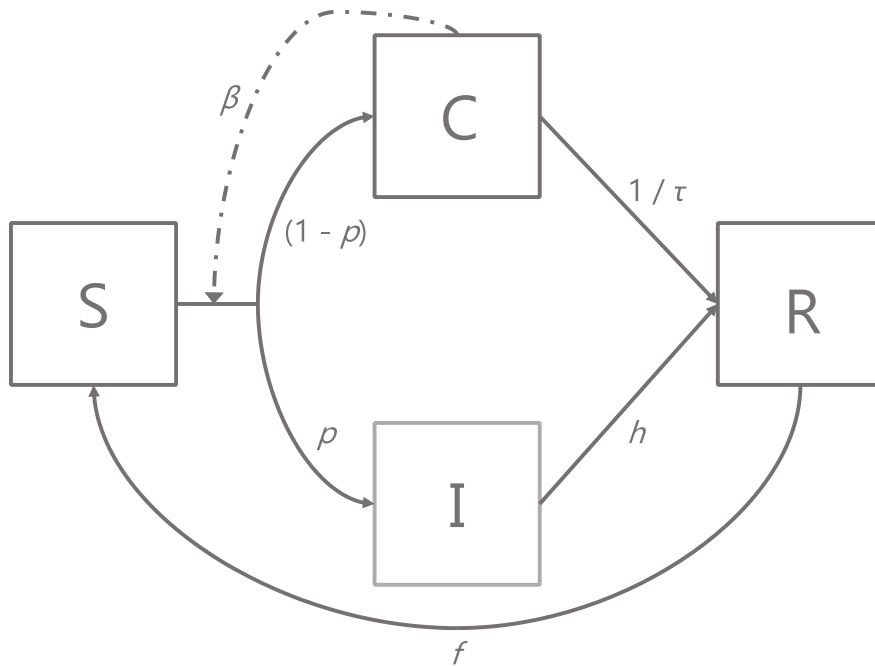


Figure C.3: Schematic diagram of the epidemiological model. Boxes represent proportions of hosts in mutually exclusive states: susceptible (S), infected asymptomatic carriers (C), invasive (I) or recovered and immune (R). Solid arrows represent flows of individuals between states, and dashed arrows represent factors influencing those flows. Equations describing the system are presented in Materials and Methods (methods equations 4.1), along with parameter definitions (Table C.2).

Table C.2: 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

Table C.2 continued

$R(t)$	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
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.

REFERENCES

- AIRGOOD-OBRYCKI, W. & RIEGER, S. 2019. Defining suburbs: How definitions shape the suburban landscape. *Joint Center of Housing Studies of Harvard University*.
- ALANEE, S. R., MCGEE, L., JACKSON, D., CHIOU, C. C., FELDMAN, C., MORRIS, A. J., ORTQVIST, A., RELLO, J., LUNA, C. M., BADDOUR, L. M., IP, M., YU, V. L., KLUGMAN, K. P. & 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.
- ALIZON, S., HURFORD, A., MIDEO, N. & VAN BAALEN, M. 2009. Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future. *J Evol Biol*, 22, 245-59.
- ALLEN, R. C., POPAT, R., DIGGLE, S. P. & BROWN, S. P. 2014. Targeting virulence: can we make evolution-proof drugs? *Nat Rev Microbiol*, 12, 300-8.
- ALTHOUSE, B. M., HAMMITT, L. L., GRANT, L., WAGNER, B. G., REID, R., LARZELERE-HINTON, F., WEATHERHOLTZ, R., KLUGMAN, K. P., RODGERS, G. L., O'BRIEN, K. L. & HU, H. 2017. Identifying transmission routes of *Streptococcus pneumoniae* and sources of acquisitions in high transmission communities. *Epidemiol Infect*, 145, 2750-2758.
- ANDERSON, R. M. & MAY, R. M. 1991. *Infectious diseases of humans : dynamics and control*, Oxford ; New York, Oxford University Press.
- ARRANZ, L., DE VICENTE, A., MUÑOZ, M. & DE LA FUENTE, M. 2009. Impaired immune function in a homeless population with stress-related disorders. *Neuroimmunomodulation*, 16, 251-60.
- BADIAGA, S., RAOULT, D. & BROUQUI, P. 2008. Preventing and controlling emerging and reemerging transmissible diseases in the homeless. *Emerg Infect Dis*, 14, 1353-9.
- 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.
- BARTLETT, M. S. 1960. The critical community size for measles in the United States. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 123, 37-44.

- BEALL, B., CHOCHUA, S., LI, Z., TRAN, T., VARGHESE, J., MCGEE, L., LI, Y. & METCALF, B. 2022a. Invasive pneumococcal disease clusters disproportionately impact persons experiencing homelessness, injecting drug users, and the western United States. *J Infect Dis*.
- BEALL, B., CHOCHUA, S., LI, Z., TRAN, T., VARGHESE, J., MCGEE, L., LI, Y. & METCALF, B. J. 2022b. Invasive Pneumococcal Disease Clusters Disproportionally Impact Persons Experiencing Homelessness, Injecting Drug Users, and the Western United States. *J Infect Dis*, 226, 332-341.
- BEGON, M., BENNETT, M., BOWERS, R. G., FRENCH, N. P., HAZEL, S. M. & TURNER, J. 2002. A clarification of transmission terms in host-microparasite models: numbers, densities and areas. *Epidemiol Infect*, 129, 147-53.
- BERSCH, B., BOUGAULT, C., ROUX, L., FAVIER, A., VERNET, T. & DURMORT, C. 2013. New insights into histidine triad proteins: solution structure of a *Streptococcus pneumoniae* PhtD domain and zinc transfer to AdcAII. *PLoS One*, 8, e81168.
- BERTHIER, K., LANGLAIS, M., AUGER, P. & PONTIER, D. 2000. Dynamics of a feline virus with two transmission modes within exponentially growing host populations. *Proc Biol Sci*, 267, 2049-56.
- BJORNSTAD, O. A. F. B. A. G. B. 2002. Dynamics of Measles Epidemics: Estimating Scaling of Transmission Rates Using a Time Series SIR Model. *Ecological Monographs*, 72, 169-184.
- BLOMBERG, C., DAGERHAMN, J., DAHLBERG, S., BROWALL, S., FERNEBRO, J., ALBIGER, B., MORFELDT, E., NORMARK, S. & HENRIQUES-NORMARK, B. 2009. Pattern of accessory regions and invasive disease potential in *Streptococcus pneumoniae*. *J Infect Dis*, 199, 1032-42.
- BORREMANS, B., REIJNIERS, J., HENS, N. & LEIRS, H. 2017. The shape of the contact-density function matters when modelling parasite transmission in fluctuating populations. *R Soc Open Sci*, 4, 171308.
- BRISSAC, T., MARTÍNEZ, E., KRUCKOW, K. L., RIEGLER, A. N., GANAIE, F., IM, H., BAKSHI, S., ARROYO-DIAZ, N. M., SPENCER, B. L., SAAD, J. S., NAHM, M. H. & ORIHUELA, C. J. 2021. Capsule Promotes Intracellular Survival and Vascular Endothelial Cell Translocation during Invasive Pneumococcal Disease. *mBio*, 12, e0251621.
- BRUEGGEMANN, A. B., GRIFFITHS, D. T., MEATS, E., PETO, T., CROOK, D. W. & SPRATT, B. G. 2003. Clonal relationships between invasive and carriage *Streptococcus pneumoniae* and serotype- and clone-specific differences in invasive disease potential. *J Infect Dis*, 187, 1424-32.

- BRUEGGEMANN, A. B., PETO, T. E., CROOK, D. W., BUTLER, J. C., KRISTINSSON, K. G. & SPRATT, B. G. 2004. Temporal and geographic stability of the serogroup-specific invasive disease potential of *Streptococcus pneumoniae* in children. *J Infect Dis*, 190, 1203-11.
- BRYNILDSDRUD, O., BOHLIN, J., SCHEFFER, L. & ELDHOLM, V. 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol*, 17, 238.
- BURCKHARDT, I., PANITZ, J., BURCKHARDT, F. & ZIMMERMANN, S. 2017. Identification of *Streptococcus pneumoniae*: development of a standardized protocol for optochin susceptibility testing using total lab automation. *BioMed research international*, 2017.
- BUREAU, U. C. 2021. Historical population density data (1910–2020).
- BURNHAM, K. P., ANDERSON, D. R. & HUYVAERT, K. P. 2011. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral ecology and sociobiology*, 65, 23-35.
- CARROL, E. D., GUIVER, M., NKHOMA, S., MANKHAMBO, L. A., MARSH, J., BALMER, P., BANDA, D. L., JEFFERS, G., WHITE, S. A., MOLYNEUX, E. M., MOLYNEUX, M. E., SMYTH, R. L., HART, C. A. & GROUP, I. S. 2007. High pneumococcal DNA loads are associated with mortality in Malawian children with invasive pneumococcal disease. *Pediatr Infect Dis J*, 26, 416-22.
- CASSIR, N., LUCIANI, L., FERRIEUX, D., LEVASSEUR, A., FOURNIER, P. E. & PAROLA, P. 2021. Pneumococcal pneumonia among shipyard workers: Inside the features of disease onset. *Travel Med Infect Dis*, 44, 102183.
- CHAGUZA, C., ANDAM, C. P., HARRIS, S. R., CORNICK, J. E., YANG, M., BRICIO-MORENO, L., KAMNG'ONA, A. W., PARKHILL, J., FRENCH, N., HEYDERMAN, R. S., KADIOGLU, A., EVERETT, D. B., BENTLEY, S. D. & HANAGE, W. P. 2016. Recombination in *Streptococcus pneumoniae* Lineages Increase with Carriage Duration and Size of the Polysaccharide Capsule. *mBio*, 7.
- CHAGUZA, C., SENGHORE, M., BOJANG, E., LO, S. W., EBRUKE, C., GLADSTONE, R. A., TIENCHEU, P. E., BANCROFT, R. E., WORWUI, A., FOSTER-NYARKO, E., CEESAY, F., OKOI, C., MCGEE, L., KLUGMAN, K. P., BREIMAN, R. F., BARER, M. R., ADEGBOLA, R. A., ANTONIO, M., BENTLEY, S. D. & KWAMBANA-ADAMS, B. A. 2020. Carriage Dynamics of Pneumococcal Serotypes in Naturally Colonized Infants in a Rural African Setting During the First Year of Life. *Front Pediatr*, 8, 587730.
- CHANDRA, S. A. K.-N. E. A. K. G. A. V. J. 2013. A geographic analysis of population density thresholds in the influenza pandemic of 1918–19. *International journal of health geographics*, 12, 9.

- CROSS, P. C., CREECH, T. G., EBINGER, M. R., MANLOVE, K., IRVINE, K., HENNINGSEN, J., ROGERSON, J., SCURLOCK, B. M. & CREEL, S. 2013. Female elk contacts are neither frequency nor density dependent. *Ecology*, 94, 2076-86.
- CROSS, P. C., HEISEY, D. M., SCURLOCK, B. M., EDWARDS, W. H., EBINGER, M. R. & BRENNAN, A. 2010. Mapping brucellosis increases relative to elk density using hierarchical Bayesian models. *PLoS One*, 5, e10322.
- DAWOOD, F. S., AMBROSE, J. F., RUSSELL, B. P., HAWKSWORTH, A. W., WINCHELL, J. M., GLASS, N., THURMAN, K., SOLTIS, M. A., MCDONOUGH, E., WARNER, A. K., WESTON, E., CLEMMONS, N. S., ROSEN, J., MITCHELL, S. L., FAIX, D. J., BLAIR, P. J., MOORE, M. R. & LOWERY, J. 2011. Outbreak of pneumonia in the setting of fatal pneumococcal meningitis among US Army trainees: potential role of *Chlamydia pneumoniae* infection. *BMC Infect Dis*, 11, 157.
- DENG, Y., HERBERT, J. A., SMITH, C. M. & SMYTH, R. L. 2018. An in vitro transepithelial migration assay to evaluate the role of neutrophils in Respiratory Syncytial Virus (RSV) induced epithelial damage. *Sci Rep*, 8, 6777.
- DEVINE, V. T., CLEARY, D. W., JEFFERIES, J. M., ANDERSON, R., MORRIS, D. E., TUCK, A. C., GLADSTONE, R. A., O'DOHERTY, G., KURUPARAN, P., BENTLEY, S. D., FAUST, S. N. & CLARKE, S. C. 2017. The rise and fall of pneumococcal serotypes carried in the PCV era. *Vaccine*, 35, 1293-1298.
- DOERN, C. D. & CAREY-ANN, B. D. 2010. It's not easy being green: the viridans group streptococci, with a focus on pediatric clinical manifestations. *Journal of clinical microbiology*, 48, 3829-3835.
- DOMENECH DE CELLÈS, M., OPATOWSKI, L., SALOMON, J., VARON, E., CARBON, C., BOËLLE, P. Y. & GUILLEMOT, D. 2011. Intrinsic epidemicity of *Streptococcus pneumoniae* depends on strain serotype and antibiotic susceptibility pattern. *Antimicrob Agents Chemother*, 55, 5255-61.
- DOMÍNGUEZ-HÜTTINGER, E., BOON, N. J., CLARKE, T. B. & TANAKA, R. J. 2017. Mathematical Modeling of *Streptococcus pneumoniae* Colonization, Invasive Infection and Treatment. *Front Physiol*, 8, 115.
- DONATI, C., HILLER, N. L., TETTELIN, H., MUZZI, A., CROUCHER, N. J., ANGIUOLI, S. V., OGGIONI, M., DUNNING HOTOPP, J. C., HU, F. Z., RILEY, D. R., COVACCI, A., MITCHELL, T. J., BENTLEY, S. D., KILIAN, M., EHRLICH, G. D., RAPPUOLI, R., MOXON, E. R. & MASIGNANI, V. 2010. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol*, 11, R107.
- DONKOR, E. S., STABLER, R. A., HINDS, J., ADEGBOLA, R. A., ANTONIO, M. & WREN, B. W. 2012. Comparative phylogenomics of *Streptococcus pneumoniae*

- isolated from invasive disease and nasopharyngeal carriage from West Africans. *BMC Genomics*, 13, 569.
- DOWELL, S. F., WHITNEY, C. G., WRIGHT, C., ROSE, C. E. & SCHUCHAT, A. 2003. Seasonal patterns of invasive pneumococcal disease. *Emerg Infect Dis*, 9, 573-9.
- DRIJKONINGEN, J. J. & ROHDE, G. G. 2014. Pneumococcal infection in adults: burden of disease. *Clin Microbiol Infect*, 20 Suppl 5, 45-51.
- EDWARDS, J. M., WATSON, N., FOCHT, C., WYNN, C., TODD, C. A., WALTER, E. B., HEINE, R. P. & SWAMY, G. K. 2019. Group B Streptococcus (GBS) colonization and disease among pregnant women: a historical cohort study. *Infectious diseases in obstetrics and gynecology*, 2019.
- FEIKIN, D. R. & KLUGMAN, K. P. 2002. Historical changes in pneumococcal serogroup distribution: implications for the era of pneumococcal conjugate vaccines. *Clin Infect Dis*, 35, 547-55.
- FERRARI, M. J., PERKINS, S. E., POMEROY, L. W. & BJØRNSTAD, O. N. 2011. Pathogens, social networks, and the paradox of transmission scaling. *Interdiscip Perspect Infect Dis*, 2011, 267049.
- FINKENSTÄDT, B., KEELING, M. & GRENFELL, B. 1998. Patterns of density dependence in measles dynamics. *Proc Biol Sci*, 265, 753-62.
- GARRETT, T. 2007. Economic Effects of the 1918 Influenza Pandemic Implications for a Modern-day Pandemic. *Working paper CA0721*.
- GENO, K. A., GILBERT, G. L., SONG, J. Y., SKOVSTED, I. C., KLUGMAN, K. P., JONES, C., KONRADSEN, H. B. & NAHM, M. H. 2015. Pneumococcal Capsules and Their Types: Past, Present, and Future. *Clin Microbiol Rev*, 28, 871-99.
- GLADSTONE, R. A., SIIRA, L., BRYNILDSRUD, O. B., VESTRHEIM, D. F., TURNER, P., CLARKE, S. C., SRIFUENGFUNG, S., FORD, R., LEHMANN, D., EGOROVA, E., VOROPAEVA, E., HARALDSSON, G., KRISTINSSON, K. G., MCGEE, L., BREIMAN, R. F., BENTLEY, S. D., SHEPPARD, C. L., FRY, N. K., CORANDER, J., TOROPAINEN, M., STEENS, A. & CONSORTIUM, G. P. S. 2022. International links between Streptococcus pneumoniae vaccine serotype 4 sequence type (ST) 801 in Northern European shipyard outbreaks of invasive pneumococcal disease. *Vaccine*, 40, 1054-1060.
- GRASSLY, N. C. & FRASER, C. 2008. Mathematical models of infectious disease transmission. *Nat Rev Microbiol*, 6, 477-87.
- GRAY, B. M., CONVERSE, G. M. & DILLON, H. C. 1980. Epidemiologic studies of Streptococcus pneumoniae in infants: acquisition, carriage, and infection during the first 24 months of life. *J Infect Dis*, 142, 923-33.

- GREGORICH, M., STROHMAIER, S., DUNKLER, D. & HEINZE, G. 2021. Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution. *Int J Environ Res Public Health*, 18.
- HARDIE, J. M. & WHILEY, R. A. 1997. Classification and overview of the genera *Streptococcus* and *Enterococcus*. *J Appl Microbiol*, 83, 1S-11S.
- HAUSDORFF, W. P., FEIKIN, D. R. & KLUGMAN, K. P. 2005. Epidemiological differences among pneumococcal serotypes. *The Lancet infectious diseases*, 5, 83-93.
- HENRIQUES-NORMARK, B. & TUOMANEN, E. I. 2013. The pneumococcus: epidemiology, microbiology, and pathogenesis. *Cold Spring Harb Perspect Med*, 3.
- HERBOLD, W., MAUS, R., HAHN, I., DING, N., SRIVASTAVA, M., CHRISTMAN, J. W., MACK, M., REUTERSHAN, J., BRILES, D. E., PATON, J. C., WINTER, C., WELTE, T. & MAUS, U. A. 2010. Importance of CXC chemokine receptor 2 in alveolar neutrophil and exudate macrophage recruitment in response to pneumococcal lung infection. *Infect Immun*, 78, 2620-30.
- HILLERINGMANN, M., KOHLER, S., GÁMEZ, G. & HAMMERSCHMIDT, S. 2015. Chapter 17 - Pneumococcal Pili and Adhesins. In: BROWN, J., HAMMERSCHMIDT, S. & ORIHUELA, C. (eds.) *Streptococcus Pneumoniae*. Amsterdam: Academic Press.
- HÖGBERG, L., GELI, P., RINGBERG, H., MELANDER, E., LIPSITCH, M. & EKDAHL, K. 2007. Age- and serogroup-related differences in observed durations of nasopharyngeal carriage of penicillin-resistant pneumococci. *J Clin Microbiol*, 45, 948-52.
- HYAMS, C., TRZCINSKI, K., CAMBERLEIN, E., WEINBERGER, D. M., CHIMALAPATI, S., NOURSADEGHI, M., LIPSITCH, M. & BROWN, J. S. 2013. *Streptococcus pneumoniae* capsular serotype invasiveness correlates with the degree of factor H binding and opsonization with C3b/iC3b. *Infect Immun*, 81, 354-63.
- JAILLARD, M., LIMA, L., TOURNOUD, M., MAHÉ, P., VAN BELKUM, A., LACROIX, V. & JACOB, L. 2018. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genet*, 14, e1007758.
- JEFFERIES, J. M., JOHNSTON, C. H., KIRKHAM, L. A., COWAN, G. J., ROSS, K. S., SMITH, A., CLARKE, S. C., BRUEGGEMANN, A. B., GEORGE, R. C., PICHON, B., PLUSCHKE, G., PFLUGER, V. & MITCHELL, T. J. 2007. Presence of nonhemolytic pneumolysin in serotypes of *Streptococcus pneumoniae* associated with disease outbreaks. *J Infect Dis*, 196, 936-44.

- JOCHEMS, S. P., WEISER, J. N., MALLEY, R. & FERREIRA, D. M. 2017. The immunological mechanisms that control pneumococcal carriage. *PLoS Pathog*, 13, e1006665.
- KALIZANG'OMA, A., CHAGUZA, C., GORI, A., DAVISON, C., BELEZA, S., ANTONIO, M., BEALL, B., GOLDBLATT, D., KWAMBANA-ADAMS, B., BENTLEY, S. D. & HEYDERMAN, R. S. 2021. Streptococcus pneumoniae serotypes that frequently colonise the human nasopharynx are common recipients of penicillin-binding protein gene fragments from Streptococcus mitis. *Microb Genom*, 7.
- KELLNER, J. D., RICKETSON, L. J., DEMCZUK, W. H. B., MARTIN, I., TYRRELL, G. J., VANDERKOOI, O. G. & MULVEY, M. R. 2021. Whole-Genome Analysis of Streptococcus pneumoniae Serotype 4 Causing Outbreak of Invasive Pneumococcal Disease, Alberta, Canada. *Emerg Infect Dis*, 27, 1867-1875.
- KELLOGG, J. A., BANKERT, D. A., ELDER, C. J., GIBBS, J. L. & SMITH, M. C. 2001. Identification of Streptococcus pneumoniae revisited. *J Clin Microbiol*, 39, 3373-5.
- KILIAN, M. 2010. Streptococcus and Lactobacillus. *Topley and Wilson's microbiology and microbial infections*. Hodder Arnold, London, United Kingdom.
- KIRKPATRICK, S. I. & TARASUK, V. 2011. Housing circumstances are associated with household food access among low-income urban families. *Journal of urban health*, 88, 284-296.
- KOPPE, U., SUTTORP, N. & OPITZ, B. 2012. Recognition of Streptococcus pneumoniae by the innate immune system. *Cell Microbiol*, 14, 460-6.
- KRZYŚCIAK, W., PLUSKWA, K., JURCZAK, A. & KOŚCIELNIAK, D. 2013. The pathogenicity of the Streptococcus genus. *European Journal of Clinical Microbiology & Infectious Diseases*, 32, 1361-1376.
- KURTZ, S., PHILLIPPY, A., DELCHER, A. L., SMOOT, M., SHUMWAY, M., ANTONESCU, C. & SALZBERG, S. L. 2004. Versatile and open software for comparing large genomes. *Genome Biol*, 5, R12.
- LANGLEY, G., SCHAFFNER, W., FARLEY, M. M., LYNFIELD, R., BENNETT, N. M., REINGOLD, A., THOMAS, A., HARRISON, L. H., NICHOLS, M., PETIT, S., MILLER, L., MOORE, M. R., SCHRAG, S. J., LESSA, F. C., SKOFF, T. H., MACNEIL, J. R., BRIERE, E. C., WESTON, E. J. & VAN BENEDEN, C. 2015. Twenty Years of Active Bacterial Core Surveillance. *Emerg Infect Dis*, 21, 1520-8.
- LANNES-COSTA, P. S., DE OLIVEIRA, J. S. S., DA SILVA SANTOS, G. & NAGAO, P. E. 2021. A current review of pathogenicity determinants of Streptococcus sp. *J Appl Microbiol*, 131, 1600-1620.

- LEES, J. A., CROUCHER, N. J., GOLDBLATT, D., NOSTEN, F., PARKHILL, J., TURNER, C., TURNER, P. & BENTLEY, S. D. 2017. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. *Elife*, 6.
- LEES, J. A., FERWERDA, B., KREMER, P. H. C., WHEELER, N. E., SERÓN, M. V., CROUCHER, N. J., GLADSTONE, R. A., BOOTSMA, H. J., ROTS, N. Y., WIJMEGA-MONSUUR, A. J., SANDERS, E. A. M., TRZCIŃSKI, K., WYLLIE, A. L., ZWINDERMAN, A. H., VAN DEN BERG, L. H., VAN RHEENEN, W., VELDINK, J. H., HARBOE, Z. B., LUNDBO, L. F., DE GROOT, L. C. P. G., VAN SCHOOR, N. M., VAN DER VELDE, N., ÄNGQUIST, L. H., SØRENSEN, T. I. A., NOHR, E. A., MENTZER, A. J., MILLS, T. C., KNIGHT, J. C., DU PLESSIS, M., NZENZE, S., WEISER, J. N., PARKHILL, J., MADHI, S., BENFIELD, T., VON GOTTBURG, A., VAN DER ENDE, A., BROUWER, M. C., BARRETT, J. C., BENTLEY, S. D. & VAN DE BEEK, D. 2019. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nat Commun*, 10, 2176.
- LEES, J. A., GALARDINI, M., BENTLEY, S. D., WEISER, J. N. & CORANDER, J. 2018. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, 34, 4310-4312.
- LEHTINEN, S., BLANQUART, F., CROUCHER, N. J., TURNER, P., LIPSITCH, M. & FRASER, C. 2017. Evolution of antibiotic resistance is linked to any genetic mechanism affecting bacterial duration of carriage. *Proc Natl Acad Sci U S A*, 114, 1075-1080.
- LEMAY, J. A., RICKETSON, L. J., ZWICKER, L. & KELLNER, J. D. 2019. Homelessness in Adults with Invasive Pneumococcal Disease (IPD) in Calgary, Canada. *Open Forum Infect Dis*, 6.
- LEWNARD, J. A. & HANAGE, W. P. 2019. Making sense of differences in pneumococcal serotype replacement. *Lancet Infect Dis*, 19, e213-e220.
- LI, Y., WEINBERGER, D. M., THOMPSON, C. M., TRZCIŃSKI, K. & LIPSITCH, M. 2013. Surface charge of *Streptococcus pneumoniae* predicts serotype distribution. *Infect Immun*, 81, 4519-24.
- LØCHEN, A., CROUCHER, N. J. & ANDERSON, R. M. 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.
- LØCHEN, A., TRUSCOTT, J. E. & CROUCHER, N. J. 2022. Analysing pneumococcal invasiveness using Bayesian models of pathogen progression rates. *PLoS Comput Biol*, 18, e1009389.

- LY, T. D. A., PERIERES, L., HOANG, V. T., DAO, T. L. & GAUTRET, P. 2021. Pneumococcal infections and homelessness. *J Prev Med Hyg*, 62, E950-E957.
- MAESTRO, B. & SANZ, J. M. 2016. Choline Binding Proteins from *Streptococcus pneumoniae*: A Dual Role as Enzybiotics and Targets for the Design of New Antimicrobials. *Antibiotics (Basel)*, 5.
- 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*.
- MARRIOTT, H. M., MITCHELL, T. J. & DOCKRELL, D. H. 2008. Pneumolysin: a double-edged sword during the host-pathogen interaction. *Curr Mol Med*, 8, 497-509.
- MELEGARO, A., GAY, N. J. & MEDLEY, G. F. 2004. Estimating the transmission parameters of pneumococcal carriage in households. *Epidemiol Infect*, 132, 433-41.
- MELIN, M., TRZCIQSKI, 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. *Infection and Immunity*, 78, 5262 - 5270.
- METCALF, B. J., CHOCHUA, S., GERTZ, R. E., LI, Z., WALKER, H., TRAN, T., HAWKINS, P. A., GLENNEN, A., LYNFIELD, R., LI, Y., MCGEE, L., BEALL, B. & TEAM, A. B. C. S. 2016a. Using whole genome sequencing to identify resistance determinants and predict antimicrobial resistance phenotypes for year 2015 invasive pneumococcal disease isolates recovered in the United States. *Clin Microbiol Infect*, 22, 1002.e1-1002.e8.
- METCALF, B. J., CHOCHUA, S., WALKER, H., TRAN, T., LI, Z., VARGHESE, J., SNIPPES VAGNONE, P. M., 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.
- METCALF, B. J., GERTZ, R. E., GLADSTONE, R. A., WALKER, H., SHERWOOD, L. K., JACKSON, D., LI, Z., LAW, C., HAWKINS, P. A., CHOCHUA, S., SHETH, M., RAYAMAJHI, N., BENTLEY, S. D., KIM, L., WHITNEY, C. G., MCGEE, L., BEALL, B. & TEAM, A. B. C. S. 2016b. Strain features and distributions in pneumococci from children with invasive disease before and after 13-valent conjugate vaccine implementation in the USA. *Clin Microbiol Infect*, 22, 60.e9-60.e29.

- METCALF, B. J., WALDETOFT, K. W., BEALL, B. W. & BROWN, S. P. 2023. Variation in pneumococcal invasiveness metrics is driven by serotype carriage duration and initial risk of disease. *Epidemics*, 45, 100731.
- MILLS, C. E., ROBINS, J. M. & LIPSITCH, M. 2004. Transmissibility of 1918 pandemic influenza. *Nature*, 432, 904-6.
- MITCHELL, A. M. & MITCHELL, T. J. 2010. Streptococcus pneumoniae: virulence factors and variation. *Clin Microbiol Infect*, 16, 411-8.
- MOFFA, M., CRONK, R., FEJFAR, D., DANCAUSSE, S., PADILLA, L. A. & BARTRAM, J. 2019. A systematic scoping review of environmental health conditions and hygiene behaviors in homeless shelters. *Int J Hyg Environ Health*, 222, 335-346.
- MORELAND, J. G. & BAILEY, G. 2006. Neutrophil transendothelial migration in vitro to Streptococcus pneumoniae is pneumolysin dependent. *Am J Physiol Lung Cell Mol Physiol*, 290, L833-40.
- MOSITES, E., ZULZ, T., BRUDEN, D., NOLEN, L., FRICK, A., CASTRODALE, L., MCLAUGHLIN, J., VAN BENEDEN, C., HENNESSY, T. W. & BRUCE, M. G. 2019. Risk for Invasive Streptococcal Infections among Adults Experiencing Homelessness, Anchorage, Alaska, USA, 2002-2015. *Emerg Infect Dis*, 25, 1911-8.
- MOSTOWY, R. J., CROUCHER, N. J., DE MAIO, N., CHEWAPREECHA, C., SALTER, S. J., TURNER, P., AANENSEN, D. M., BENTLEY, S. D., DIDELOT, X. & FRASER, C. 2017. Pneumococcal Capsule Synthesis Locus cps as Evolutionary Hotspot with Potential to Generate Novel Serotypes by Recombination. *Mol Biol Evol*, 34, 2537-2554.
- NISHIURA, H. A. C. G. 2008. Rurality and pandemic influenza: Geographic heterogeneity in the risks of infection and death in Kanagawa, Japan (1918-1919). *The New Zealand medical journal*, 121, 18-27.
- NOBBS, A. H., LAMONT, R. J. & JENKINSON, H. F. 2009. Streptococcus adherence and colonization. *Microbiology and molecular biology reviews*, 73, 407-450.
- O'BRIEN, K. L., WOLFSON, L. J., WATT, J. P., HENKLE, E., DELORIA-KNOLL, M., MCCALL, N., LEE, E., MULHOLLAND, K., LEVINE, O. S., 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.
- OBERT, C., SUBLETT, J., KAUSHAL, D., HINOJOSA, E., BARTON, T., TUOMANEN, E. I. & ORIHUELA, C. J. 2006. Identification of a Candidate Streptococcus pneumoniae core genome and regions of diversity correlated with invasive pneumococcal disease. *Infect Immun*, 74, 4766-77.

- OBOLSKI, U., GORI, A., LOURENÇO, J., THOMPSON, C., THOMPSON, R., FRENCH, N., HEYDERMAN, R. & GUPTA, S. 2018. Identifying *Streptococcus pneumoniae* genes associated with invasive disease using pangenome-based whole genome sequence typing. *bioRxiv*, 314666.
- PAGE, A. J., CUMMINS, C. A., HUNT, M., WONG, V. K., REUTER, S., HOLDEN, M. T., FOOKES, M., FALUSH, D., KEANE, J. A. & PARKHILL, J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31, 3691-3.
- PAUDEL, S., BARAL, P., GHIMIRE, L., BERGERON, S., JIN, L., DECORTE, J. A., LE, J. T., CAI, S. & JEYASEELAN, S. 2019. CXCL1 regulates neutrophil homeostasis in pneumonia-derived sepsis caused by *Streptococcus pneumoniae* serotype 3. *Blood*, 133, 1335-1345.
- PEARL, J. & MACKENZIE, D. 2018. The book of why : the new science of cause and effect. New York: Basic Books,.
- PEYRANI, P., MANDELL, L., TORRES, A. & TILLOTSON, G. S. 2019. The burden of community-acquired bacterial pneumonia in the era of antibiotic resistance. *Expert Rev Respir Med*, 13, 139-152.
- PILISHVILI, T., LEXAU, C., FARLEY, M. M., HADLER, J., HARRISON, L. H., BENNETT, N. M., REINGOLD, A., THOMAS, A., SCHAFFNER, W., CRAIG, A. S., SMITH, P. J., BEALL, B. W., WHITNEY, C. G., MOORE, M. R. & NETWORK, A. B. C. S. E. I. P. 2010. Sustained reductions in invasive pneumococcal disease in the era of conjugate vaccine. *J Infect Dis*, 201, 32-41.
- R CORE TEAM 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
- RELLER, L. B., WEINSTEIN, M. P., WERNO, A. M. & MURDOCH, D. R. 2008. Laboratory diagnosis of invasive pneumococcal disease. *Clinical infectious diseases*, 46, 926-932.
- ROBINSON, D. A., TURNER, J. S., FACKLAM, R. R., PARKINSON, A. J., BREIMAN, R. F., GRATTEN, M., STEINHOFF, M. C., HOLLINGSHEAD, S. K., BRILES, D. E. & CRAIN, M. J. 1999. Molecular characterization of a globally distributed lineage of serotype 12F *Streptococcus pneumoniae* causing invasive disease. *J Infect Dis*, 179, 414-22.
- ROMNEY, M. G., HULL, M. W., GUSTAFSON, R., SANDHU, J., CHAMPAGNE, S., WONG, T., NEMATALLAH, A., FORSTING, S. & DALY, P. 2008. Large community outbreak of *Streptococcus pneumoniae* serotype 5 invasive infection in an impoverished, urban population. *Clin Infect Dis*, 47, 768-74.
- SANTOS, G. & VERA, J. 2020. The Role of Age, Neutrophil Infiltration and Antibiotics Timing in the Severity of. *Int J Mol Sci*, 21.

- SCHISTERMAN, E. F., PERKINS, N. J., MUMFORD, S. L., AHRENS, K. A. & MITCHELL, E. M. 2017. Collinearity and Causal Diagrams: A Lesson on the Importance of Model Specification. *Epidemiology*, 28, 47-53.
- SHAHSIKALA, N. 2008. *Occurrence of Sero Groups of Beta Haemolytic Streptococci in Pharynx of Asymptomatic School Children*. Rajiv Gandhi University of Health Sciences (India).
- SHEN, P., LEES, J. A., BEE, G. C. W., BROWN, S. P. & WEISER, J. N. 2019. Pneumococcal quorum sensing drives an asymmetric owner-intruder competitive strategy during carriage via the competence regulon. *Nat Microbiol*, 4, 198-208.
- SLEEMAN, K. L., GRIFFITHS, D., SHACKLEY, F., DIGGLE, L., GUPTA, S., MAIDEN, M. C., MOXON, E. R., CROOK, D. W. & PETO, T. E. 2006. Capsular serotype-specific attack rates and duration of carriage of *Streptococcus pneumoniae* in a population of children. *J Infect Dis*, 194, 682-8.
- SONG, J. Y., NAHM, M. H. & MOSELEY, M. A. 2013. Clinical implications of pneumococcal serotypes: invasive disease potential, clinical presentations, and antibiotic resistance. *J Korean Med Sci*, 28, 4-15.
- STONE, M. 2010. *Shelter poverty: New ideas on housing affordability*, Temple University Press.
- SUBRAMANIAN, K., DANIEL, N., HESHAM, M., LAURA, S., KHANDAKER, S., GIORGIA, M., EMMA, D., ALUN, K., MARIE, Y., ADNANE, A., JOHAN, N., PER-ÅKE, N., LAURA, P., ARAS, K. & BIRGITTA, H.-N. 2019. Pneumolysin binds to the mannose receptor C type 1 (MRC-1) leading to anti-inflammatory responses and enhanced pneumococcal survival. *Nature Microbiology*.
- SY, K. T. L., WHITE, L. F. & NICHOLS, B. E. 2021. Population density and basic reproductive number of COVID-19 across United States counties. *PLoS One*, 16, e0249271.
- TOMCZYK, S., ARRIOLA, C. S., BEALL, B., BENITEZ, A., BENOIT, S. R., BERMAN, L., BRESEE, J., DA GLORIA CARVALHO, M., COHN, A., CROSS, K., DIAZ, M. H., FRANCOIS WATKINS, L. K., GIERKE, R., HAGAN, J. E., HARRIS, A. M., JAIN, S., KIM, L., KOBAYASHI, M., LINDSTROM, S., MCGEE, L., MCMORROW, M., METCALF, B. L., MOORE, M. R., MOURA, I., NIX, W. A., NYANGOMA, E., OBERSTE, M. S., OLSEN, S. J., PIMENTA, F., SOCIAS, C., THURMAN, K., WALLER, J., WATERMAN, S. H., WESTERCAMP, M., WHARTON, M., WHITNEY, C. G., WINCHELL, J. M., WOLFF, B. & KIM, C. 2016. Multistate Outbreak of Respiratory Infections Among Unaccompanied Children, June 2014-July 2014. *Clin Infect Dis*, 63, 48-56.
- TÖNNIES, T., KAHL, S. & KUSS, O. 2022. Collider Bias in Observational Studies. *Dtsch Arztebl Int*, 119, 107-122.

- VAN SELM, S., VAN CANN, L. M., KOLKMAN, M. A., VAN DER ZEIJST, B. A. & VAN PUTTEN, J. P. 2003. Genetic basis for the structural difference between *Streptococcus pneumoniae* serotype 15B and 15C capsular polysaccharides. *Infect Immun*, 71, 6192-8.
- VANDERKOOI, O. G., CHURCH, D. L., MACDONALD, J., ZUCOL, F. & KELLNER, J. D. 2011. Community-based outbreaks in vulnerable populations of invasive infections caused by *Streptococcus pneumoniae* serotypes 5 and 8 in Calgary, Canada. *PLoS One*, 6, e28547.
- VILHENA, C., DU, S., BATTISTA, M., WESTERMANN, M., KOHLER, T., HAMMERSCHMIDT, S. & ZIPFEL, P. F. 2023. The choline-binding proteins PspA, PspC, and LytA of *Streptococcus pneumoniae* and their interaction with human endothelial and red blood cells. *Infection and Immunity*, 91, e00154-23.
- VÖGELE, M., BHASKARA, R. M., MULVIHILL, E., VAN PEE, K., YILDIZ, Ö., KÜHLBRANDT, W., MÜLLER, D. J. & HUMMER, G. 2019. Membrane perforation by the pore-forming toxin pneumolysin. *Proc Natl Acad Sci U S A*, 116, 13352-13357.
- WALKER, M. J., BARNETT, T. C., MCARTHUR, J. D., COLE, J. N., GILLEN, C. M., HENNINGHAM, A., SRIPRAKASH, K. S., SANDERSON-SMITH, M. L. & NIZET, V. 2014. Disease manifestations and pathogenic mechanisms of Group A *Streptococcus*. *Clin Microbiol Rev*, 27, 264-301.
- WALTER, N. D., TAYLOR, T. H., DOWELL, S. F., MATHIS, S., MOORE, M. R. & TEAM, A. B. C. S. S. 2009. Holiday spikes in pneumococcal disease among older adults. *N Engl J Med*, 361, 2584-5.
- WEIGHT, C. M., VENTURINI, C., POJAR, S., JOCHEMS, S. P., REINÉ, J., NIKOLAOU, E., SOLÓRZANO, C., NOURSADEGHI, M., BROWN, J. S., FERREIRA, D. M. & HEYDERMAN, R. S. 2019. Microinvasion by *Streptococcus pneumoniae* induces epithelial innate immunity during colonisation at the human mucosal surface. *Nat Commun*, 10, 3060.
- WEINBERGER, D. M., TRZCIŃSKI, K., LU, Y. J., BOGAERT, D., BRANDES, A., GALAGAN, J., ANDERSON, P. W., MALLEY, R. & LIPSITCH, M. 2009. Pneumococcal capsular polysaccharide structure predicts serotype prevalence. *PLoS Pathog*, 5, e1000476.
- WEISER, J. N., FERREIRA, D. M. & PATON, J. C. 2018. *Streptococcus pneumoniae*: transmission, colonization and invasion. *Nat Rev Microbiol*, 16, 355-367.
- WOLTER, N., TEMPIA, S., COHEN, C., MADHI, S. A., VENTER, M., MOYES, J., WALAZA, S., MALOPE-KGOKONG, B., GROOME, M., DU PLESSIS, M., MAGOMANI, V., PRETORIUS, M., HELLFERSCEE, O., DAWOOD, H., KAHN, K., VARIAVA, E., KLUGMAN, K. P. & VON GOTTFERG, A. 2014.

High nasopharyngeal pneumococcal density, increased by viral coinfection, is associated with invasive pneumococcal pneumonia. *J Infect Dis*, 210, 1649-57.

YILDIRIM, I., HANAGE, W. P., LIPSITCH, M., SHEA, K. M., STEVENSON, A., FINKELSTEIN, J., HUANG, S. S., LEE, G. M., KLEINMAN, K. & PELTON, S. I. 2010. Serotype specific invasive capacity and persistent reduction in invasive pneumococcal disease. *Vaccine*, 29, 283-8.

YOTHER, J. 2011. Capsules of *Streptococcus pneumoniae* and other bacteria: paradigms for polysaccharide biosynthesis and regulation. *Annu Rev Microbiol*, 65, 563-81.

ZAFAR, M. A., WANG, Y., HAMAGUCHI, S. & WEISER, J. N. 2017. Host-to-Host Transmission of *Streptococcus pneumoniae* Is Driven by Its Inflammatory Toxin, Pneumolysin. *Cell Host Microbe*, 21, 73-83.

ZIVICH, P. N., GRABENSTEIN, J. D., BECKER-DREPS, S. I. & WEBER, D. J. 2018. *Streptococcus pneumoniae* outbreaks and implications for transmission and control: a systematic review. *Pneumonia (Nathan)*, 10, 11.

ZULZ, T., WENGER, J. D., RUDOLPH, K., ROBINSON, D. A., RAKOV, A. V., BRUDEN, D., SINGLETON, R. J., BRUCE, M. G. & HENNESSY, T. W. 2013. Molecular characterization of *Streptococcus pneumoniae* serotype 12F isolates associated with rural community outbreaks in Alaska. *J Clin Microbiol*, 51, 1402-7.