bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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: research.2016@x14n.org**

Abstract

Persistence and extinction are key processes in infectious disease dynamics that, due to incomplete reporting, are seldom directly observable. For fully-immunizing diseases, reporting probabilities can be readily 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., marginal probability of non-zero incidence), which provides an upper bound on the marginal probability of disease extinction. We examine measles and pertussis in pre-vaccine era U.S. cities, and describe a conserved scaling relationship between population size, reporting probability, and observed presence (i.e., non-zero case reports). We use this relationship to estimate disease presence given perfect reporting, and define cryptic presence as the difference between estimated and observed 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. While the methods employed here are specific to fully-immunizing diseases, our results suggest that cryptic incidence deserves careful attention, particularly in diseases with low case counts, poor reporting, and longer infectious periods.

Keywords

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

Introduction

Epidemic Dynamics of Childhood Diseases

Measles and pertussis (whooping cough) are acutely infectious diseases caused by obligate human pathogens: the measles virus and *Bordetella pertussis*, respectively. These well-studied childhood diseases are fully immunizing but highly infectious, with a low average age of infection (< 10 years) in the pre-vaccine era [1]. Both diseases have fast life cycles compared to human host demographics [1].

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Recurrent epidemics are a common feature of these diseases, driven by long-term host demographics and periodic forcing of disease transmission via changes in host density, such as school terms [2–4] or economic migration [5, 6]. 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].

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]. In pertussis, the contribution of waning immunity to observed dynamics has been a subject of extensive debate, both in infection-derived and vaccine-derived immunity [17–19]. In the pre-vaccine era, however, pertussis dynamics are consistent with the dynamics of infections that confer relatively long-lasting immunity, irrespective of whether the mechanism is long-term protection or natural immune boosting [20, 21].

Reporting probabilities vary widely between these diseases, as well as between locations [3, 22–24]. Measles infection causes characteristic symptoms of fever, rash, and pathognomonic Koplik's spots [25]. Pertussis, on the other hand, exhibits age-dependent severity, and shares symptoms with many other common respiratory diseases [26, 27]. In addition, reports of pertussis in adults was generally absent in the pre-vaccine era [18]. Consequently, pertussis reporting is generally less complete and more variable than measles reporting [22, 24]. 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, evolution, and 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 [28–31], and for pathogens with short infectious periods [31, 32]. Indeed, previous work has shown that local disease persistence scales approximately log-linearly with population size [31, 28, 33–35, 20, 36]. Likewise, theory predicts that, when all else is equal, longer latent and infectious periods and higher birth rates should increase local persistence [33, 31].

At the metapopulation level, host migration allows for imported infections to "rescue" local chains of infection [37, 36, 32, 38]. Intermediate levels of connectivity can aid rescue effects and metapopulation persistence [37], while very high levels of connectivity can synchronize populations and decrease rescue effects [39]. Low connectivity between populations, on the other hand, can favor boom-bust cycles. Here, disease importation is 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.

Seasonal forcing plays a central role in these diseases, and can synchronize and/or accentuate periodic troughs of

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

incidence in populations and metapopulations [11, 40, 41]. Yet the interplay between periodic forcing and temporal patterns of incidence can be be complex [42]. We note that previous work on measles has focused largely on England & Wales, and assumed a unified school calendar [2, 4, 40]. U.S. school calendars are set by local municipalities, and have varied considerably over both time and space [43–45]. As such, metapopulation patterns of extinction in response to seasonal forcing likely differ between these countries.

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 2) 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. Life-long immunity provides a key dynamical constraint, allowing us to reliably estimate reporting probability. Finally, the absence 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 can seldom be directly observed. Previous work has estimated disease persistence from case reports, either in distinct human populations (e.g. cities, Conlan et al. [36]) or in metapopulations (e.g. countries, Metcalf et al. [32]). Lacking, however, are quantitative assessments of the impact of observational uncertainty on persistence estimates.

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 [46, 47]. In assemblages of species, sampling effort can be accounted for via accumulation or rarefaction curves that quantify presence via asymptotic richness [47–49]. Related work has explored the interdependence between detection probability and spatiotemporal resolution [50], and has quantified the expected additional sampling required to achieve asymptotic detection [51].

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" [50] by aggregating case reports over a range of successively longer reporting windows. 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.

Overview

We compare metapopulation patterns of presence between two diseases (measles and pertussis) within a single metapopulation. 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 use weekly, per-city disease case reports (C_{obs}) and reporting probabilities to estimate the marginal (weekly) probability of disease presence (P). We show that city population size (N) and reporting rate (r) predict observed presence (non-zero case reports, P_{obs}). We use this relationship to estimate the (weekly) probability of disease presence given full reporting (i.e.

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

probability of non-zero incidence, P_{est}). We find an increase in pertussis P_{est} relative to that of measles across a range of population sizes. In addition, we show that the observed scaling of P_{obs} with N and r is robust to temporal aggregation of case reports over longer reporting windows.

We define cryptic presence (P_c) as the difference between estimated and observed presence: the (estimated, weekly) probability of unobserved presence. We show that P_c scales with both population size and reporting probability, and is particularly common in small populations with low reporting probability.

Methods

All subsequent analyses were conducted separately for each disease. All rates and probabilities are per week, unless otherwise noted. See Table 1 for definitions. Data are described in detail in Gunning et al. [24].

Name	Symbol	Definition	Comments
Reporting window	W		time (weeks)
Case reports	C_{obs}		per time (W^{-1})
Case count	C		per time (W^{-1})
Reporting probability	r		Assumed constant
Population size	N		1930 census
Monitored population	N_m	$N \times r$	
Observed presence	P_{obs}	$\Pr(C_{obs} > 0)$	per time (W^{-1})
Estimated presence	P_{est}	$\Pr(C > 0)$	per time (W^{-1})
Cryptic presence	P_c	$P_{est} - P_{obs}$	per time (W^{-1})

Table 1: Definitions. Unless otherwise noted, a reporting window of one week was used (for case reports, case count, and presence probabilities).

Incomplete observation

For each city and disease, we estimate a single, time-marginalized reporting probability (r) from case reports and demographic records, as in Gunning et al. [24]. 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, 22]. As discussed in Gunning et al. [24], no strong evidence of time-variable reporting is apparent in this system over the period of record, and reporting probability is assumed to be constant over time.

We assume that case reports are generated via binomial sampling of cases: $C_{obs} \sim \text{Bin}(C, r)$. Thus, each city's r can be estimated from the ratio total case reports to total surviving births, summed over the period of observation: $r = \sum C_{obs} / \sum \text{births}$. As described in Gunning et al. [24], total surviving births are estimated from yearly per capita state birth rates and infant mortality rates, along with yearly city populations. We also estimate approximate confidence intervals on r by bootstrapping (yearly) birth and infant mortality rates.

Given a binomial sampling process, we can estimate cases as the ratio of case reports to reporting probability: $C = C_{obs}/r$. Yet this correction fails for $C_{obs} = 0$, where our best estimate of C is zero. For low reporting

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

probability and low incidence, a non-trivial proportion of observed zeros (i.e., $C_{obs} = 0$) result from unobserved non-zero incidence. Consider, for example, $r = 0.1$ and $C = 10$, such that $\Pr(C_{obs} = 0) = (1 - r)^C \approx 0.35$. Here, approximately 35% of case reports will be zero, thus yielding erroneous under-estimates of $C = 0$. In short, observed zeros result from a "mixed process" of disease absence ($C = 0$), together with unobserved, cryptic incidence: $(C > 0 \cap C_{obs} = 0)$. It is this unobserved presence of disease that we seek to quantify.

Estimated and Cryptic Presence

As noted above, we focus here on the marginal, per time probability of disease presence. We first exclude cities where a disease was always present ($\Pr(C_{obs} > 0) = 1$). We define the monitored population (N_m) as the full population scaled by the reporting probability: $N_m = N \times r$. Note that, at full reporting, $N = N_m$.

We employ a binomial generalized linear model (B-GLM) to model the response of P_{obs} to $\log M_m$, where disease presence ($C_{obs} > 0$) is equated with the binomial trial's "success". Reporting weeks with NAs were excluded, and each city was weighted by the number of non-excluded reporting weeks. We estimate a single 2-coefficient B-GLM (slope + intercept) for each disease using R's `glm` interface [52].

We use a complementary log-log (cloglog) link function (f): $f(P_{obs}) \sim \log N_m$. Unlike the more common logit link, the cloglog link hypothesizes an asymmetric response to the predictor. That is, at large population sizes, cities approach complete presence ($P_{obs} = 1$) more rapidly than a logit link predicts. This accords with biological intuition, where mechanistically distinct processes dominate near complete presence versus near complete absence. Further discussion of mechanistic biological interpretations of this model formulation is included below.

We use the resulting B-GLMs to extrapolate estimated disease presence from full population size: $f(P_{est}) \sim \log N$. As noted, 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): $P_{est} = P_{obs} + P_c$. As such, we estimate cryptic presence as the difference between estimated and observed presence: $P_c = P_{est} - P_{obs}$. Non-zero cryptic presence, in turn, provides evidence of cryptic incidence.

Model Exploration

We also explore the effect of the length of reporting windows. We sum case reports over reporting windows of varying widths W, ranging from 2 to 16 weeks. To compute reporting window sums ($C_{obs,W}$), NA weeks were omitted, and windows containing only NA weeks were excluded (excluded windows were common for pertussis). $P_{obs,W}$ was computed as the proportion of non-zero window sums: $\Pr(C_{obs,W} > 0)$. We then build a new model for each disease using both $\log N_m$ and W as predictors of $C_{obs,W}$.

The cloglog link function $f(x) = \log(-\log(1 - x))$ also provides a biologically relevant hazard analysis interpretation of the postulated model formulation. 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 window W is $\Pr(C = 0) = 1 - P \approx \exp(-\lambda SW)$. For a given population size N, the susceptible proportion is then S/N, and $P = \Pr(C > 0) = 1 - \exp(-\lambda N(S/N)W)$. The cloglog link then yields: $f(P) = \log(\lambda) + \log(S/N) + \log(N) + \log(W)$. Thus, we expect the transformed response ($f(P)$) to change linearly in both $\log(N)$ and $\log(W)$. In truth, the pop-

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ulations studied here are not at equilibrium, such that λ and S/N instead oscillate over time around a long-term mean. Nonetheless, the above analysis hypothesizes a functional relationship between P, N, and W, which we explore further below.

Estimating uncertainty

Both P_{est} and P_c are influenced by uncertainty arising from estimates of r, as well as GLM predictions. For each disease, we used a two-step process of bootstrap resampling to estimate the combined impact of reporting probability and B-GLM prediction uncertainty. Note that population size N changes over the period of record, and no uncertainty or variation therein is accounted for here.

First, bootstrap draws of r (henceforth r_b) were taken via non-parametric resampling. For each draw, yearly state birth rates and national infant mortality rates were resampled, and total births thus summed (see Gunning et al. [24] for details). The resulting r_b were used to compute N_m from N, and a B-GLM was fit to the result. In this way, 1e+04 models were fit.

These models were then used to extrapolate P_{est} from N. The following was conducted for each bootstrap model (above), and for each city within that model. To incorporate per-city variance of r into model predictions, N was back-estimated from a (new) bootstrapped N_m, which was then divided by the estimated reporting rate: $N = N \times (r_b/r)$. The model's expected value of P_{est} was then extrapolated from N. A random binomial sample was then taken, where the number of trials equaled the number of non-NA weeks for that city, with $\Pr(success) = E(P_{est})$. The bootstrap draw of P_{est} is then the proportion of successes.

Finally, P_c was computed from P_{obs} and the re-sampled P_{est}. The resulting bootstrap samples were used to construct prediction intervals (95% PI) for P_{est} and P_c.

Disease	Measles	Pertussis
Total Cities (with $P_{obs} < 1$)	82	79
Total Weeks	1148	1043
Date Range	1924 - 1945	1924 - 1943
Reporting Probability (r)	0.31 [0.18-0.41, 0.56]	0.11 [0.04-0.16, 0.77]
Observed Presence (P_{obs})	0.63 [0.49-0.79, 0.33]	0.68 [0.47-0.90, 0.40]
Estimated Presence (P_{est})	0.84 [0.75-0.95, 0.17]	0.98 [0.99-1.00, 0.04]
Cryptic Presence (P_c)	0.21 [0.11-0.29, 0.66]	0.31 [0.10-0.50, 0.82]

Table 2: Overview: number of included cities, time period of record (inclusive), and summary statistics for studied cities: (Mean [Q1-Q3, CV]), including observed, estimated, and cryptic presence (P_{obs}, P_{est}, and P_c, resp.).

Results

For reference, time series of sum case reports, as well as variance-scaled case reports of select cities, are shown in Figures S2 and S3. Summary statistics and observation counts are shown in Table 2. Overall, the average

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

reporting probability of pertussis is much lower than for measles, with a higher coefficient of variation among cities, as discussed in Gunning et al. [24]. Figure 1 shows P_{obs}, P_{est}, and P_c (rows) versus N and N_m (columns). Figure 1 also provides a visual illustration of $P_{est} - P_{obs} = P_c$, i.e., panels $E = C - A$, and $F = D - B$.

Regardless of disease, we expect a lower probability of presence in smaller populations [28–31]. Indeed, we find that population size (N) predicts observed presence (P_{obs}). As shown in Figure 1A, no difference between diseases is evident when solely case reports are considered. When reporting is considered, however, monitored population size (N_m) yields an excellent predictor of P_{obs} (Figure 1B): pseudo-R^2 = 0.908 (measles) and 0.958 (pertussis). We use the models shown in Figure 1A to predict P_{est} from N. The resulting predictions are shown in Figure 1C (plotted against N) and Figure 1D (plotted against N_m). Cryptic persistence (P_c) is simply $P_{est} - P_{obs}$, such that Figure 1E is the difference between Figure 1C and Figure 1A.

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 [33, 31], a pattern obscured by pertussis' low and variable reporting. Correcting for incomplete reporting, we estimate that pertussis presence (P_{est}) 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 case count 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 case count (Table 2, Figure 1E).

We expect cryptic presence to be a function of both reporting probability and the underlying distribution of case counts. Indeed, we observe increasing P_c with decreasing r for both diseases (Figure 2). We also find marked differences between diseases: for a given r, measles generally experiences higher P_c, possibly due to prolonged periods of low incidence. Finally, conditioned on r, larger populations exhibit lower P_c than smaller populations, particularly for pertussis (Figure 2, inset).

Temporal aggregation

Model fits, including the effects of temporal aggregation, are shown in Figure S1 and Table S1. As predicted, we find that (cloglog-transformed) P_{est} increases linearly in both $\log(N)$ and $\log(W)$. The slope of P_{est} 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 also shows the gradual decay of model fidelity with increasing temporal aggregation, along with an associated reduction in sample counts, as cities with complete presence ($P_{obs} = 1$) 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 reporting windows, limiting the range of values that P_{obs} can adopt. At $W = 16$ weeks, for example, the maximum number of (non-excluded) reporting windows per city is 61 (pertussis) and 71 (measles), such that the maximum incomplete P_{obs} is approximately 0.984 and 0.986, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

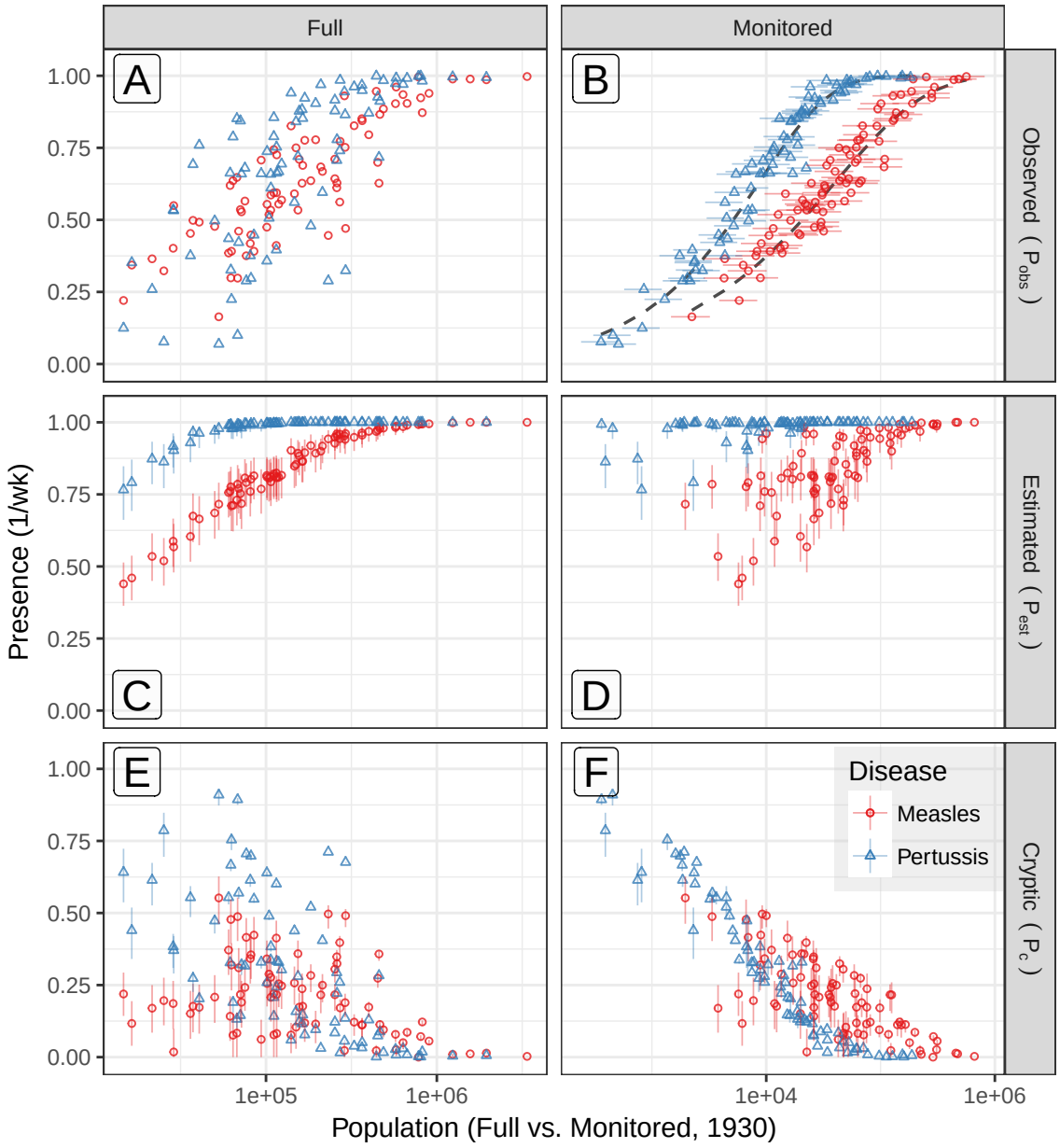

Figure 1: **Presence (P) by population size (N).** Columns: full population (N) and monitored population ($N_m = N \times r$). Rows: observed (P_{obs}), estimated (P_{est}), and cryptic ($P_c = P_{est} - P_{obs}$) presence. **A:** Empirical observations of P_{obs} versus N. **B:** N is scaled by incomplete reporting to yield N_m. Horizontal bars show uncertainty in r (95% CI). The response of P_{obs} to N_m is modeled with a binomial GLM (cloglog link, one model per disease). Dashed black lines show model fits (see Figure S1 for details). **C:** The resulting models are used to predict disease presence at full reporting ($P_{est}|r = 1$), along with 95% PI. Here, P_{est} of pertussis is higher than measles in all but the largest cities. **D:** As in C, but with N_m (see below). **E,F:** For each city, cryptic presence (P_c) is the difference between the previous two rows: $E = C - A$, and $F = D - B$. **E:** P_c is uncommon in large cities, likely due to the larger number of total cases per week. Panel **F** shows that, for pertussis, P_c increases predictably with both population size and reporting. Measles, on the other hand, shows considerable variation in the response of P_c to N_m, suggesting a non-linear response of disease incidence to city size.

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

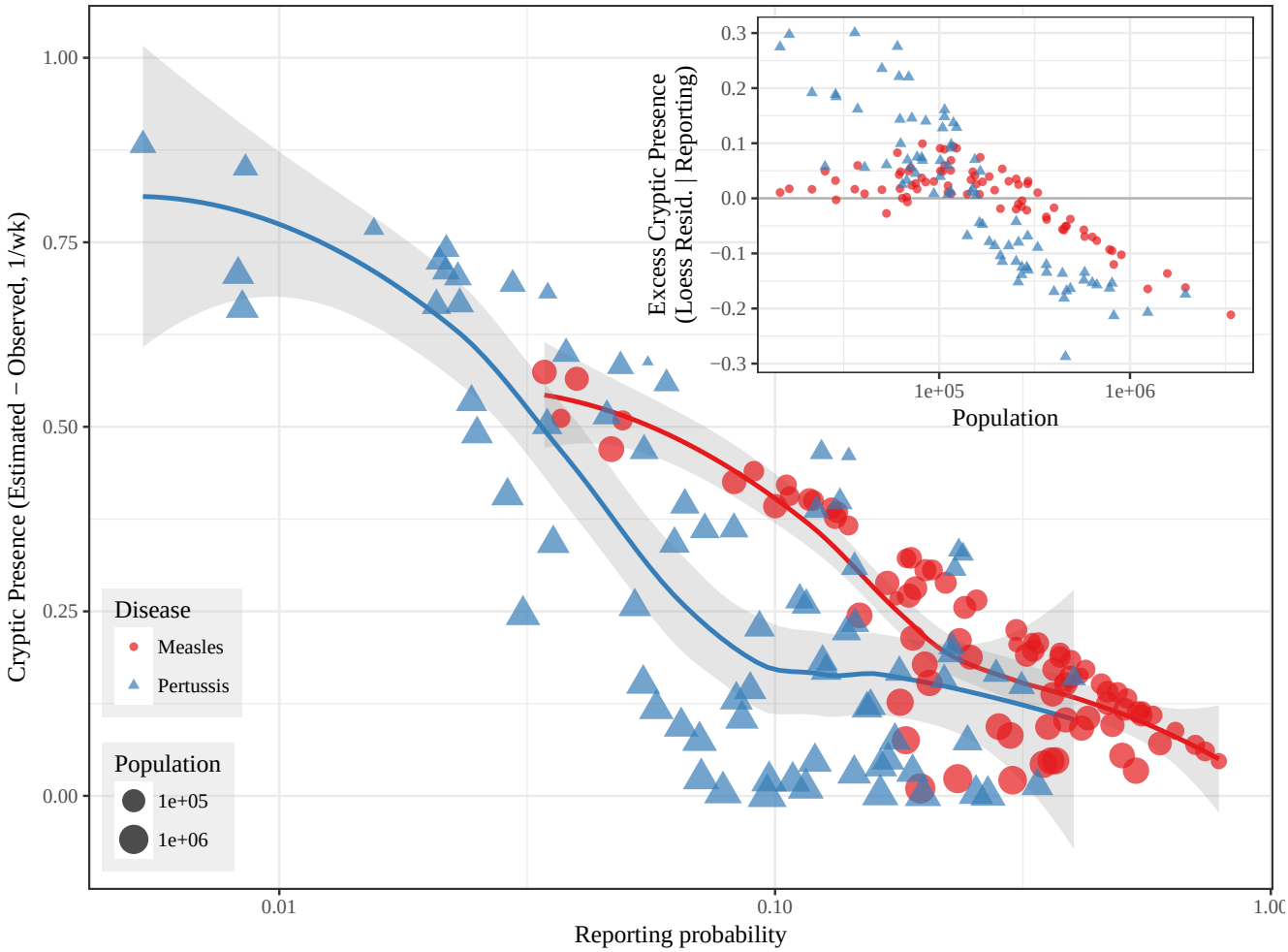

Figure 2: **Cryptic presence (P_c) by reporting probability (r).** Reporting of pertussis is less complete and more variable than measles; cryptic presence also varies widely in pertussis. A superimposed LOESS regression shows that, at low reporting probabilities, cryptic presence is strongly correlated 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.

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 [53, 54], and importation of infection back into previously disease-free populations and metapopulations continues [55–57].

Where, when, and why vaccine-preventable diseases persist are key ecological questions with important modern

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

epidemiological consequences. As we have shown, incomplete disease reporting substantially affects common measures of disease presence, particularly for low reporting probability and low case counts. This impedes inference 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 [58]. Unfortunately, low birth registration coverage is common in modern developing nations [59], where incidence of vaccine-preventable diseases such as measles is currently highest [60]. In addition, completeness of birth registration varies greatly by geographic region and socioeconomic status [59, 61]. 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 [62].

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

Incomplete reporting is a common feature in human diseases, but has received relatively little attention. A comprehensive review of incomplete reporting in this system is given in Gunning et al. [24]. Recent analysis of historical U.S. polio concludes that "absence of clinical disease is not a reliable indicator of polio transmission" due to unobserved incidence [63], highlighting the critical role incomplete reporting can play in modern disease control.

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 [31, 64, 38]. 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_{obs} and P_{est}, resp.) exceeds 95% (i.e., min(Population) given $P > 0.95$; Figure 3). The effects of incomplete reporting here are dramatic: for measles, CCS changes from $\approx$ 580 thousand (P_{obs}) to $\approx$ 330 thousand (P_{est}), while for pertussis, CCS changes from $\approx$ 210 thousand (P_{obs}) to 50 thousand (P_{est}).

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

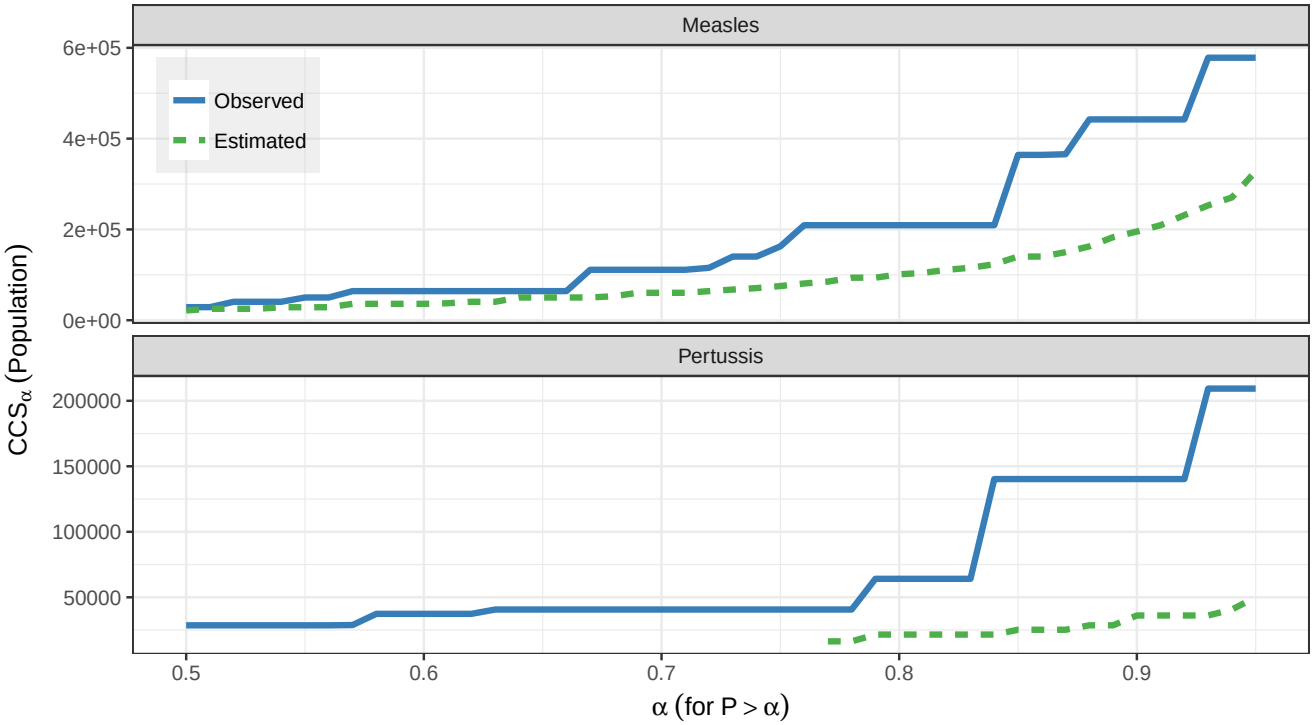

Figure 3: **Empirical estimates of** CCS_α: the minimum population size (N) such that $\alpha < P$ (for $0 < \alpha < 1$). Results shown for both observed presence (P_{obs}, blue dashed line) and estimated presence (P_{est}, green solid line). Thus, CCS_α is the minimum N where the disease is present 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).

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [65, 41] 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 [38]. Our estimates of widespread cryptic presence in cities experiencing low case counts 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 [66] 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 case counts and poor reporting commonly co-occur [63]. 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 case counts adds uncertainty to ongoing disease control efforts. As case counts drop, 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 [42].

Here we show that cryptic presence can have a complex and disease-specific relationship with population size; we also provide a method for estimating this relationship from historical surveillance records in fully immunizing diseases such as polio or rubella. While these *methods* are not directly applicable to multi-strain diseases such as influenza or dengue, or repeat infections such as malaria, we suggest that a similar relationship between incomplete monitoring and undetected incidence is likely, based purely on intrinsic sampling stochasticity in real-world disease detection.

In practice, these methods could assist in the optimal allocation of resources for an active surveillance strategy of a fully-immunizing disease: 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 [67, 42, 68]. In some locations, low incidence has led to the cessation of routine disease surveillance [26]. Active surveillance, on the other hand, has revealed widespread unreported incidence [26], including asymptomatic infection and subsequent transmission [69, 70].

For immunizing diseases, cryptic incidence does serve to increase natural immune boosting, even in the presence of widespread vaccination. The well-known "honeymoon period" [71] refers to the combined benefits of disease-

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [58]. 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 [27, 21] 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 awareness of pathogen persistence could, in turn, inform phylodynamic models that seek to couple the ecology and evolution of human diseases [72–74], 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* [75, 76], whose spread represents a "major threat to global public health" [77]. In the case of wild-type polio, eradication of the last few cases has proved both expensive and logistically challenging [78]. 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 [66]. An improved understanding of the relationship between disease monitoring effort and cryptic disease presence can benefit modern 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.

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data Accessibility

- U.S. case report data and demographics available at Data Dryad, `doi:10.5061/dryad.92p46`

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.

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[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] Aaron M Wendelboe, Annelies Van Rie, Stefania Salmaso, and Janet A Englund. Duration of immunity against pertussis after natural infection or vaccination. *The Pediatric infectious disease journal*, 24(5):S58–S61, 2005.

[18] James D Cherry. Adult pertussis in the pre-and post-vaccine eras: lifelong vaccine-induced immunity? *Expert review of vaccines*, 13(9):1073–1080, 2014.

[19] Matthieu Domenech de Cellès, Felicia M. G. Magpantay, Aaron A. King, and Pejman Rohani. The pertussis enigma: reconciling epidemiology, immunology and evolution. *Proc. R. Soc. B*, 283(1822), 2016. ISSN 0962-8452. doi: 10.1098/rspb.2015.2309. URL `http://rspb.royalsocietypublishing.org/content/283/1822/20152309`.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] Derrick Baxby. The diagnosis of the invasion of measles from a study of the exanthema as it appears on the buccal mucous membrane. *Reviews in Medical Virology*, 7(2):71, 1997.

[26] 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.

[27] 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.

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

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

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

[32] 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.

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

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

[35] 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.

[36] 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.

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

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

[39] D.J.D. Earn, S.A Levin, and P. Rohani. Coherence and Conservation. *Science*, 290(5495):1360–1364, 2000. doi: 10.1126/science.290.5495.1360.

[40] B.T. Grenfell, O.N. Bjørnstad, and J. Kappey. Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414(6865):716–723, 2001.

[41] 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.

[42] 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.

[43] B. Metzker. School Calendars. Technical Report EDO-EA-02-03, Educational Resources Information Center, 2002.

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[44] James M Pedersen. The history of school and summer vacation. *Journal of Inquiry and Action in Education*, 5(1):4, 2012.

[45] Joel Weiss and Robert S. Brown. *Telling Tales Over Time*, pages 23–54. SensePublishers, Rotterdam, 2013. ISBN 978-94-6209-263-1. doi: 10.1007/978-94-6209-263-1_3. URL `http://dx.doi.org/10.1007/978-94-6209-263-1_3`.

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

[47] 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.

[48] 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.

[49] 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.

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

[51] 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.

[52] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

[53] 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.

[54] 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.

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

[56] 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.

[57] 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.

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[58] 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.

[59] 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.

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

[61] 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.

[62] 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`.

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

[64] 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.

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

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

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

[68] 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.

[69] 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.

[70] 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.

[71] 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.

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplemental Information for

Evidence of cryptic incidence in childhood diseases

Christian E. Gunning, Matthew J. Ferrari, Erik Erhardt, and Helen J. Wearing, 2017.

Disease	Window Length (W, weeks)	# of Cities	# of Windows	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	1042	0.958
Pertussis	2	73	521	0.937
Pertussis	4	59	259	0.888
Pertussis	8	46	130	0.848
Pertussis	16	32	61	0.732

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

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

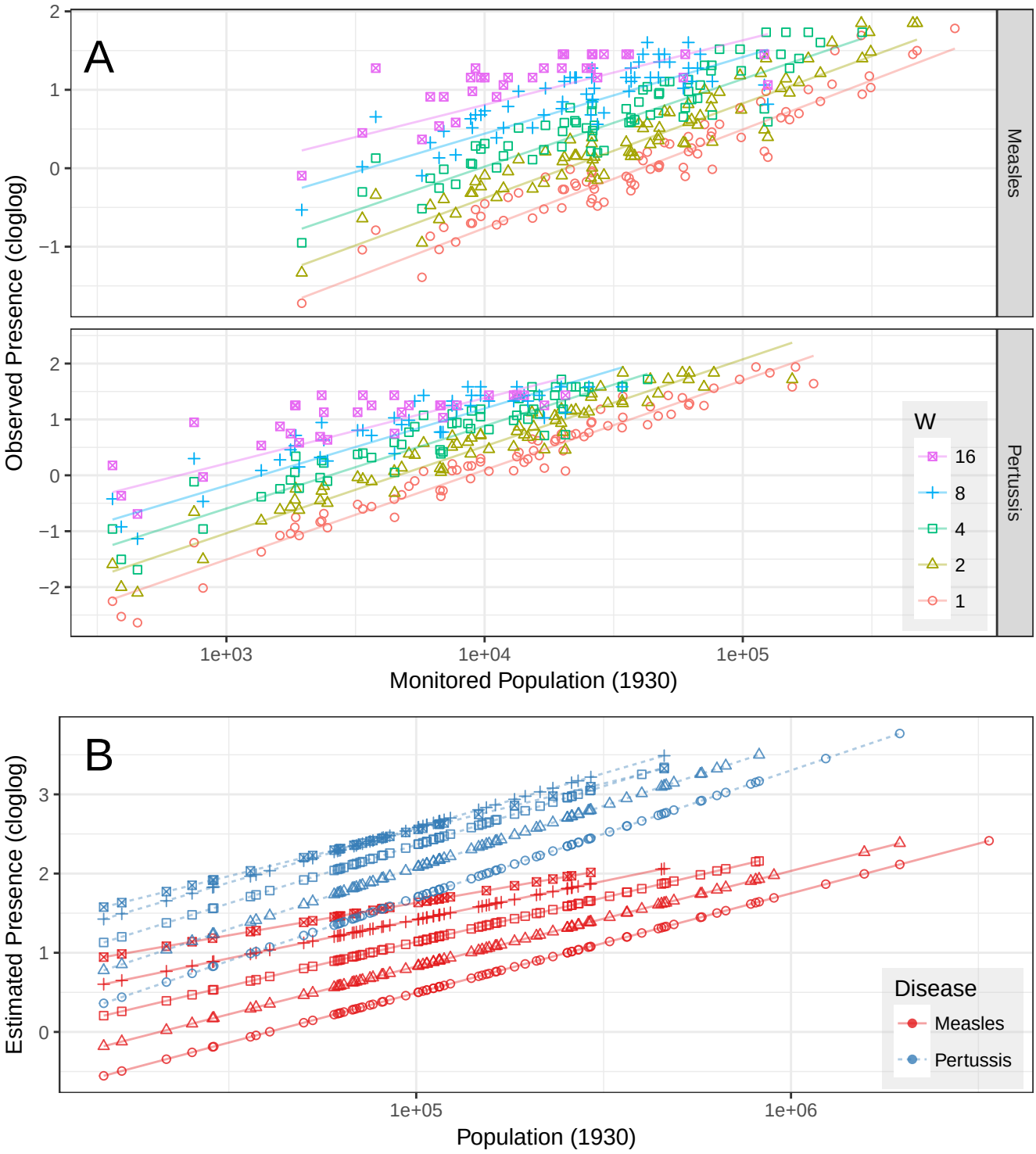

Figure S1: Binomial GLMs: **A: Observed Presence.** Response of observed presence (P_{obs}) to monitored population size (N_m) and reporting window length (W, weeks). Lines show model estimates, with a separate model for each disease, using a cloglog link (y-axis, f): $f(P_{obs}) \sim \log N_m | \log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

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

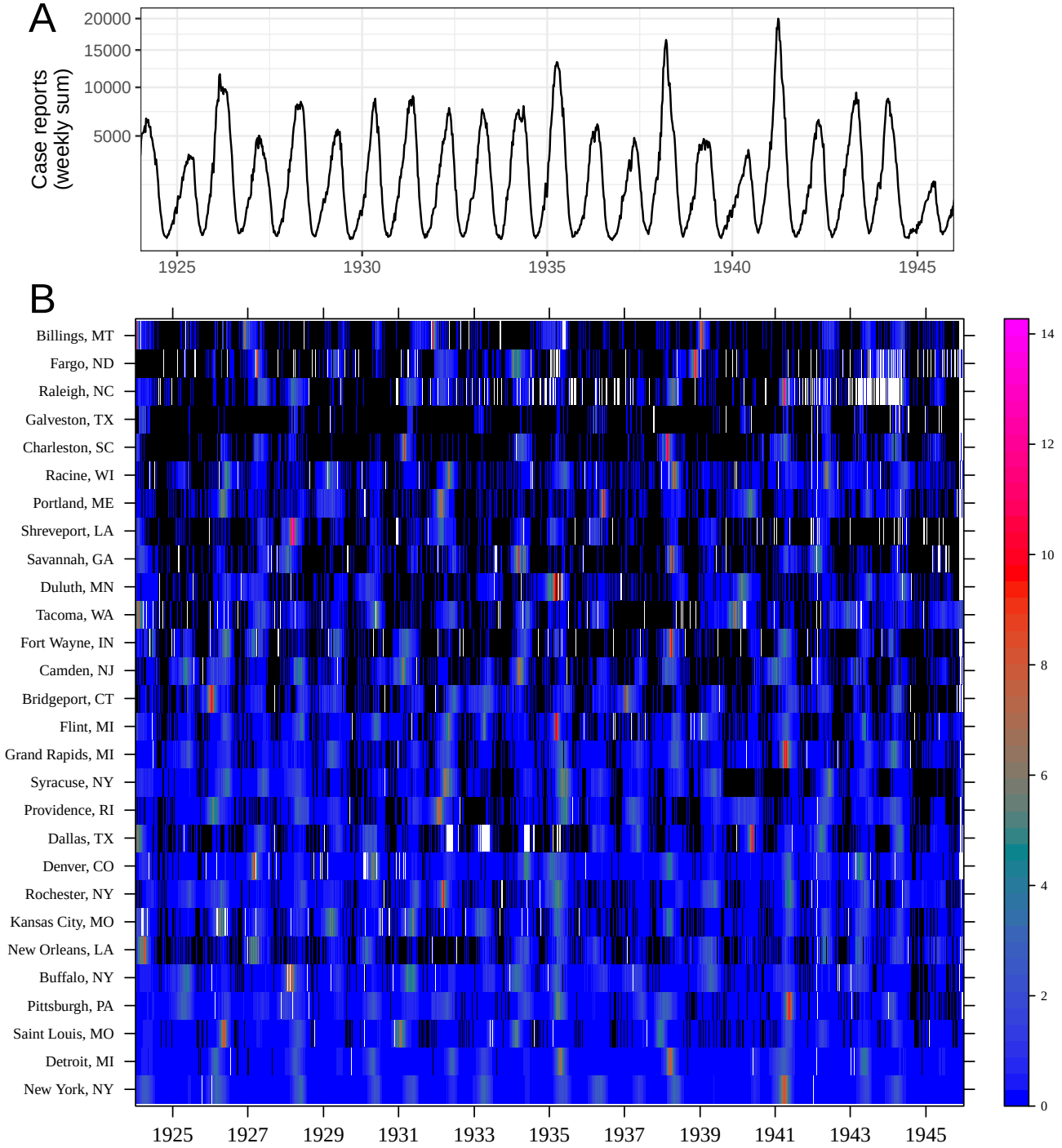

Figure S2: **Measles.** **A**. Total weekly case reports (all cities, sqrt-transformed). **B**. Weekly case reports (black = 0; white = missing). One city per row, ordered by population size, showing every third city. To facilitate comparison between cities, each city's observations are scaled by that city's temporal variance.

bioRxiv preprint doi: https://doi.org/10.1101/079194; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

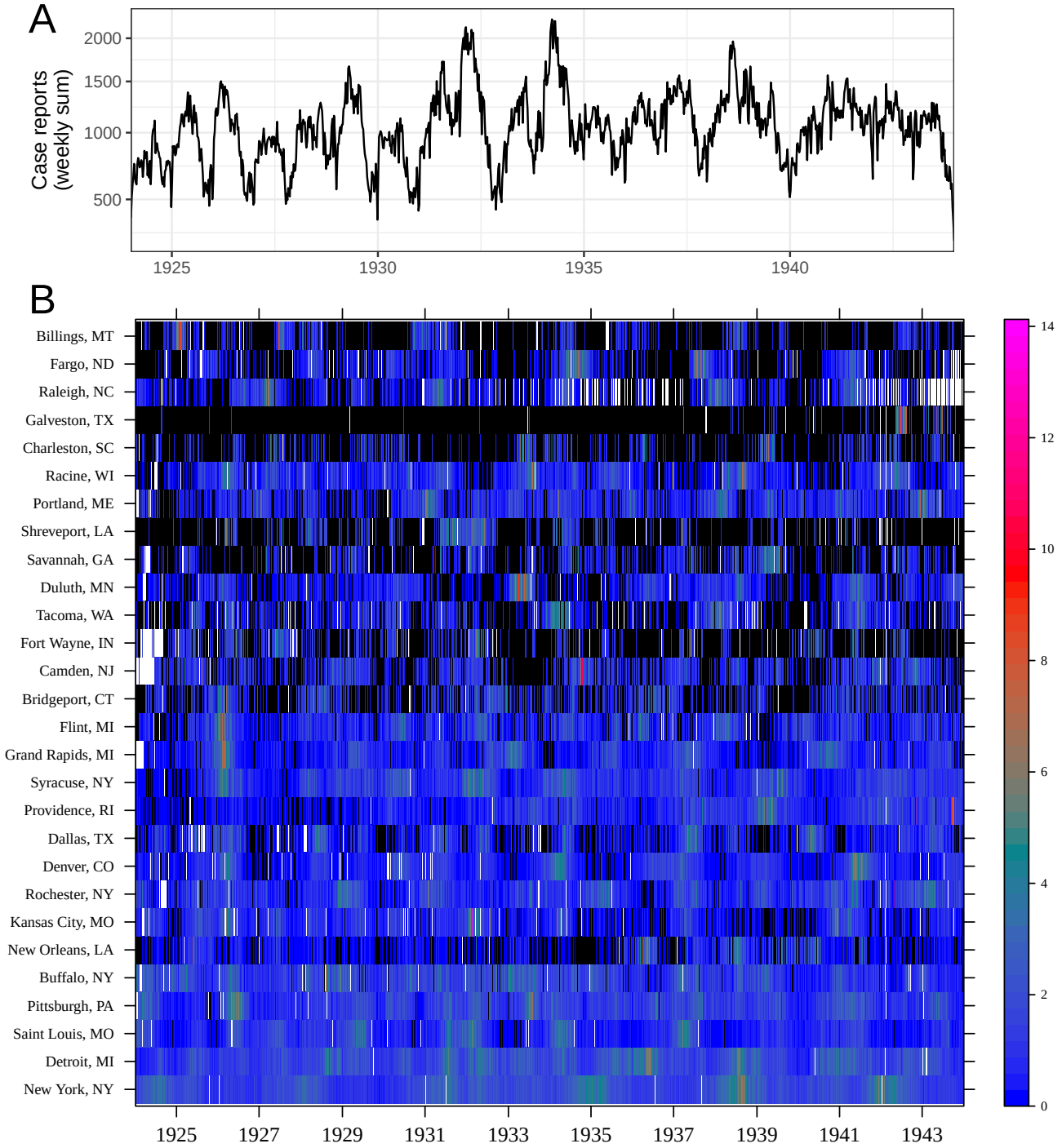

Figure S3: **Pertussis.** **A**. Total weekly case reports (all cities, sqrt-transformed). **B**. Weekly case reports (black = 0; white = missing). One city per row, ordered by population size, showing every third city. To facilitate comparison between cities, each city's observations are scaled by that city's temporal variance.