Evidence of cryptic incidence in childhood diseases

Christian E. Gunning[1,4*] Matthew J. Ferrari[2], Erik Erhardt[3], Helen J. Wearing[1,3]

1 Department of Biology, University of New Mexico, Albuquerque, New Mexico, USA

2 Center for Infectious Disease Dynamics, Pennsylvania State University, University Park, Pennsylvania, USA

3 Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico, USA

4 Currently at Department of Entomology, North Carolina State University, Raleigh, North Carolina, USA

∗ **E-mail: ceg.2015@x14n.org**

Abstract

Persistence and extinction are key processes in infectious disease dynamics that, due to incomplete reporting, cannot be directly observed. For fully-immunizing diseases, reporting probabilities can be estimated from demographic records and case reports. Yet reporting probabilities are not sufficient to unambiguously reconstruct disease incidence from case reports. Here, we focus on disease presence (i.e., non-zero incidence), which provides an upper bound on disease extinction. We examine measles and pertussis in pre-vaccine era U.S. cities, and describe a conserved scaling relationship between population size, reporting rate, and observed presence (i.e., non-zero case reports). Using this relationship, we estimate disease presence given perfect reporting, and define cryptic presence as the difference between observed and estimated presence. We estimate that, in early 20^{th} century U.S. cities, pertussis presence was higher than measles presence across a range of population sizes, and that cryptic presence was common in small cities with imperfect reporting. Our results suggest that unobserved, cryptic incidence deserves careful attention, particularly in "colonizer" diseases with longer infectious periods and lower transmission rates. Indeed, cryptic presence could paradoxically increase in response to control measures such as vaccination campaigns.

Keywords

disease persistence, stochastic extinction, incomplete observation, critical community size, measles, pertussis

Introduction

Epidemic Dynamics of Childhood Diseases

Measles and pertussis (whooping cough) are acutely infectious and fully immunizing diseases caused by obligate human pathogens: the measles virus and *Bordetella pertussis*, respectively. These so-called "childhood diseases" have high reproductive ratios ($R_0 \approx 20$) and, in the pre-vaccine era, infected an overwhelming majority of humans at a young age [1]. Recurrent epidemics are a common feature of these diseases, driven by periodic forcing of disease transmission via changes in host density, such as school terms [2–4] or economic migration [5, 6].

The parasite life cycles of both measles and pertussis are fast compared to human host demographics: the combined latent and infectious period is approximately 15 days for measles and 30 days for pertussis [1]. At high incidence, susceptible hosts are rapidly depleted, leading to subsequent inter-epidemic troughs of low incidence, where stochastic extinction can occur. When infection is low or absent from a population, susceptible replenishment proceeds via the host demographic processes of birth and migration. These forces combine to yield characteristic yearly and multi-annual epidemic cycles in a range of diseases and human populations [7–14].

While broadly similar, the life histories of measles and pertussis differ significantly in pace. Measles has a shorter life cycle, is more "invasive", and experiences more pronounced epidemics, while pertussis is the superior "colonizer". The slower life history of pertussis is expected to dampen the effects of isolation relative to measles, and is predicted to enhance dynamical stochasticity [15, 16].

Reporting probabilities also vary widely between diseases, as well as between locations [3, 17–19]. Measles has unambiguous symptoms that remain approximately constant with age. Pertussis, on the other hand, exhibits age-dependent severity, and shares symptoms with many other common respiratory diseases [20, 21]. Consequently, pertussis reporting is generally less complete and more variable than measles reporting [17, 19]. Such observational differences complicate meaningful comparisons between diseases, particularly in the presence of dynamical uncertainty.

Determinants of Persistence

Persistence and stochastic extinction are key processes that affect pathogen ecology and evolution, along with disease control efforts. As an ecological outcome, disease persistence arises from a complex interplay between local, within-population processes and metapopulation-level interactions among populations. Disentangling the impact of local and metapopulation processes on disease dynamics has proved challenging. At the local level, stochasticity in host and pathogen demographic processes commonly results in local extinction, particularly in small populations [22–25], and for pathogens with short infectious periods [25, 26]. Indeed, previous work has shown that local disease persistence scales approximately log-linearly with population size [25, 22, 27–31]. Likewise, theory predicts that, when all else is equal, longer latent and infectious periods and higher birth rates should increase local persistence [27, 25].

At the metapopulation level, host migration allows for pathogen reintroduction, which can restart (or "rescue") local chains of infection from extinction [32, 31, 26, 33]. High levels of connectivity among populations can increase rescue effects and metapopulation persistence [32], while low connectivity can favor a boom-bust cycle. In the latter case, "rescues" are uncommon, and prolonged periods of local extinction allow susceptible individuals to accumulate far above equilibrium. Eventual pathogen re-introduction causes explosive epidemics that, in turn, reduce susceptible individuals far below equilibrium, thus favoring stochastic extinction.

Here we focus on measles and pertussis in pre-vaccine era U.S. cities, where explosive epidemics and prolonged periods of low incidence and / or stochastic extinction are common. We examine a record of more than two decades of continuous weekly disease monitoring (1924-1945, Table 1) that includes the majority of U.S. urban areas in this era. The early 20^{th} century U.S. provides an attractive model system: high-quality demographic records are available, and a diverse range of population sizes, ethnic compositions, and levels of geographic isolation are represented here. In addition, lack of vaccination eliminates uncertainty associated with vaccine uptake and efficacy.

Estimating Disease Presence

Due to imperfect reporting, the dynamical processes of persistence and stochastic extinction cannot be directly observed. Previous work has estimated disease persistence from case reports, either in distinct human populations (e.g. cities, Conlan et al. [31]) or in metapopulations (e.g. countries, Metcalf et al. [26]). Lacking, however, are quantitative assessments of the impact of observational uncertainty.

Species *presence* is a related quantity that has received considerable attention from community and conservation ecologists seeking reliable measures of species composition or richness. Here, sampling effort and species abundance have long been recognized to affect species detection probabilities [34, 35]. In assemblages of species, sampling effort can be accounted for via accumulation or rarefaction curves that quantify presence via asymptotic richness [35–37]. Related work has explored the interdependence between detection probability and spatiotemporal resolution [38], and has quantified the expected additional sampling required to achieve asymptotic detection [39].

Here we address a related problem: the reliable detection of a *single* species' presence. In this case, reporting probability provides a proxy for sampling effort, while disease incidence is analogous to species abundance. We also explore the impact of temporal "grain size" [38] by aggregating case reports over a range of successively longer time periods. We suggest that the long-term, per-population probability of disease presence yields a lower bound on disease incidence, and provides an upper bound on time spent in an extinct state.

Outline

In this work, we use empirical observations of disease case reports (C) to estimate disease presence (P). We compare measures of presence between two childhood diseases within a single metapopulation. Marginalizing over time, and using cities as the basic units of population, we show that population size (N) and reporting rate (r) predict observed presence (non-zero case reports, P_o). We then use this relationship to estimate disease presence at full reporting (i.e. non-zero incidence, P), and demonstrate increased presence of pertussis relative to measles across a range of population sizes. We also examine the impact of temporal aggregation of disease case reports on estimates of disease presence.

We next examine cryptic presence (P_c), which we define as the difference between observed and estimated presence: the estimated per-time probability of unobserved presence. We show that cryptic presence scales with both population size and reporting probability, and is particularly common in small populations with low reporting probability. Consequently, patterns of cryptic presence in this system differ markedly between diseases.

Methods

Overview

In the following, we indicate estimated quantities with a *tilde* superscript: $\widetilde{X}$. Thus P_o is observed presence, while $\widetilde{P}$ is an estimate of presence. All subsequent analyses were conducted for each disease separately. Definitions: case reports, C; incidence, I; reporting probability, r; population size (1930), N.

Disease incidence varies greatly over time, both within and across years; here we marginalize over time, and focus on long-term differences between populations and diseases. We define observed presence as the (marginal) per-time probability of non-zero case reports ($P_o = Pr(C > 0)$), and estimate presence as the (marginal) per-time probability of non-zero incidence ($\widetilde{P} \approx Pr(I > 0)$). Unless otherwise noted, we use the epidemiological reporting week as the basic unit of time.

For each city, we estimate reporting probabilities ($\widetilde{r}$) from case reports and demographic records. We define the *monitored* population size as the full population size scaled by the reporting rate: $\widetilde{N_m} = N \times \widetilde{r}$).

We employ a binomial generalized linear model (B-GLM) to model the response of observed presence to monitored population. We use this model to estimate presence ($\widetilde{P}$) from full population size (N), and examine how the temporal aggregation of case reports affects model fits and predictions.

We define cryptic presence as the difference between estimated and observed presence ($\widetilde{P_c} = \widetilde{P} - P_o$). We examine the dependence of cryptic presence on both population size and reporting rate, and compare metapopulation patterns of estimated and cryptic presence between disease.

Incomplete observation

To estimate reporting probabilities ($\widetilde{r}$), we assume that each population's proportion of susceptibles is in quasi-equilibrium over the period of record, and that the lifetime probability of infection is close to unity [3, 17]. As discussed in Gunning et al. [19], no strong evidence of time-variable reporting is apparent in this system over the period of record, and reporting probability is thus assumed to be constant over time.

In summary, we assume that case reports are generated via binomial sampling of incidence: $C \sim \text{Bin}(I, r)$. Thus, each city's estimated reporting probability is the total surviving births divided by total case reports, summed over the period of observation: $\widetilde{r} = \sum C / \sum \text{births})$. As described in Gunning et al. [19], total surviving births are estimated from yearly per capita state birth rates and infant mortality rates, along with yearly city populations. We estimate approximate confidence intervals by bootstrapping (yearly) birth and infant mortality rates.

Assuming that case reports are generated via a binomial sampling process, estimated incidence is simply case reports divided by the reporting probability: $\widetilde{I} = C/\widetilde{r}$. Yet this correction fails for $C = 0$, where the maximum likelihood estimator (MLE) of incidence is zero. In the case of low reporting probability and low incidence, a non-trivial proportion of observed zeros ($C = 0$) result from cryptic presence (P_c).

Assuming, for example, that $r = 0.1$ and $I = 10$, then $Pr(C = 0 | r = 0.1, I = 10) = (1 - 0.1)^{10} \sim 0.35$. In this case, some 35% of case reports will be zero, despite absolute incidence that is far from zero. In short, time periods with zero case reports result from a "mixed process" of disease absence ($I = 0$), together with unobserved, cryptic presence: $(I > 0 \cap C = 0) \Rightarrow P_c > 0$. It is this unobserved presence that we seek to quantify.

Estimated and Cryptic Presence

We first exclude cities where a disease was always present ($Pr(C > 0) = 1$). We define full population (N) as the 1930 census population, and the monitored population (N_m) as the full population scaled by the reporting probability: $\widetilde{N_m} = N \times \widetilde{r}$.

The response of observed presence ($P_o = Pr(C > 0)$) to log monitored population was modeled with a binomial GLM (B-GLM), with a complementary log-log (cloglog) link function ($f(\dot{})$): $f(P_o) \sim \log \widetilde{N_m}$. Populations were weighted by the number of non-excluded time periods (i.e., number of "successes", where $C > 0 \Rightarrow$ success). Unlike the more commonly employed logit link, the cloglog link hypothesizes an asymmetric response to the predictor. That is, at large population sizes, cities approach complete presence ($P_o = 1$) more rapidly than a logit link predicts. This accords with biological intuition, where mechanistically distinct processes dominate near complete presence versus near complete absence.

We use the resulting B-GLMs to extrapolate an estimate of presence ($\widetilde{P} \approx P$), from full population size: $f(\widetilde{P}) \sim \log N$. Conceptually, N simply equals the monitored population under complete reporting, motivating our choice.

Estimated presence is the sum of observed presence and cryptic (unobserved) presence (P_c): $\widetilde{P} = P_o + P_c$. As such, our best estimate of cryptic presence is the difference between estimated and observed presence: $\widetilde{P_c} = \widetilde{P} - P_o$. Cryptic presence, in turn, provides evidence of cryptic incidence.

Model Exploration

To explore the effect of time period, observed presence was recomputed from temporally aggregated case reports, and a new set of B-GLMs were fit. Case reports were summed over time periods of varying lengths (Δ_t), ranging from 2 to 16 weeks. To compute time period sums, NA weeks were omitted, and periods containing only NA weeks are excluded (excluded periods were common for pertussis). Presence was computed as the proportion of non-zero periods. Here, both $\log \widetilde{N_m}$ and Δ_t were predictors.

Note that the cloglog link function $f(x) = \log(-\log(1 - x))$ provides a useful biological interpretation of the postulated model formulation: $f(P) \sim \log(N)$. For each population, assume a constant rate of infection (λ) and total susceptible population (S). Then the probability of no new infections in a given time period Δ_t is $Pr(I = 0) = 1 - P \approx \exp^{(-\lambda S \Delta_t)}$. For a given population size N, the susceptible proportion is then S/N, and $P = Pr(I > 0) = 1 - \exp^{(-\lambda N(S/N)\Delta_t)}$. The cloglog link then yields: $f(P) = \log(\lambda) + \log(S/N) + \log(N) + \log(\Delta_t)$. Thus, we expect the transformed response ($f(P)$) to change linearly in both $\log(N)$ and $\log(\Delta_t)$. In truth, the populations studied here are not at equilibrium, such that λ and S/N instead oscillate around quasi-equilibria. Nonetheless, the above analysis hypothesizes a functional relationship between P, N, and Δ_t, which we explore further below.

Estimating uncertainty

Estimated presence and cryptic presence are influenced by uncertainty arising from reporting probability estimates and linear model predictions. We used a two-step process of bootstrap sampling to estimate the combined impact of reporting probability and B-GLM prediction uncertainty.

First, reporting probability draws were estimated via non-parametric sampling. For each bootstrap draw, yearly state birth rates and national infant mortality rates were resampled, and total births thus summed. The resulting draws of $\widetilde{r}$ were then used to compute $\widetilde{N_m}$ from N, and a new B-GLM was fit for each model.

To incorporate per-city variance of r into model predictions of P, a population size was back-estimated from $\widetilde{N_m}$ using two independent draws of reporting probability: $\widetilde{N} = N \times (\widetilde{r}_i / \widetilde{r}_j)$. The resulting $\widetilde{N}$ was used to estimate a new $\widetilde{P}$, as above. In this way, bootstrap samples of $\widetilde{P}$ are drawn that incorporate a "worst-case" uncertainty in r. Finally, $\widetilde{P_c}$ was computed from P_o and the re-sampled $\widetilde{P}$, and the resulting bootstrap samples were used to construct 95% prediction intervals for $\widetilde{P}$ and $\widetilde{P_c}$.

Disease	Measles	Pertussis
Total Weeks	1148	1148
Total Cities with $P_o < 1$	82	79
Start Date	1924-01-05	1924-01-05
End Date	1945-12-29	1945-12-29
Reporting Probability (r)	0.31 [0.18-0.41, 0.56]	0.10 [0.04-0.15, 0.77]
Observed Presence (P_o)	0.64 [0.48-0.79, 0.32]	0.68 [0.49-0.92, 0.39]
Estimated Presence ($\widetilde{P}$)	0.84 [0.75-0.95, 0.17]	0.98 [0.99-1.00, 0.04]
Cryptic Presence ($\widetilde{P_c}$)	0.20 [0.10-0.28, 0.64]	0.30 [0.08-0.48, 0.81]

Table 1: Overview of sampled cities and case report weeks, followed by summary statistics of main results (Mean [Q1-Q3, CV]), including observed (P_o), estimated ($\widetilde{P}$), and cryptic ($\widetilde{P_c}$) presence.

Results

Summary statistics and observation counts are shown in Table 1. Overall, the average reporting probability of pertussis is much lower than for measles, with a higher coefficient of variation amongst cities. Figure 1 shows P_o, $\widetilde{P}$, and $\widetilde{P_c}$ (rows) versus N and $\widetilde{N_m}$ (columns). Figure 1 also provides a visual illustration of $\widetilde{P} - P_o = \widetilde{P_c}$, i.e., panels $E = C - A$, and $F = D - B$.

Regardless of disease, we expect less frequent presence in smaller populations [22–25]. We do, in fact, find population size (N) to be a reasonable predictor of observed presence (P_o), though no differences between diseases are evident (Figure 1A). When reporting is considered, monitored population size (N_m) yields an excellent predictor of P_o (Figure 1B): pseudo-$R^2 = 0.908$ (measles) and 0.962 (pertussis). It is these models that we use for subsequent estimates of presence ($\widetilde{P}$, Figure 1C and D).

Theory predicts that pertussis, with a longer infectious period and lower transmission rate, should exhibit less frequent stochastic extinction than measles for a given population size [27, 25], a pattern obscured by pertussis' low and variable reporting. Correcting for incomplete reporting, we estimate that pertussis presence ($\widetilde{P}$) is indeed higher across a wide range of population sizes (Figure 1C).

As expected, cryptic presence of both diseases is rare in large populations (Figure 1E), where absolute incidence is high. In the remaining populations, however, the two diseases differ. For measles, cryptic presence is common across a wide range of population sizes, though true absence (i.e. via stochastic extinction) appears to dominate in smaller cities (Figure 1C). For pertussis, cryptic presence is most common in smaller cities, where frequent failures to detect disease arise from a combination of low reporting and low absolute incidence (Table 1, Figure 1E).

We expect cryptic presence to be a function of both reporting probability and the underlying distribution of absolute incidence. For both diseases, we do indeed observe increased cryptic presence for lower reporting probabilities (Figure 2). We find marked differences between diseases: for a given reporting probability, measles generally experiences higher cryptic presence, possibly due to prolongued periods of low absolute incidence. Finally, conditioned on reporting probability, larger populations exhibit less cryptic presence than smaller populations, particularly for pertussis (Figure 2, inset).

Temporal aggregation

Model fits, including the effects of temporal aggregation, are shown in Figure S1. As predicted, we find that (cloglog-transformed) $\widetilde{P}$ increases linearly in both $\log(N)$ and $\log(\Delta_t)$. The slope of $\widetilde{P}$ in response to N is steeper in pertussis, suggesting that pertussis reaches complete presence more quickly than measles with increasing population size (as theory predicts).

Table S1 shows the gradual decay of model fidelity with increasing temporal aggregation, along with an associated reduction in sample counts, as cities with complete presence ($P_o = 1$) are omitted. A close inspection of Figure S1A also reveals, at high levels of aggregation, poor model fits in large populations, where observed presence is far below model predictions (i.e. large negative residuals). This pattern likely results from the small number of available time periods, which limits the range of values that P_o can adopt. At $\Delta_t = 16$ weeks, for example, the maximum number of (non-excluded) periods per city is 65 (pertussis) and 71 (measles), such that the maximum incomplete P_o is approximately 0.985 and 0.986, respectively.

Discussion

Despite widespread availability of inexpensive and effective vaccines, childhood diseases have resisted elimination efforts. Classic epidemiological theory proposes that reducing the susceptible proportion of a population below $1/R_0$ should interrupt disease transmission, leading to local extinction [1]. Yet metapopulation elimination of disease has proven elusive and expensive: morbidity and mortality from vaccine-preventable diseases remains high in developing nations [40, 41], and importation of infection back into previously disease-free populations and metapopulations continues [42–44].

Where, when, and why vaccine-preventable diseases persist are key ecological questions with important modern epidemiological consequences. As we have shown, incomplete disease reporting substantially affects common measures of disease presence, particularly for low reporting probability and low absolute incidence. This impedes in-

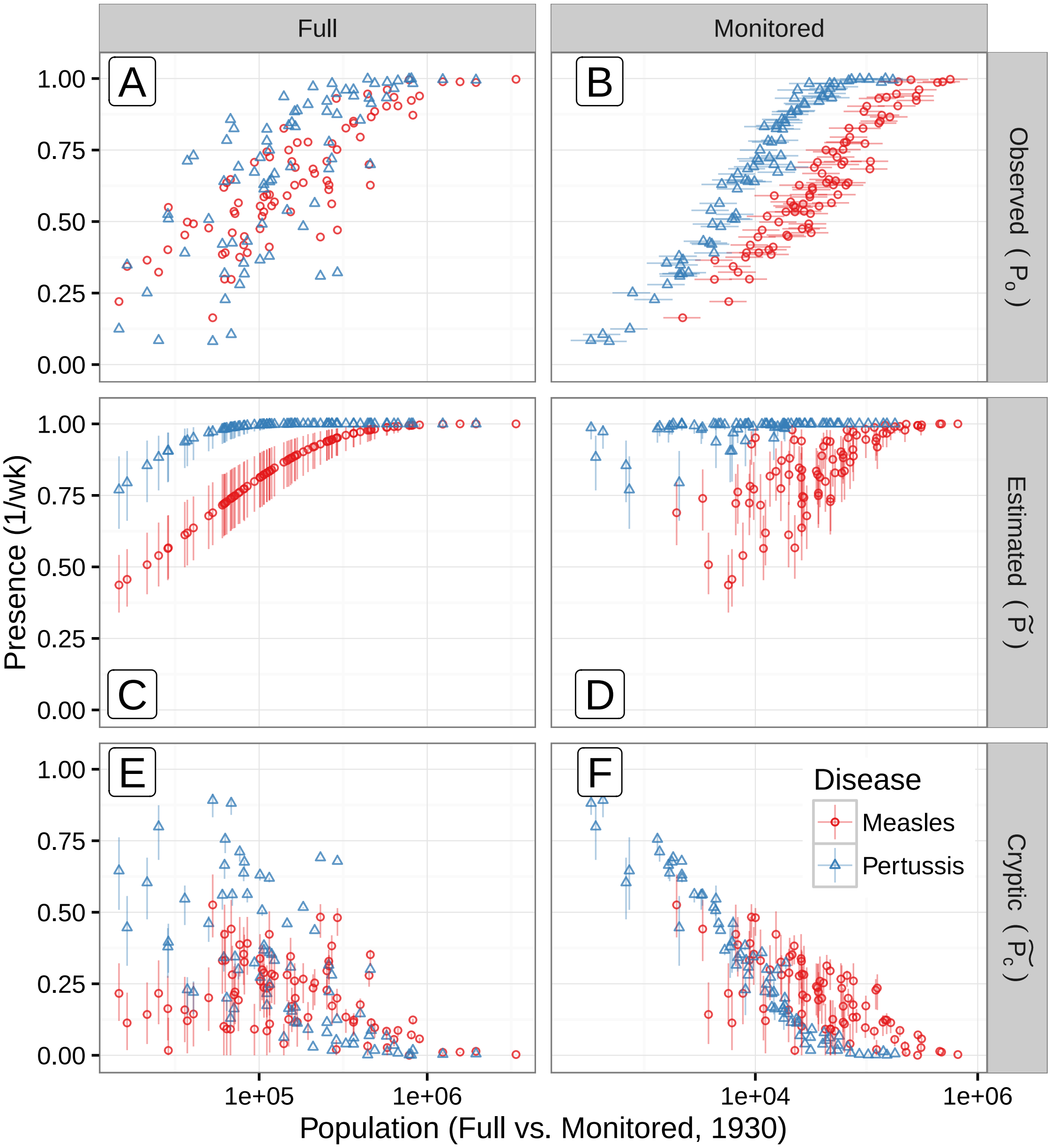

Figure 1: **Presence (P) by population size (N).** Columns: full population (N) and monitored population ($\widetilde{N_m} = N \times \widetilde{r}$). Rows: observed ($P_o$), estimated ($\widetilde{P}$), and cryptic ($\widetilde{P_c} = \widetilde{P} - P_o$) presence. **A:** Empirical observations of P_o versus N. **B:** N is scaled by incomplete reporting to yield $\widetilde{N_m}$. The response of P_o to $\widetilde{N_m}$ is modeled with a binomial GLM (cloglog link, one model per disease). **C:** The resulting models are used to estimate presence at full reporting ($\widetilde{P}$). Here, estimated presence of pertussis is higher than measles in all but the largest cities. **E,F:** For each city, cryptic presence ($\widetilde{P_c}$) is the difference between the previous two rows: $E = C - A$, and $F = D - B$. Note that movement between columns preserves Y-axis measures, and movement between rows preserves X-axis measures. **E:** Cryptic presence is uncommon in large cities, likely due to higher absolute incidence. See Figure S1 for model fits. Panel **F** show that, for pertussis, $\widetilde{P_c}$ increases predictably with both population size and reporting. Measles, on the other hand, shows considerable variation in the response of $\widetilde{P_c}$ to $\widetilde{N_m}$, suggesting a non-linear response of disease incidence to city size.

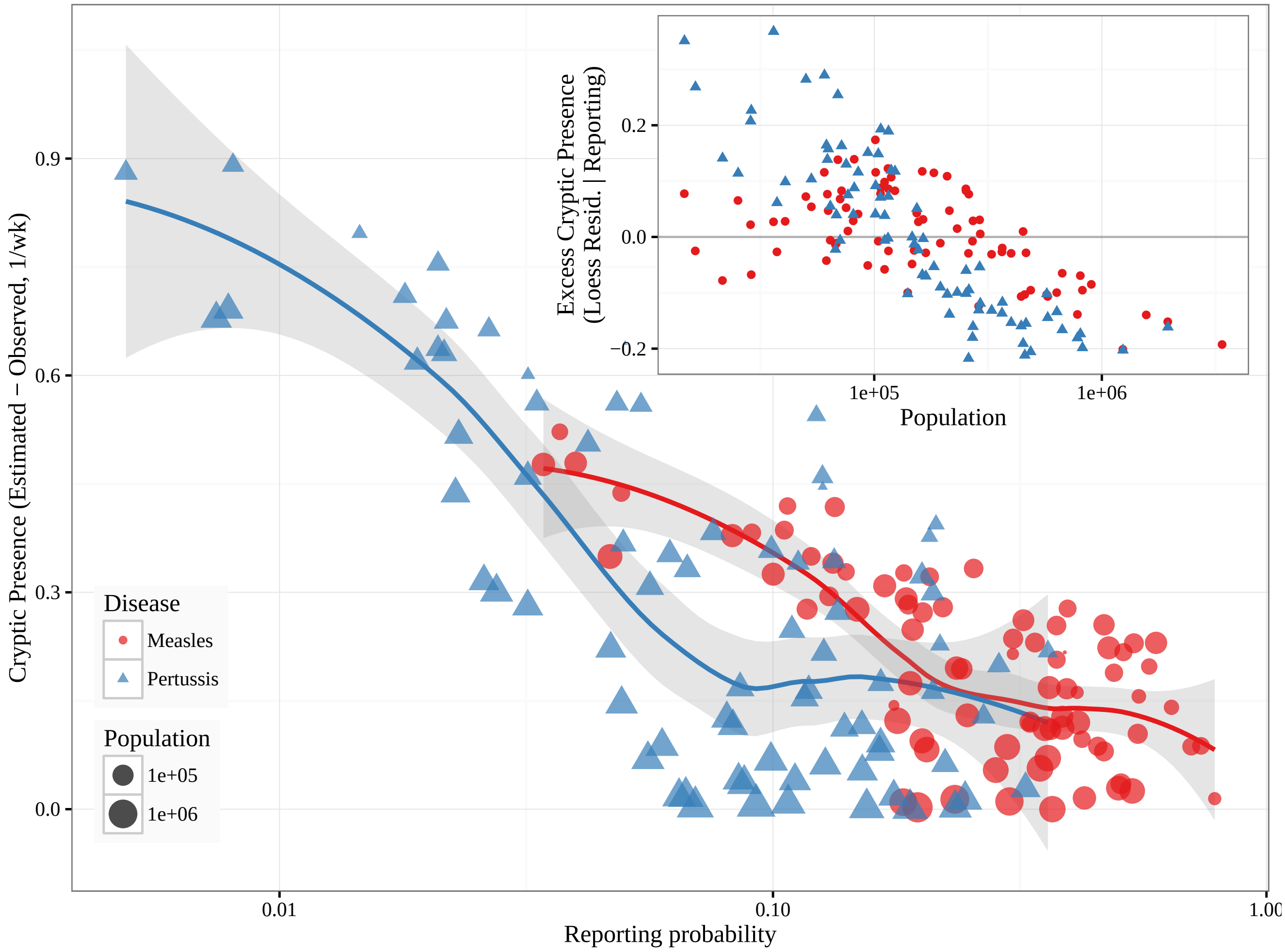

Figure 2: Cryptic presence ($\widetilde{P_c}$) by reporting probability ($\widetilde{r}$). Reporting of pertussis is less complete and more variable than pertussis; cryptic presence also varies widely in pertussis. A superimposed LOESS regression shows that, at low reporting probabilities, cryptic presence is strongly correlated with reporting probability. The residuals of the LOESS regression are also plotted against population size (inset figure). For a given reporting probability, larger cities generally exhibit less cryptic presence than smaller cities, particularly for pertussis. Cryptic presence is essentially absent in the largest cities, regardless of disease or reporting probability.

ference about disease dynamics at the local scale, and complicates comparisons between diseases or metapopulations with different reporting probabilities.

One particular area of practical concern that warrants increased attention is the fidelity of available demographic records in the modern era. Birth and migration rates help constrain reporting estimates and inform control measures [45]. Unfortunately, low birth registration coverage is common in modern developing nations [46], where modern incidence of vaccine-preventable diseases such as measles is highest [47]. In addition, completeness of birth registration varies greatly by geographic region and socioeconomic status [46, 48]. In some cases, multiple independent sources of demographic records can be employed to validate findings, such as the use of both government census records and survey-based Demographic and Health Surveys [49].

A key challenge in disease ecology is the unraveling of complex feedbacks between metapopulations and their constituent populations. Local disease persistence is driven both by local processes (birth, disease transmission) and metapopulation processes (host migration, disease importation). This study system pairs two different diseases within the same metapopulation, highlighting differences due to pathogen life history.

Here we estimate that cryptic presence is widespread in both diseases. We expect that cryptic presence is concentrated in cities that exhibit long periods of low but non-zero incidence, teetering on the edge of stochastic extinction. Yet the characteristics of these "refuge" populations differ markedly between diseases. We find that cryptic presence is concentrated at smaller populations in pertussis than in measles (Figures 1E and 2). This accords with epidemiological theory, which predicts that measles' high transmission rate and short infectious period leads to rapid susceptible depletion in small populations. Thus, small populations are expected to commonly experience measles extinction. Pertussis, on the other hand, can sustain low but non-zero incidence in much smaller populations than measles due to a longer infectious period and lower transmission rate.

Relation to Previous Work

Critical community size is one commonly employed threshold measure of disease persistence. CCS in particular, and threshold measures of extinction in general, has been widely criticized as poorly specified and difficult to measure [25, 50, 33]. In addition, cryptic presence should artificially inflate CCS estimates, as larger populations appear to undergo stochastic extinction. Nonetheless, the CCS of a disease remains a commonly reported "feature" of empirical data. For comparison, we present a simple empirical definition of CCS: the minimum population size where observed or estimated presence (P_o and $\widetilde{P}$, resp.) exceeds 95% (i.e., min(Population) given $P > 0.95$; see also Figure S2). The effects of incomplete reporting here are dramatic: for measles, CCS changes from ≈ 580 thousand (P_o) to ≈ 330 thousand ($\widetilde{P}$), while for pertussis, CCS changes from ≈ 210 thousand (P_o) to 50 thousand ($\widetilde{P}$).

We expect that lower metapopulation incidence should, in general, decrease local persistence by reducing disease importation. How local persistence scales up to metapopulation persistence is less clear. Conventional epidemiological wisdom [51, 52] holds that metapopulation persistence depends on local persistence in focal cities above a critical size (i.e., CCS). Recent work suggests that aggregates of medium-sized cities exhibit patterns of persistence similar to individual cities of comparable size [33]. Our estimates of widespread cryptic presence in cities experiencing low absolute incidence further emphasizes the role that "non-focal" cities can play in metapopulation persistence.

Implications for Disease Detection and Control

Recent detection of wild-type polio in Nigeria [53] clearly illustrates that failure to account for cryptic presence can lead to biased assessments of control effort efficacy, and mistaken allocation of control efforts away from areas where disease remains present. Previous work has demonstrated the high likelihood of unobserved incidence in polio, where low absolute incidence and poor reporting commonly co-occur [54]. Our results provide a clear warning against overly optimistic interpretations of apparent disease absence, and the critical importance of ongoing surveillance efforts.

More generally, the observed interdependence between cryptic presence, incomplete reporting, and absolute incidence adds uncertainty to ongoing disease control efforts. As disease incidence is reduced, the frequency of cryptic presence is expected to become more sensitive to incomplete reporting. On the other hand, successful control measures will potentially lower cryptic presence in small populations, as those populations transition from low but non-zero incidence into true extinction. Indeed, this pattern has been observed in pertussis in England & Wales [55].

Here we show that cryptic presence may have a complex and disease-specific relationship with population size; we also provide a method for estimating this relationship from historical surveillance records. In practice, these methods could be used to define an optimal allocation of resources for an active surveillance strategy: first, to identify when elimination has been achieved at the meta-population scale and second, to monitor the maintenance of elimination. The results presented here suggest, for example, that additional resources for pertussis monitoring would be best allocated towards active surveillance in smaller populations (Figure 1E).

An additional complication is that disease monitoring intensity is commonly tied to disease incidence. This could lead to the paradoxical increase in cryptic presence as a disease approaches elimination due to reduced monitoring efforts. One example is pertussis, where high vaccination rates in developed nations have decreased incidence to very low levels [56, 55, 57]. In some locations, low incidence has led to the cessation of routine disease surveillance [20]. Active surveillance, on the other hand, has revealed widespread unreported incidence [20], including asymptomatic infection and subsequent transmission [58, 59].

For immunizing diseases, cryptic incidence does serve to increase natural immune boosting, even in the presence of widespread vaccination. The well-known "honeymoon period" [60] refers to the combined benefits of disease-induced and vaccine-induced immunity in a population shortly after the introduction of vaccination. As disease incidence falls, however, disease-induced immunity drops, leading to paradoxical negative feedback between vaccine-induced immunity and immunity from natural infection [45]. Cryptic incidence again adds an element of uncertainty regarding the long-term immune status of populations. Our results suggest that active monitoring could be used to identify sero-conversion or immune boosting [21, 61] from cryptic incidence, which could, in turn, inform ongoing control efforts.

The above-noted uncertainties highlight the need for novel, cost-effective monitoring to assess the frequency of cryptic presence. One example is genetic sequence monitoring, which could provide evidence of local or metapopulation persistence. Increased understanding of pathogen persistence could, in turn, inform phylodynamic models that seek to couple the ecology and evolution of human diseases [62–64], as well as provide novel insight into patterns of host metapopulation connectivity.

Here we use a well-studied, highly-constrained system to show that cryptic presence was both common and explicable in the pre-vaccine era U.S. Our work, along with recent public health developments, suggests that attention to cryptic presence in other disease systems is warranted. Widespread asymptomatic malaria incidence in Southeast Asia, for example, has been suggested as a potential reservoir of artemisinin-resistant *Plasmodium falciparum* [65, 66], whose spread represents a "major threat to global public health" [67]. In the case of wild-type polio, eradication of the last few cases has proved both expensive and logistically challenging [68]. After two years of apparent disease absence, the recent detection wild-type polio virus' endemic persistence in Nigeria argues strongly against complacency in disease surveillance efforts [53]. Clearly, an improved understanding cryptic diseases presence stands to benefit both ongoing and future disease control efforts.

Author Contributions

CEG and HJW designed the study. CEG performed the analyses and wrote the first manuscript draft. EBE contributed to design and execution of analyses. All authors contributed to subsequent manuscript revisions.

Acknowledgments

Comments from four anonymous reviewers greatly improved the manuscript.

The authors would like to thank Natalie Wright, Michael A. Robert, Michael Chang, James H. Brown, and Melanie E. Moses for their support and assistance.

MF was supported under a grant from the Bill and Melinda Gates Foundation and the RAPIDD Program of the Science and Technology Directorate of the Department of Homeland Security. CG was supported by a fellowship in the Program in Interdisciplinary Biological and Biomedical Sciences at the University of New Mexico. This publication was made possible by Grant Numbers P20RR018754 from the National Center for Research Resources (NCRR), T32EB009414 from the National Institute of Biomedical Imaging and Bioengineering (NIBIB), components of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NCRR, NIBIB, or NIH.

Data Accessibility

- U.S. case report data and demographics available at Data Dryad

References

[1] R.M. Anderson and R.M. May. Directly transmitted infectious diseases: control by vaccination. *Science*, 215 (4536):1053–1060, 1982.

[2] M.S. Bartlett. Deterministic and stochastic models for recurrent epidemics. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 4, pages 81–109. University of California Press Berkeley, 1956.

[3] W.P. London and J.A. Yorke. Recurrent outbreaks of measles, chickenpox and mumps: I. Seasonal variation in contact rates. *Am. J. Epidemiol.*, 98(6):453–468, 1973.

[4] P.E.M. Fine and J.A. Clarkson. Measles in England and Wales. I. An analysis of factors underlying seasonal patterns. *Int J Epidemiol*, 11(1):5–14, 1982.

[5] M.J. Ferrari, R.F. Grais, N. Bharti, A.J.K. Conlan, O.N. Bjørnstad, L.J. Wolfson, P.J. Guerin, A. Djibo, and B.T. Grenfell. The dynamics of measles in sub-Saharan Africa. *Nature*, 451(7179):679–684, 2008.

[6] N. Bharti, A.J. Tatem, M.J. Ferrari, R.F. Grais, A. Djibo, and B.T. Grenfell. Explaining seasonal fluctuations of measles in Niger using nighttime lights imagery. *Science*, 334(6061):1424–1427, 2011.

[7] P. Rohani, D.J.D. Earn, and B.T. Grenfell. Opposite patterns of synchrony in sympatric disease metapopulations. *Science*, 286(5441):968, 1999.

[8] M.S. Bartlett. Measles periodicity and community size. *J R Stat Soc Ser A*, 120(1):48–70, 1957.

[9] R.M. Anderson, B.T. Grenfell, and R.M. May. Oscillatory fluctuations in the incidence of infectious disease and the impact of vaccination: time series analysis. *J Hyg (Lond)*, 93(03):587–608, 1984.

[10] M.C. Gomes, J.J. Gomes, and A.C. Paulo. Diphtheria, pertussis, and measles in Portugal before and after mass vaccination: A time series analysis. *Eur. J. Epidemiol.*, 15(9):791–798, 1999.

[11] D.J.D. Earn, P. Rohani, B.M. Bolker, and B.T. Grenfell. A simple model for complex dynamical transitions in epidemics. *Science*, 287(5453):667, 2000.

[12] C.T. Bauch and D.J.D. Earn. Transients and attractors in epidemics. *Proc. R. Soc. B*, 270(1524):1573–1578, 2003.

[13] L. Stone, R. Olinky, and A. Huppert. Seasonal dynamics of recurrent epidemics. *Nature*, 446(7135):533–536, 2007.

[14] H. Broutin, C. Viboud, B.T. Grenfell, M.A. Miller, and P. Rohani. Impact of vaccination and birth rate on the epidemiology of pertussis: a comparative study in 64 countries. *Proc. R. Soc. B*, pages 1–7, 2010. doi: 10.1098/rspb.2010.0994.

[15] P. Rohani, M.J. Keeling, and B.T. Grenfell. The interplay between determinism and stochasticity in childhood diseases. *Am. Nat.*, 159(5):469–481, 2002.

[16] H.T.H. Nguyen and P. Rohani. Noise, nonlinearity and seasonality: the epidemics of whooping cough revisited. *J R Soc Interface*, 5(21):403–413, 2008. doi: 10.1098/rsif.2007.1168.

[17] J.A. Clarkson and P.E.M. Fine. The efficiency of measles and pertussis notification in England and Wales. *Int J Epidemiol*, 14(1):153–168, 1985.

[18] D. He, E.L. Ionides, and A.A. King. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *J R Soc Interface*, 7(43):271–283, 2010.

[19] C.E. Gunning, E. Erhardt, and H.J. Wearing. Conserved patterns of incomplete reporting in pre-vaccine era childhood diseases. *Proc. R. Soc. B*, 281(1794):20140886, 2014.

[20] S. Baron, E. Njamkepo, E. Grimprel, P. Begue, J.C. Desenclos, J. Drucker, and N. Guiso. Epidemiology of pertussis in French hospitals in 1993 and 1994: thirty years after a routine use of vaccination. *Pediatr. Infect. Dis. J.*, 17(5):412–418, 1998.

[21] S. Mattoo and J.D. Cherry. Molecular pathogenesis, epidemiology, and clinical manifestations of respiratory infections due to Bordetella pertussis and other Bordetella subspecies. *Clin. Microbiol. Rev.*, 18(2):326–382, 2005. doi: 10.1128/CMR.18.2.326382.2005.

[22] M.S. Bartlett. The critical community size for measles in the United States. *J R Stat Soc Ser A*, 123(1):37–44, 1960.

[23] F.L. Black. Measles endemicity in insular populations: critical community size and its evolutionary implication. *J. Theor. Biol.*, 11(2):207–211, 1966.

[24] M.J. Keeling and B.T. Grenfell. Disease extinction and community size: modeling the persistence of measles. *Science*, 275(5296):65, 1997.

[25] I. Nåsell. A new look at the critical community size for childhood infections. *Theor Popul Biol*, 67(3):203–216, 2005.

[26] C.J.E. Metcalf, K. Hampson, A.J. Tatem, B.T. Grenfell, and O.N. Bjørnstad. Persistence in Epidemic Metapopulations: Quantifying the Rescue Effects for Measles, Mumps, Rubella and Whooping Cough. *PloS ONE*, 8 (9):e74696, 2013.

[27] I. Nåsell. On the time to extinction in recurrent epidemics. *Proc. R. Soc. B*, 61(2):309–330, 1999.

[28] H. Andersson and T. Britton. Stochastic epidemics in dynamic populations: quasi-stationarity and extinction. *J Math Biol*, 41(6):559–580, 2000.

[29] A.L. Lloyd. Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics. *Theor Popul Biol*, 60(1):59–71, 2001.

[30] H.J. Wearing and P. Rohani. Estimating the duration of pertussis immunity using epidemiological signatures. *PLoS Pathog*, 5(10):e1000647, 2009.

[31] A.J.K. Conlan, P. Rohani, A.L. Lloyd, M. Keeling, and B.T. Grenfell. Resolving the impact of waiting time distributions on the persistence of measles. *J R Soc Interface*, 7(45):623, 2010.

[32] I. Hanski. Metapopulation dynamics. *Nature*, 396(6706):41–49, 1998.

[33] C.E. Gunning and H.J. Wearing. Probabilistic measures of persistence and extinction in measles (meta)populations. *Ecol. Lett.*, 16:985–994, 2013.

[34] Edward F Connor and Daniel Simberloff. Species number and compositional similarity of the galapagos flora and avifauna. *Ecological Monographs*, 48(2):219–248, 1978.

[35] Bruno Andreas Walther, P Cotgreave, RD Price, RD Gregory, and Dale H Clayton. Sampling effort and parasite species richness. *Parasitology Today*, 11(8):306–310, 1995.

[36] Nicholas J Gotelli and Robert K Colwell. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology letters*, 4(4):379–391, 2001.

[37] Patrick D Schloss and Jo Handelsman. Introducing dotur, a computer program for defining operational taxonomic units and estimating species richness. *Applied and environmental microbiology*, 71(3):1501–1506, 2005.

[38] Carsten Rahbek. The role of spatial scale and the perception of large-scale species-richness patterns. *Ecology letters*, 8(2):224–239, 2005.

[39] Anne Chao, Robert K Colwell, Chih-Wei Lin, and Nicholas J Gotelli. Sufficient sampling for asymptotic minimum species richness estimators. *Ecology*, 90(4):1125–1133, 2009.

[40] N.S. Crowcroft, C. Stein, P. Duclos, and M. Birmingham. How best to estimate the global burden of pertussis? *Lancet Infect Dis*, 3(7):413–418, 2003. doi: 10.1016/S1473-3099(03)00669-8.

[41] R.E. Black, S. Cousens, H.L. Johnson, J.E. Lawn, I. Rudan, D.G. Bassani, P. Jha, H. Campbell, C.F. Walker, R. Cibulskis, T. Eisele, L. Liu, and C. Mathers. Global, regional, and national causes of child mortality in 2008: a systematic analysis. *Lancet*, 375(9730):1969–1987, 2010. doi: 10.1016/S0140-6736(10)60549-1.

[42] M.N. Mulders, A.T. Truong, and C.P. Muller. Monitoring of measles elimination using molecular epidemiology. *Vaccine*, 19(17):2245–2249, 2001.

[43] S.L. Katz, J.I. Santos, M.A. Nakamura, M.V. Godoy, P. Kuri, C.A. Lucas, and R.T. Conyer. Measles in Mexico, 1941–2001: Interruption of Endemic Transmission and Lessons Learned. *J. Infect. Dis.*, 189(Supplement 1): S243–S250, 2004. doi: 10.1086/378520.

[44] E. Kaliner, J. Moran-Gilad, I. Grotto, E. Somekh, E. Kopel, M. Gdalevich, E. Shimron, Y. Amikam, A. Leventhal, B. Lev, and R. Gamzu. Silent reintroduction of wild-type poliovirus to Israel, 2013–risk communication challenges in an argumentative atmosphere. *Euro Surveill*, 19(7):207030, 2014.

[45] M.J. Ferrari, B.T. Grenfell, and P.M. Strebel. Think globally, act locally: the role of local demographics and vaccination coverage in the dynamic response of measles infection to control. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 368(1623):20120141, 2013.

[46] The United Nations Children's Fund (UNICEF). The 'Rights' start to life: a statistical analysis of birth registration, 2005. URL `http://www.unicef.org/publications/index_25248.html`. Accessed Sep 2014.

[47] World Health Organization. *World Health Statistics*. WHO Press, World Health Organization, Geneva, Switzerland, 2010. ISBN 978 92 4 156398 7.

[48] P.W. Setel, S.B. Macfarlane, S. Szreter, L. Mikkelsen, P. Jha, S. Stout, and C. AbouZahr. A scandal of invisibility: making everyone count by counting everyone. *The Lancet*, 370(9598):1569–1577, 2007.

[49] B. Schoumaker. Quality and consistency of DHS fertility estimates, 1990 to 2012. Technical Report DHS Methodological Reports No. 12, ICF International, Rockville, Maryland, USA, 2014. URL `http://dhsprogram.com/pubs/pdf/MR12/MR12.pdf`.

[50] J.O. Lloyd-Smith, P.C. Cross, C.J. Briggs, M. Daugherty, W.M. Getz, J. Latto, M.S. Sanchez, A.B. Smith, and A. Swei. Should we expect population thresholds for wildlife disease? *Trends Ecol. Evol.*, 20(9):511–519, 2005.

[51] W.H. McNeill. *Plagues and Peoples.* Anchor, 1977.

[52] A.J.K. Conlan and B.T. Grenfell. Seasonality and the persistence and invasion of measles. *Proc. R. Soc. B*, 274(1614):1133–1141, 2007.

[53] Leslie Roberts. New polio cases in nigeria spur massive response. *Science*, 353(6301):738–738, 2016.

[54] Micaela Martinez-Bakker, Aaron A King, and Pejman Rohani. Unraveling the transmission ecology of polio. *PLoS Biol*, 13(6):e1002172, 2015.

[55] P. Rohani, D.J.D. Earn, and B.T. Grenfell. Impact of immunisation on pertussis transmission in England and Wales. *Lancet*, 355(9200):285–286, 2000.

[56] J.W. Bass and S.R Stephenson. The return of pertussis. *Pediatr. Infect. Dis. J.*, 6(2):141–144, 1987.

[57] J.C. Blackwood, D.A.T. Cummings, H. Broutin, S. Iamsirithaworn, and P. Rohani. Deciphering the impacts of vaccination and immunity on pertussis epidemiology in Thailand. *Proc. Natl. Acad. Sci. U.S.A.*, 110(23): 9595–9600, 2013.

[58] S.C. de Greeff, F.R. Mooi, A. Westerhof, J.M.M. Verbakel, M.F. Peeters, C.J. Heuvelman, D.W. Notermans, L.H. Elvers, J.F.P. Schellekens, and H.E. de Melker. Pertussis Disease Burden in the Household: How to Protect Young Infants. *Clin. Infect. Dis.*, 50(10):1339–1345, 2010.

[59] A.M. Wendelboe, E. Njamkepo, A. Bourillon, D.D. Floret, J. Gaudelus, M. Gerber, E. Grimprel, D. Greenberg, S. Halperin, J. Liese, et al. Transmission of Bordetella pertussis to young infants. *The Pediatric infectious disease journal*, 26(4):293–299, 2007.

[60] A.R. McLean and R.M. Anderson. Measles in developing countries. Part II. The predicted impact of mass vaccination. *Epidem. Inf.*, 100:419–442, 1988.

[61] J.S. Lavine, A.A. King, and O.N. Bjørnstad. Natural immune boosting in pertussis dynamics and the potential for long-term vaccine failure. *Proc. Natl. Acad. Sci. U.S.A.*, 108(17):7259–7264, 2011.

[62] B.T. Grenfell, O.G. Pybus, J.R. Gog, J.L.N. Wood, J.M. Daly, J.A. Mumford, and E.C. Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656):327–332, 2004.

[63] K. Koelle, S. Cobey, B. Grenfell, and M. Pascual. Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science*, 314(5807):1898–1903, 2006.

[64] Stacy O Scholle, Rolf JF Ypma, Alun L Lloyd, and Katia Koelle. Viral substitution rate variation can arise from the interplay between within-host and epidemiological dynamics. *The American Naturalist*, 182(4):494–513, 2013.

[65] H. Noedl, D. Socheat, and W. Satimai. Artemisinin-Resistant Malaria in Asia. *N. Engl. J. Med.*, 361(5):540–541, 2009. doi: 10.1056/NEJMc0900231. URL `http://www.nejm.org/doi/full/10.1056/NEJMc0900231`. PMID: 19641219.

[66] B. Wang, S. Han, C. Cho, J. Han, Y. Cheng, S. Lee, G.N.L. Galappaththy, K. Thimasarn, M.T. Soe, H.W. Oo, et al. Comparison of Microscopy, Nested-PCR, and Real-Time-PCR Assays Using High-Throughput Screening of Pooled Samples for Diagnosis of Malaria in Asymptomatic Carriers from Areas of Endemicity in Myanmar. *J. Clin. Microbiol.*, 52(6):1838–1845, 2014. doi: 10.1128/JCM.03615-13.

[67] Ambrose O Talisuna, Corine Karema, Bernhards Ogutu, Elizabeth Juma, John Logedi, Andrew Nyandigisi, Modest Mulenga, Wilfred F Mbacham, Cally Roper, Philippe J Guerin, et al. Mitigating the threat of artemisinin resistance in africa: improvement of drug-resistance surveillance and response systems. *The Lancet infectious diseases*, 12(11):888–896, 2012.

[68] S.G.F. Wassilak, M.S. Oberste, R.H. Tangermann, O.M. Diop, H.S. Jafari, and G.L. Armstrong. Progress toward global interruption of wild poliovirus transmission, 2010–2013, and tackling the challenges to complete eradication. *Journal of Infectious Diseases*, 210(suppl 1):S5–S15, 2014.

Supplemental Information for *Evidence of cryptic incidence in childhood diseases*

Christian E. Gunning, Matthew J. Ferrari, Erik Erhardt, and Helen J. Wearing, 2016.

Disease	Δ_t (Weeks)	# Cities	# Δ_t (Max.)	Pseudo-R^2
Measles	1	82	1147	0.908
Measles	2	80	573	0.886
Measles	4	74	286	0.827
Measles	8	59	143	0.715
Measles	16	34	71	0.635
Pertussis	1	79	1146	0.962
Pertussis	2	74	572	0.946
Pertussis	4	61	283	0.905
Pertussis	8	46	141	0.857
Pertussis	16	30	65	0.726

Table S1: Binomial GLMs predict the response of observed presence (P_o) to monitored population size ($\log \widetilde{N_m}$), for different temporal aggregation lengths (Δ_t, weeks). To aggregate case reports, missing weeks are omitted, and periods containing only missing weeks are excluded. For each Δ_t and city, $P_o = Pr[(\sum_{\Delta_t} C) > 0]$. Cities with $P_o = 1$ (disease always present) are omitted, and cities are weighted by the number of non-excluded periods. As aggregation increases, fewer cities are available, while the number of observed time periods per city decreases. Pseudo-R^2s (proportion explained deviance) show a decrease in R^2 at large Δ_t (see Figure S1). For illustrative purposes, each table row shows a separate model (i.e., each disease and Δ_t).

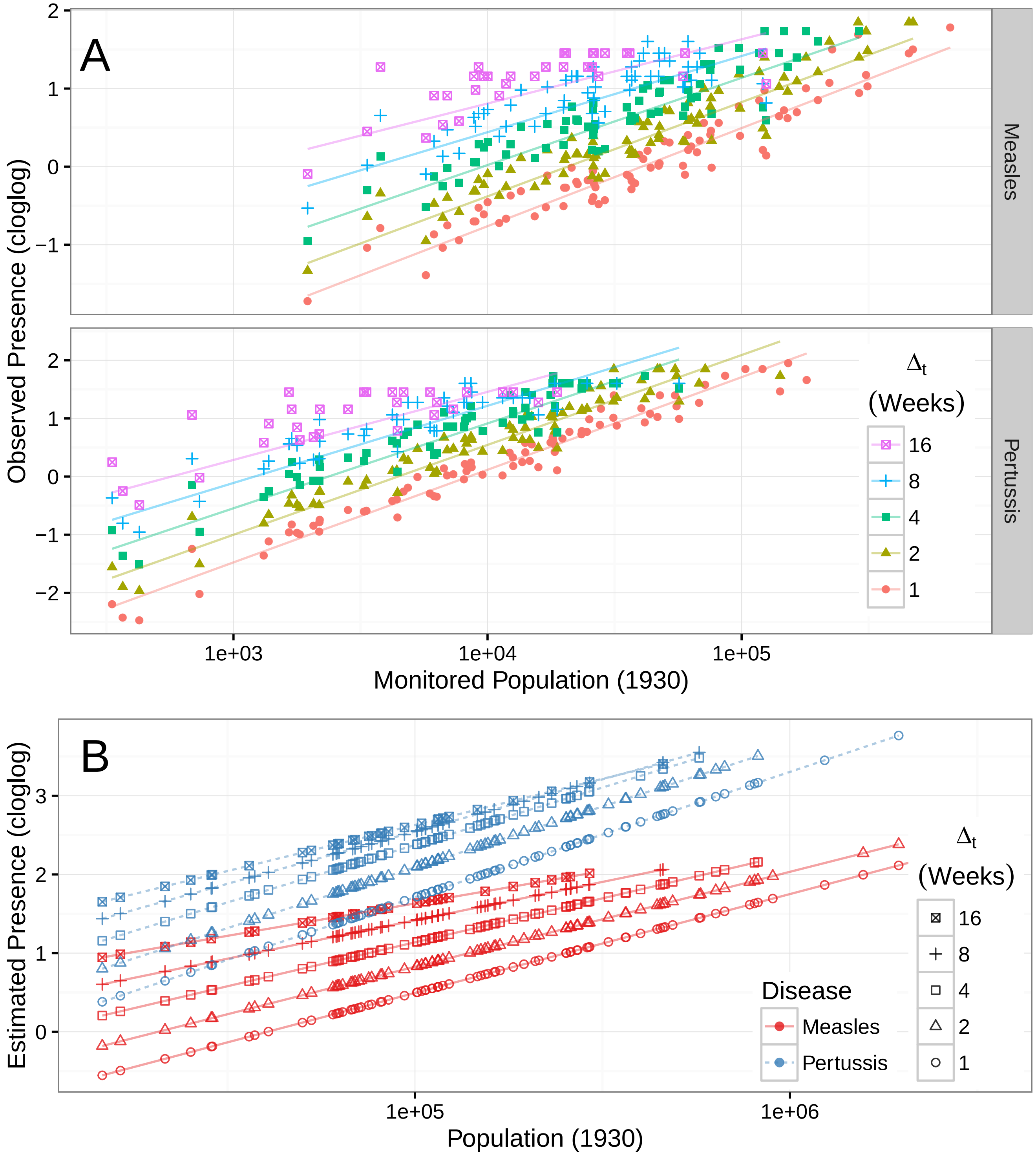

Figure S1: Binomial GLMs: **A: Observed Presence.** Response of observed presence (P_o) to monitored population size ($\widetilde{N_m}$) and aggregation period length (Δ_t, weeks). Lines show model estimates, with a separate model for each disease, using a cloglog link (y-axis, $f(\dot{})$): $f(P_o) \sim \log \widetilde{N_m})| \log \Delta_t$. See Table S1 for details and observation counts. **B: Estimated Presence.** Model estimated presence ($\widetilde{P}$) in response to population size (N, 1930): $f(\widetilde{P}) \sim \log N|log\Delta_t$.

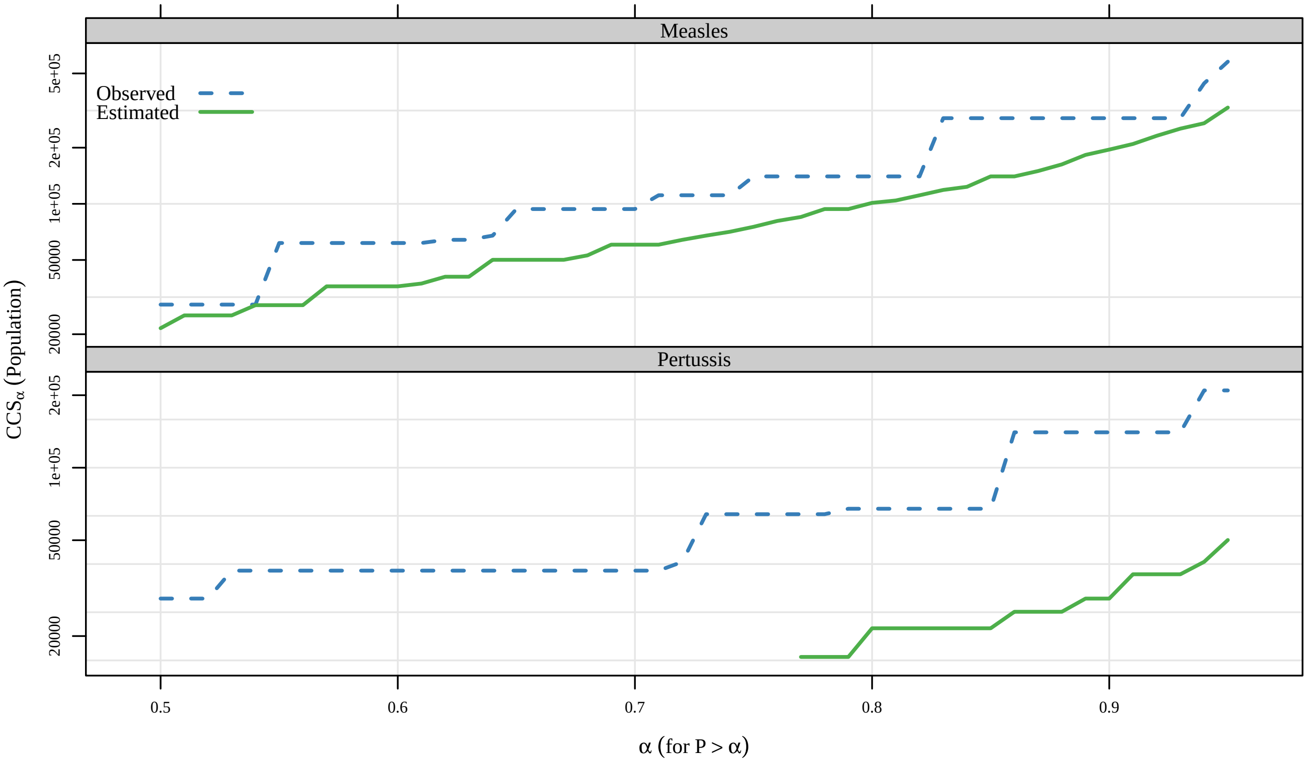

Figure S2: Empirical estimates of CCS_α based on disease presence (P), defined as the minimum population size (N) such that $\alpha < P$ (for $0 < \alpha < 1$). Results shown for both observed presence (P_o, blue dashed line) and estimated presence ($\widetilde{P}$, green solid line). Thus, CCS_α is the minimum N where the disease is present more than α proportion of sampled weeks. Pertussis is estimated to be present in all cities at $\alpha = 0.76$ (i.e., present in more than 76% of sampled weeks).

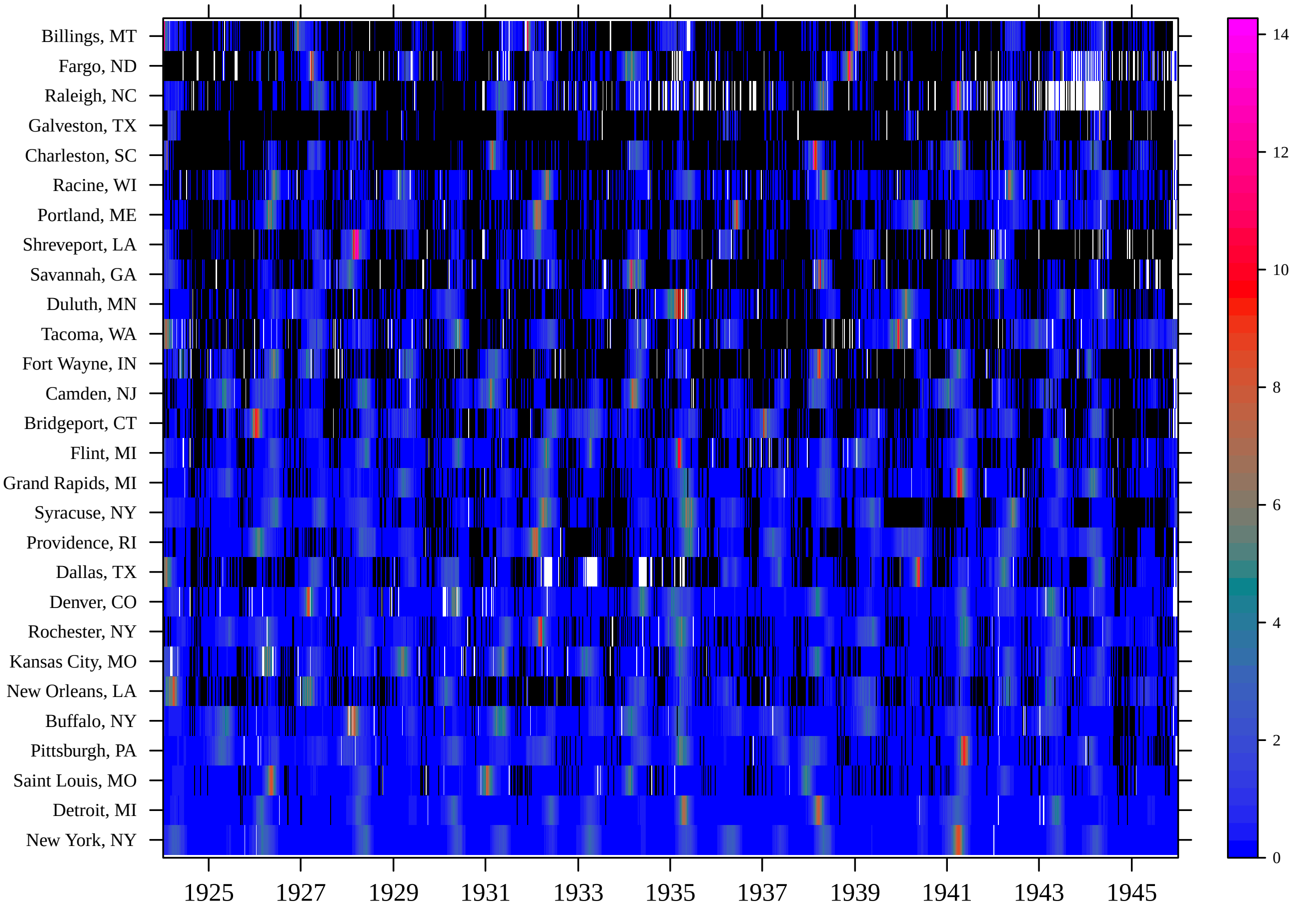

Figure S3: **Measles.** Variance-scaled case reports per sample period, with locations ordered by population size (black = 0; white = missing). Every third location is shown.

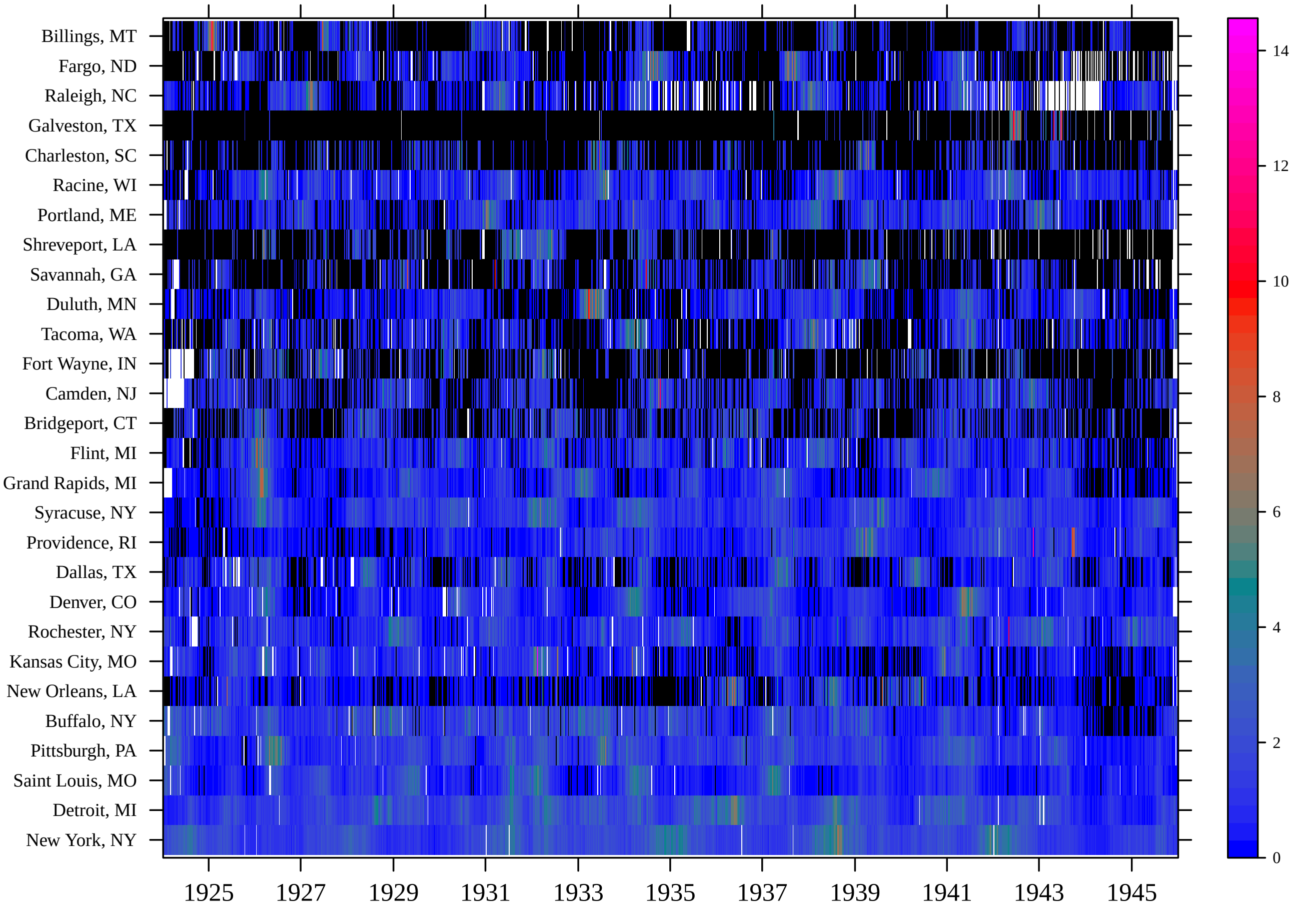

Figure S4: **Pertussis.** Variance-scaled case reports per sample period, with locations ordered by population size (black = 0; white = missing). Every third location is shown.