Misdiagnosis prevents accurate monitoring of transmission and burden for sub-critical pathogens: a case study of *Plasmodium knowlesi* malaria

John H. Huber

Department of Biological Sciences and Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA

Corresponding Author: huber.j.h@wustl.edu (JHH)

Keywords: malaria; misdiagnosis; surveillance

ABSTRACT

Maintaining surveillance of emerging infectious diseases presents challenges for monitoring their transmission and burden. Incomplete observation of infections and imperfect diagnosis reduce the observed sizes of transmission chains relative to their true sizes. Previous studies have examined the effect of incomplete observation on estimates of pathogen transmission and burden. However, each study assumed that, if observed, each infection was correctly diagnosed. Here, I leveraged principles from branching process theory to examine how misdiagnosis could contribute to bias in estimates of transmission and burden for emerging infectious diseases. Using the zoonotic *Plasmodium knowlesi* malaria as a case study, I found that, even when assuming complete observation of infections, the number of misdiagnosed cases within a transmission chain for every correctly diagnosed case could range from 0 (0 – 4) when R_0 was 0.1 to 86 (0 – 837) when R_0 was 0.9. Data on transmission chain sizes obtained using an imperfect diagnostic could consistently lead to underestimates of R_0, the basic reproduction number, and simulations revealed that such data on up to 1,000 observed transmission chains was not powered to detect changes in transmission. My results demonstrate that misdiagnosis may hinder effective monitoring of emerging infectious diseases and that sensitivity of diagnostics should be considered in evaluations of surveillance systems.

INTRODUCTION

For pathogens with sub-critical transmission (i.e., $R_0 < 1$), a robust surveillance system that identifies and correctly diagnoses infections is necessary to monitor changes in pathogen transmission and burden (1). Such pathogen surveillance is important both for measuring progress towards elimination of diseases with immediate public health importance, such as measles (2–4) and malaria (5), and for assessing the future threat of emerging infectious diseases (6), such as avian influenza (7), human monkeypox (1,8), and Middle East respiratory syndrome coronavirus (2,9).

Considerable work has been devoted to advance a mathematical framework that leverages the data collected by surveillance systems to obtain estimates of transmission and burden for pathogens with sub-critical dynamics (1,2,4,10,11). These studies have improved our understanding of a wide range of emerging infectious diseases and have critically evaluated the sensitivity of these estimates to the quality of data from the surveillance system. Crucially, each study modeled variation in surveillance quality through variation in the ascertainment fraction (i.e., the proportion of infections that are detected) and assumed that, once detected, all infections were correctly diagnosed. In reality, however, non-specific clinical and biological features are likely to limit the sensitivity of clinical diagnosis, particularly for emerging infectious diseases (12,13). The extent to which misdiagnosis affects estimates of transmission and burden for pathogens with sub-critical dynamics remains largely unaddressed.

The zoonotic *Plasmodium knowlesi* malaria offers a natural case study to examine the impact of misdiagnosis on estimates of transmission and burden. Endemic to Southeast Asia (14), *P. knowlesi* is a vector-borne disease with most or all infections in humans caused by spillover transmission from the long- and pig-tailed macaque reservoir (15,16). The extent of

transmission between humans is currently unknown (17). Due to morphological similarities with other *Plasmodium* spp., *P. knowlesi* is routinely misdiagnosed by light microscopy (18). A recent systematic review and meta-analysis estimated that the sensitivity of light microscopy for diagnosing *P. knowlesi* infections was less than 1% (19). This high rate of misdiagnosis greatly affects the quality of surveillance data on *P. knowlesi*, potentially biasing estimates of transmission and burden.

In this study, I aimed to evaluate the extent to which misdiagnosis of a pathogen affected the ability to monitor its change in transmission and burden. Using *P. knowlesi* as a case study, I leveraged an established framework based upon branching process theory to first quantify the potential magnitude of underestimation of pathogen burden on account of misdiagnosis. Next, I considered how underestimates of pathogen burden could lead to bias in estimates of R_0, the basic reproduction number. Finally, I quantified the degree to which misdiagnosis reduced the statistical power to detect changes in transmission from surveillance data for emerging infectious diseases, such as *P. knowlesi*.

METHODS

Branching Process Framework of Sub-Critical Transmission

To explore the effects of misdiagnosis on the monitoring of sub-critical transmission (i.e., $R_0 < 1$) of *P. knowlesi*, I extended a framework that uses branching process theory to estimate a pathogen's R_0 from its size distribution of stuttering transmission chains. Here, I followed Blumberg and Lloyd-Smith (1,11) and defined a transmission chain as a primary infection (i.e., a spillover infection from a zoonotic reservoir) and all secondary infections arising from that primary infection through at least one generation of pathogen transmission.

Assuming that the number of secondary infections caused through one generation of pathogen transmission followed a negative binomial distribution with mean R_0 and dispersion parameter κ, I used the branching framework to calculate summary statistics of the transmission chains. Specifically, I solved for the probability that a transmission chain was truly of size j infections, r_j, and the mean size of transmission chains, μ.

Following Blumberg and Lloyd-Smith (11), I considered two models of observation of infections: (i) independent observation and (ii) size-dependent observation. The model of independent observation assumes that each infection is subject to an independent probability of observation and correct diagnosis, p_{ind}, that is equal to the product of the observation probability, p_{det}, and the diagnostic sensitivity, se. By comparison, the model of size-dependent observation assumes that observation of transmission chains occurs through sentinel infections. Each infection within a transmission chain is a sentinel infection with probability, p_{sent}, and, if there is at least sentinel infection within the transmission chain, then all infections within the transmission chain are observed. Diagnosis of each infection occurs independently and is subject to sensitivity, se.

I then computed the mean observed transmission chain size, μ^*, as a function of the transmission parameters (R_0 and κ), the observation model (p_{det} or p_{sent}), and the diagnostic accuracy (se). This allowed me to relate the distribution of observed transmission chain sizes to the distribution of true transmission chain sizes and quantify bias in the maximum-likelihood estimates of transmission, $\hat{R}_0 = 1 - \frac{1}{\mu^*}$. If all infections are observed and correctly diagnosed, then $\mu^* = \mu$ and thus $\hat{R}_0 = R_0$. Violations of this assumption, either through incomplete observation or misdiagnosis, introduce bias into transmission estimates. A full description of the branching process framework can be found in the Supplement.

Analyses

Quantifying the Bounds of Total Burden

To first demonstrate how misdiagnosis, in addition to incomplete observation, may lead to an underestimate of *P. knowlesi* burden, I computed the probability distribution of the true size of a transmission chain conditional upon the observed size of a transmission chain. That is, given that I observed a transmission chain of size $\hat{j}$, the probability that the transmission chain is truly of size j is equal to

$$\Pr(j|\hat{j}) = \frac{\Pr(\hat{j}|j)\Pr(j)}{\Pr(\hat{j})}. \quad (1)$$

In eq. (1), Pr (j) is the probability that a transmission chain is of size j, r_j, computed using eq. (S1), and Pr $(\hat{j})$ is the probability that a transmission chain is of observed size j, s'_j, computed using eq. (S2) for the model of independent observation and using the numerator of eq. (S7) for the model of size-dependent observation. For the model of independent observation, the probability of observing a transmission chain of size $\hat{j}$ given that the transmission chain is truly of size j is

$$\Pr(\hat{j}|j) = \binom{j}{\hat{j}} \cdot p_{ind}^{\hat{j}} \cdot (1 - p_{ind})^{j-\hat{j}}, \quad (2)$$

where p_{ind} is equal to the product of the probability of detection, p_{det}, and the sensitivity of diagnosis, se. By comparison, for the model of size-dependent observation, this quantity is computed as

$$\Pr(\hat{j}|j) = \left(1 - (1 - p_{sent})^{j}\right) \cdot \binom{j}{\hat{j}} \cdot se^{\hat{j}} \cdot (1 - se)^{j-\hat{j}}. \quad (3)$$

Substituting the respective terms into eq. (1), for the model of independent observation, I computed the probability that a transmission chain is of true size j given that it is observed to be of size $\hat{j}$ as

$$\Pr(j|\hat{j}) = \frac{\binom{j}{\hat{j}} \cdot p_{ind}^{\hat{j}} \cdot (1 - p_{ind})^{j-\hat{j}} r_j}{s'_j}. \quad (4)$$

For the model of size-dependent observation, I computed this quantity as

$$\Pr(j|\hat{j}) = \frac{\left(1 - (1 - p_{sent})^{j}\right) \cdot \binom{j}{\hat{j}} \cdot se^{\hat{j}} \cdot (1 - se)^{j-\hat{j}} r_j}{s'_j}. \quad (5)$$

I used eqs. (4-5) to compute the expected true transmission chain sizes given observed transmission chains of one, two, or three cases while varying the probability of observation, p_{det} or p_{sent}, from 0.1 to 1.0 in increments of 0.1. I sampled the sensitivity of the diagnostic method from the posterior estimate of sensitivity of light microscopy for *P. knowlesi* with mean equal to 1.19×10^{-3} (19). For each combination of observed chain size and probability of observation, I

138 calculated the expected true transmission chain size, assuming that the true value of R_0 was
139 equal to 0.1, 0.5, or 0.9. These values of R_0 represent low, medium, and high values of sub-
140 critical transmission and fall within the plausible range of human-to-human transmission of *P.*
141 *knowlesi* (20). The dispersion parameter κ was assumed to be 0.1 in all scenarios, though a
142 supplementary analysis was performed where $\kappa \to \infty$.
143
144 ***Effect of Misdiagnosis on Estimates of Transmission***
145 On account of incomplete observation and misdiagnosis, the observed burden of *P. knowlesi* may
146 not reflect the true burden. It follows that the mean observed transmission chain size will not
147 equal the true mean transmission chain size, biasing our estimates of R_0. To explore the extent of
148 this bias in scenarios where R_0 was equal to 0.1, 0.5, or 0.9, I calculated the mean observed
149 transmission chain sizes while varying the probabilities of observation, p_{det} or p_{sent}, from 0.1 to
150 1.0 in increments of 0.1 and while using posterior samples of sensitivity of light microscopy for
151 *P. knowlesi* (19). I then compared the maximum-likelihood estimates of $\hat{R}_0$ to the true values of
152 R_0 for both the models of independent observation and size-dependent observation. In all
153 scenarios, I assumed that the dispersion parameter κ was 0.1, and a supplementary analysis was
154 performed were $\kappa \to \infty$.
155
156 ***Effect of Misdiagnosis on Statistical Power to Detect Changes in Transmission***
157 Bias in R_0 estimates on account of misdiagnosis could reduce the statistical power to detect
158 changes in R_0 over time using data on the size of transmission chains. To measure statistical
159 power as a function of the number of observed transmission chains, I followed an approach taken
160 by Blumberg *et al*. (2). I assumed that R_0 was historically equal to 0.1 and then increased to

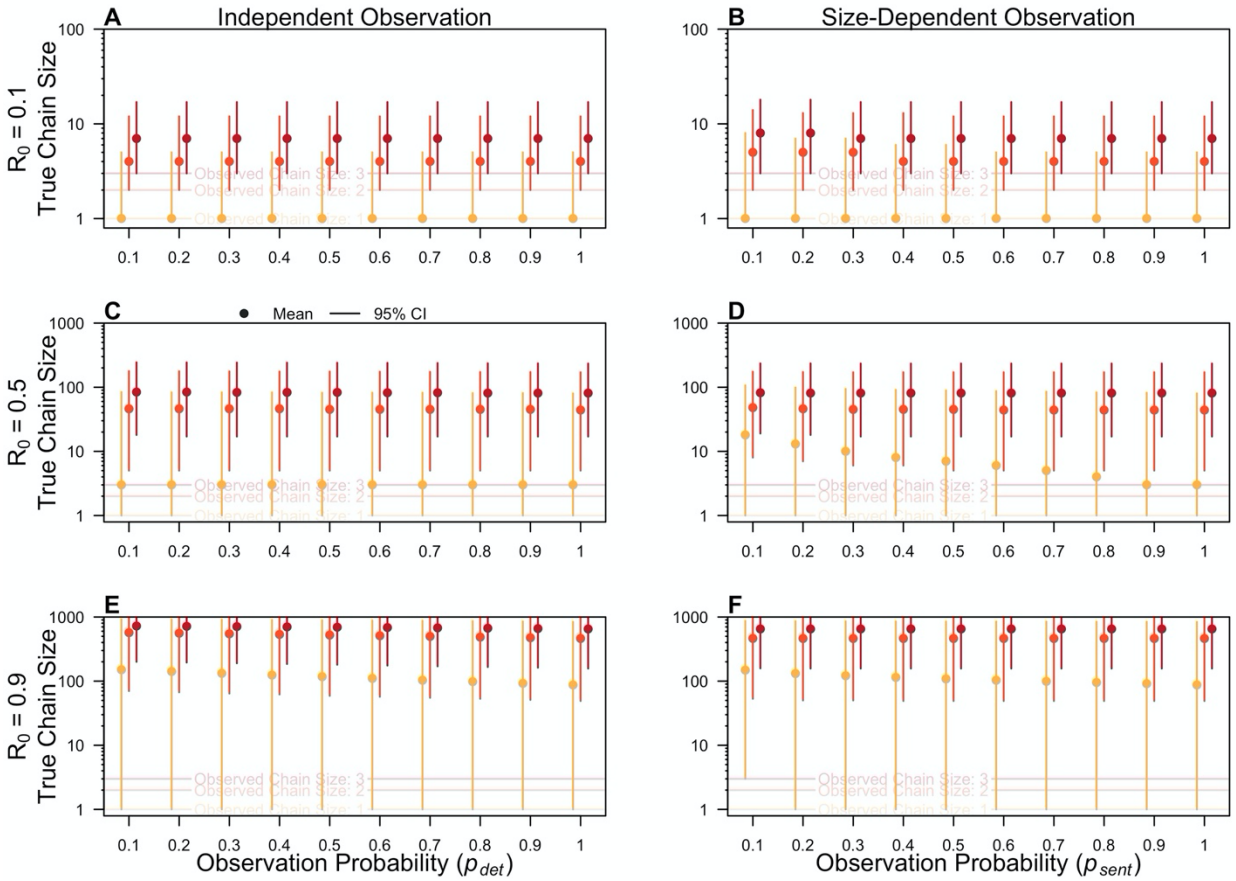
161 $R_0 + \Delta R_0$, where ΔR_0 was set to 0.1, 0.5, or 0.9. I then simulated N observed transmission chains
162 and estimated $\hat{R}_0$ while varying N from 1 to 1,000. I then compared the model in which R_0 was
163 estimated to have changed to the null hypothesis that there was no change in transmission (i.e.,
164 $\Delta R_0 = 0$) using the Akaike Information Criterion (AIC) (21). For each number of observed
165 transmission chains N, I repeated this procedure 1,000 times and computed statistical power as
166 the proportion of simulations in which I detected a change in transmission on the basis of AIC.
167 To measure the minimum effect of misdiagnosis on statistical power, I set p_{det} and p_{sent} equal
168 to 1. Because all infections are observed under this assumption and diagnosis is performed
169 independently across infections in both models, the models of independent and size-dependent
170 observation yield identical results. In all scenarios, I assumed that the dispersion parameter κ
171 was 0.1, and a supplementary analysis was performed were $\kappa \to \infty$.
172

173 **RESULTS**

174 Assuming complete observation of infections (i.e., p_{det} and p_{sent} equal to 1), misdiagnosis of *P.*
175 *knowlesi* infections would underestimate the true *P. knowlesi* burden, with the magnitude of this
176 effect depending upon R_0 (Fig. 1). For a scenario in which R_0 was 0.1, the expected true size of a
177 transmission chain is one infection (95% CI: 1 – 5) if the observed size is one case, four
178 infections (2 – 12) if the observed size is two cases, and seven infections (3 – 17) if the observed
179 size is three cases. Under an alternative scenario in which R_0 was 0.9, the expected true size of
180 the transmission chains increased to 87 (1 – 838), 461 (49 – 965), and 650 (157 – 983)
181 infections, respectively.

182 The effect of incomplete observation (i.e., p_{det} or $p_{sent} < 1$) on the expected burden was
183 most apparent at an intermediate R_0 of 0.5 and with the model of size-dependent observation

(Fig. 1D). Under this scenario, given a transmission chain of size one, the expected true transmission chain was 18 infections (1 – 107) if p_{sent} was equal to 0.1, compared to 3 infections (1 – 80) if p_{sent} was equal to 1. In all other scenarios, the expected burden did not change significantly with p_{det} or p_{sent}. This occurred because, even with complete observation (i.e., p_{det} and p_{sent} equal to 1), 99.881% of *P. knowlesi* cases were expected to be misdiagnosed, given a sensitivity of 0.119%. Therefore, irrespective of the observation probability, only a subset of true transmission chain sizes is consistent with the sizes of the observed transmission chains, given this high percentage of false negatives. If I instead assumed perfect sensitivity of the method, I observed a greater effect of the observation probability on the expected burden (Fig. S1).



196 ***Figure 1. Effect of misdiagnosis and imperfect observation on the expected burden.*** *The mean*
197 *true transmission chain size (dots) and 95% CI (segments) are shown conditional upon on an*
198 *observed transmission chain size of one (yellow), two (orange), or three (red) cases and an* R_0 *of*
199 *0.1 (A,B), 0.5 (C,D), and 0.9 (E,F). The horizontal axis is the observation probability,*
200 *representing* p_{det} *for the Model of Independent Observation (A, C, E) and* p_{sent} *for the Model of*
201 *Size-Dependent Observation (B, D, F).*
202
203 Given that misdiagnosis underestimated the burden of *P. knowlesi* in this analysis, I assessed its
204 effect on my estimates of transmission, $\hat{R}_0$. Because misdiagnosis caused the average observed
205 size of transmission chains to be less than the average true size of transmission chains, I
206 consistently underestimated $\hat{R}_0$, with the effect being more severe at lower R_0 (Fig. 2). For
207 example, with an R_0 of 0.1 and assuming perfect observation of infections (i.e., p_{det} and p_{sent}
208 equal to 1), my median $\hat{R}_0$ estimate was 1.9 x 10^{-4} (95% PPI: 2.0 x 10^{-5} – 2.9 x 10^{-3}),
209 corresponding to a 520-fold (34 – 4900) underestimate of transmission. Under an alternative
210 scenario in which R_0 was 0.9, my median $\hat{R}_0$ was 0.26 (0.046 – 0.66), corresponding to a 3.5-
211 fold (1.4 – 19.4) underestimate in transmission.
212 My estimates of transmission were sensitive to the simulated observation probability,
213 though the direction of the effect depended upon the assumed model of observation. For the
214 model of independent observation, $\hat{R}_0$ estimates increased with increasing p_{det}, because the
215 average observed size of transmission chains increased as more infections were observed (Fig. 2,
216 left column). By contrast, $\hat{R}_0$ estimates decreased with increasing p_{sent} for the model of size-
217 dependent observation (Fig. 2, right column). This counterintuitive effect can be explained by the
218 observation that, if p_{sent} is low, larger transmission chains have a greater probability that at least

one infection is a sentinel infection. This causes a bias in the size of the transmission chains that are observed at low values of p_{sent}, increasing the mean observed transmission chain size relative to that at higher values of p_{sent} and thus inflating the $\hat{R}_0$ estimate.

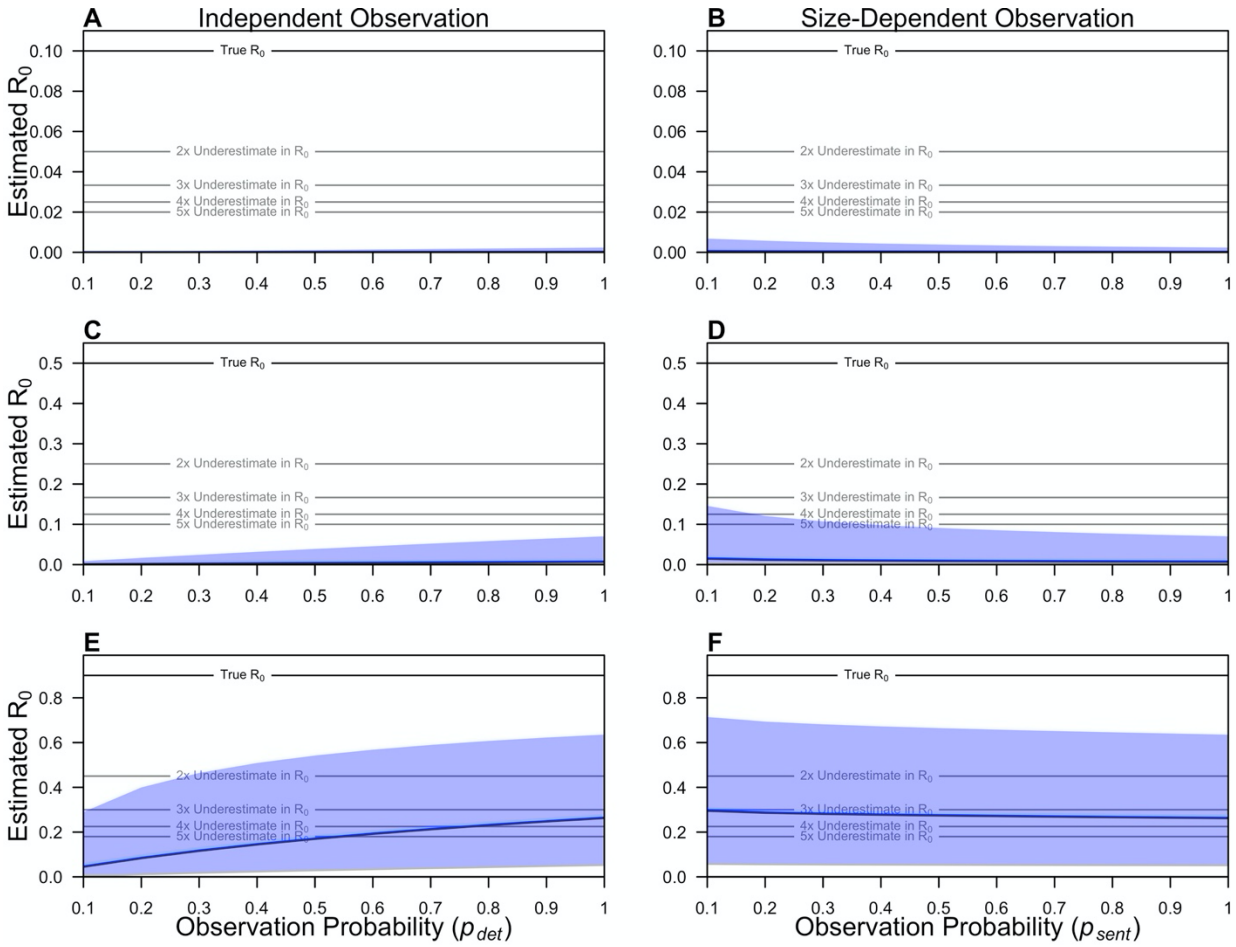


Figure 2. Effect of misdiagnosis and imperfect observation on estimates of transmission. *The posterior median (blue line) and 95% posterior prediction interval (blue shaded region) of maximum-likelihood estimates of R_0 are shown as a function of the observation probability. The observation probability represents p_{det} for the Model of Independent Observation (A, C, E) and p_{sent} for the Model of Size-Dependent Observation (B, D, F). The solid black denotes the true R_0 in each panel, and the grey lines denote two-to-five-fold underestimates of R_0 in each panel.*

The underestimates of burden (Fig. 1) and transmission (Fig. 2) indicated that misdiagnosis of *P. knowlesi* may affect the statistical power to detect changes in transmission based on the size of observed transmission chains. To test this, I simulated changes in transmission and measured the statistical power to detect that change. I observed that, under scenarios in which R_0 increased by 0.9, data on 1,000 observed transmission chains provided only 10.3% power using an imperfect diagnostic method, compared to 100% if using a perfect diagnostic method (Fig. 3). At smaller increases in R_0, data on observed transmission chain sizes obtained using an imperfect diagnostic method had effectively no power to detect a change in transmission.

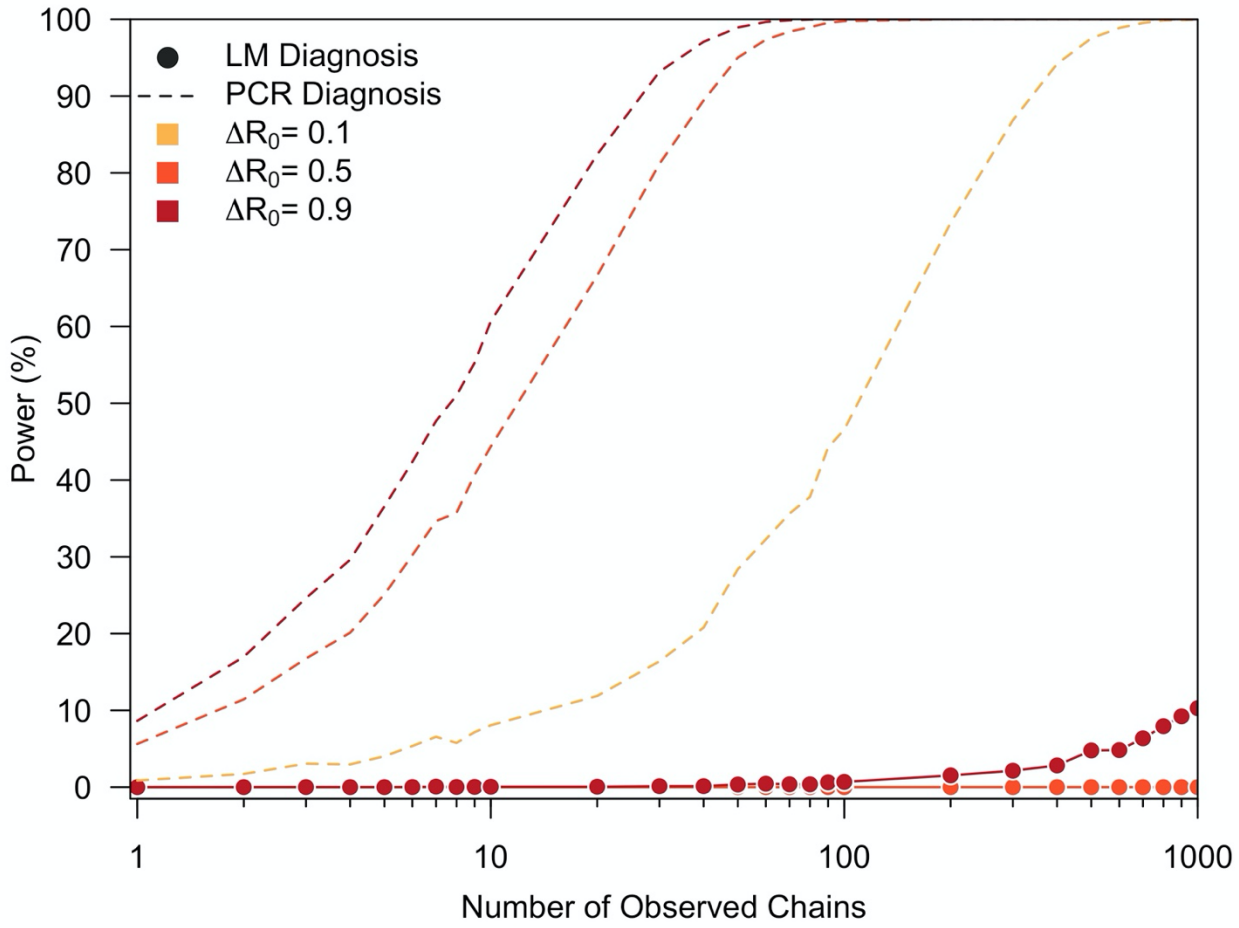


Figure 3. Effect of misdiagnosis on the statistical power to detect changes in transmission. *The statistical power (%) to detect an increase in transmission is shown as a function of the number of observed chains for a transmission increase (ΔR_0) of 0.1 (yellow), 0.5 (orange), and 0.9 (red). Solid lines and points represent an imperfect diagnostic method (i.e., LM) and the dotted lines represent perfect diagnosis (i.e., PCR).*

DISCUSSION

Obtaining accurate estimates of transmission and burden is important for monitoring the emergence of infectious diseases. Previous studies have explored the extent to which incomplete observation of infections affects the estimates of transmission for such pathogens (1,2,7,11). In

this work, I built upon these studies by considering the effect of misdiagnosis on estimates of pathogen transmission and burden. Using the zoonotic *P. knowlesi* malaria as a case study, I found that misdiagnosis—independent of incomplete observation of infections—may cause us to underestimate the transmission and burden of pathogens with sub-critical dynamics and hinders effective, prospective monitoring of changes in transmission.

My results demonstrate that, for pathogens with sub-critical transmission, misdiagnosis leads to an underestimate of overall pathogen burden. Depending upon the R_0 simulated, I found that there could be as many as 86 misdiagnosed cases, on average, for each correctly diagnosed case of *P. knowlesi*, even when assuming complete observation of infections. This effect increased under select settings when the assumption of complete observation of infections was relaxed. The underestimation of burden due to misdiagnosis has the potential to shape our epidemiological understanding of an emerging pathogen. For instance, singleton cases of a zoonotic pathogen, such as *P. knowlesi*, are commonly assessed as dead-end spillover events from the zoonotic reservoir (17). However, my simulations suggest that such singleton cases could instead represent a broad range of epidemiological outcomes, spanning dead-end spillover events to larger transmission chains.

Due to its effect on observed pathogen burden, misdiagnosis contributed a downward bias in estimates of transmission. Except for scenarios in which the true simulated R_0 was close to one, my maximum-likelihood estimates of R_0 approached zero, representing situations in which we would incorrectly conclude that human-to-human transmission of the pathogen was unlikely to be occurring. For every scenario considered, the estimate of R_0 was less than the true value, indicating that bias due to misdiagnosis exceeds the competing positive bias from incomplete observation when assuming size-dependent observation (1). For pathogens such as *P.*

275 *knowlesi*, these simulation results suggest that, in settings where misdiagnosis is common, the
276 extent of human-to-human transmission could be greater than previously thought. To date, it has
277 been believed that nearly all cases of *P. knowlesi* in humans are caused by spillover from long-
278 tailed and pig-tailed macaques, the zoonotic reservoir (17). The lack of observed human-to-
279 human transmission may be explained by multiple factors, including low parasite densities in
280 humans (16) and restricted vector habitat preference (15), and is supported by a lack of genetic
281 diversity across human *P. knowlesi* infections (22). Nevertheless, human-to-human transmission
282 of *P. knowlesi* has been demonstrated experimentally (23), and these results suggest that, if or
283 when human-to-human transmission occurs, misdiagnosis could cause us to underestimate its
284 magnitude.

285 Finally, I demonstrated that data on the sizes of transmission chains diagnosed using a
286 diagnostic with realistic sensitivity would be insufficient to monitor changes in transmission.
287 Even with 1,000 observed transmission chains, I calculated a power of only 10% to detect an
288 increase in R_0 from 0.1 to 1. This empirical power calculation assumed complete observation of
289 infections, so it represents an upper bound on the statistical power that we might expect if a
290 diagnostic with realistic sensitivity was used. Therefore, more sensitive diagnostics, such as
291 polymerase chain reaction, may be needed to detect changes in transmission that could result
292 from pathogen evolution (24), among other factors (25–27).

293 This analysis is subject to a number of limitations. First, the conclusions that I reached
294 were based upon simulated data only. I used simulations representative of *P. knowlesi* to
295 illustrate possible outcomes that may occur due to misdiagnosis (19,20), yet I lacked empirical
296 data on the distribution of transmission chain sizes for *P. knowlesi*. As such, this analysis is not
297 estimating the true extent of human-to-human transmission of *P. knowlesi*. Second, methods

298 exist to account for incomplete observation of infections in estimates of R_0 (11). However, as
299 noted by Blumberg *et al*. (2), it is challenging to estimate the proportion of infections that are
300 captured by the surveillance system. Consequently, these calculations were conditioned upon the
301 assumption of complete observation and perfect diagnoses, so violations therein should be
302 interpreted as the upper bound on the bias that would likely be observed. Finally, I considered a
303 single pathogen in isolation, though misdiagnosis is commonly due to co-circulation of related
304 pathogens (12,13,18). Accounting for the upward bias due to false-positive diagnoses from other
305 pathogens and exploring the magnitude of this effect across epidemiological settings could be
306 important directions for future work.
307

308 **ACKNOWLEDGMENTS**

309 I acknowledge support from a Graduate Research Fellowship from the National Science
310 Foundation, a Richard and Peggy Notebaert Premier Fellowship from the University of Notre
311 Dame, and the National Institute of General Medical Sciences (grant number 1R35GM143029-
312 01 to Alex Perkins). The funders had no role in the study design, in the collection, analysis, and
313 interpretation of data, in the writing of the report, or in the design to submit the article for
314 publication. I thank Alex Perkins for helpful comments on this manuscript.
315

316 **COMPETING INTERESTS**

317 I have no competing interests to declare.
318

319 **SUPPLEMENT**

320 **Methods**

Mean Transmission Chain Size

Complete Observation and Correct Diagnosis

For a pathogen with sub-critical transmission dynamics, I modeled the number of offspring caused by a single infection through one generation of transmission as a negative binomial distribution with mean R_0 and dispersion parameter κ (1,11). Therefore, it follows that transmission chains of size j occur with probability,

$$r_j = \frac{\Gamma(\kappa j + j - 1)}{\Gamma(\kappa j)\Gamma(j+1)} \frac{\left({R_0}/{\kappa}\right)^{j-1}}{\left(1 + \left({R_0}/{\kappa}\right)\right)^{\kappa j + j - 1}}. \quad (S1)$$

In eq. (S1), $\Gamma(\cdot)$ is the gamma function. Because $R_0 < 1$, the mean transmission chain size μ can be calculated as the mean of a geometric series with common ratio R_0 and is equal to $\frac{1}{1-R_0}$.

Incomplete Observation and Imperfect Diagnosis

In the case of *P. knowlesi* and many other pathogens, the size of transmission chains that are identified by a surveillance system will be affected by two factors. First, infections in a transmission chain may not present within the health system, due to a lack of symptoms or access to treatment. Second, infections in the transmission chain that do present within the health system may be misdiagnosed and thus inaccurately recorded within the surveillance system. Both factors will make the observed transmission chain appear smaller in size than the true transmission chain. Previous work by Blumberg and Lloyd-Smith (1) has quantified the effect of two models of incomplete observation on estimates of transmission and burden. Here, I build

upon this work by integrating the effect of (mis)diagnosis of infections that occurs secondary to the observation of infections within the health system.

Model of Independent Observation

The first model of incomplete observation and diagnosis assumes that each individual is subject to an independent probability p_{ind} equal to the product of observation probability, p_{det}, and the sensitivity of the diagnostic method, se. Therefore, the probability that we observe and correctly diagnose j cases from a transmission chain is equal to

$$s_j' = \sum_{k=j}^{\infty} r_k \cdot \binom{k}{j} \cdot p_{ind}^{j} \cdot (1 - p_{ind})^{k-j}, \quad (S2)$$

where r_k is the probability that a transmission chain is of true size k, calculated using eq. (S1). The probability that a transmission chain is of observed size j is equal to the normalized probability of s_j', computed as

$$r_j' = \frac{s_j'}{1 - s_0'}. \quad (S3)$$

In eq. (S3), s_0' is the probability that a transmission chain is neither observed nor correctly diagnosed. Due to incomplete observation and misdiagnosis, the probability that a transmission chain is of observed size j, r_j', is not equal to the probability that a transmission chain is of true size j, r_j. Finally, because each infection within the transmission chain is subject to an

independent probability of observation and correct diagnosis, the probability p_{obs} that a randomly sampled infection is observed and correctly diagnosed is equal to p_{ind}.

Model of Size-Dependent Observation

The alternative model assumes that transmission chains are observed through a sentinel infection, such that, if the sentinel infection is observed, then all other infections in the transmission chain will be observed. Incorporating the effect of imperfect diagnosis, the probability that we do not observe a transmission chain of size j is equal to

$$v_j = (1 - p_{sent})^j + \left(1 - (1 - p_{sent})^j\right)(1 - se)^j. \quad (S4)$$

In eq. (S4), the first term in the summation is the probability that none of the j infections of the transmission chain are a sentinel infection, and the second term in the summation is the product of the probability that at least one infection is a sentinel infection (i.e., the probability that we observe the transmission chain) and the probability that all j infections are misdiagnosed. Using eq. (S4), I calculated the probability that a transmission chain is neither observed nor correctly diagnosed as

$$s_0' = \sum_{k=1}^{\infty} r_k \cdot v_k. \quad (S5)$$

The probability that a transmission chain is of observed size j is then equal to

$$r_j' = \frac{\sum_{k=j}^{\infty} r_k \cdot (1-(1-p_{sent})^k) \cdot \binom{k}{j} \cdot se^j \cdot (1-se)^{k-j}}{1-s_0'}, \quad (S6)$$

and the probability that a randomly chosen infection is observed and correctly diagnosed is equal to

$$p_{obs} = \frac{\sum_{j=1}^{\infty} j \cdot r_j \cdot \left(1-(1-p_{sent})^j\right) \cdot se}{\mu}. \quad (S7)$$

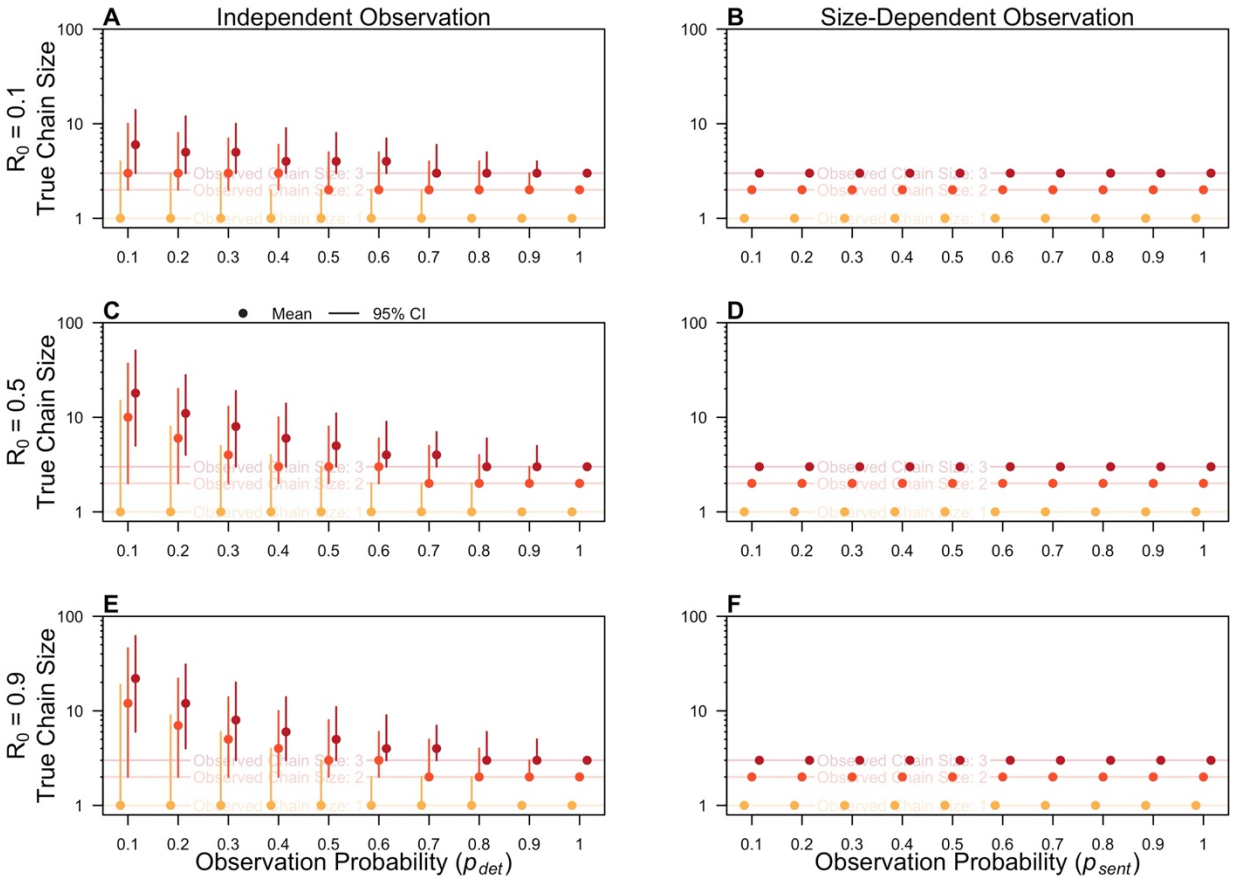
Eq. (S7) accounts for the probability that non-sentinel infections are detected, a quantity that increases as a function of the transmission chain size.

<u>Mean Transmission Chain Size</u>

For both the model of independent observation and the model of size-dependent observation, the mean observed size of transmission chains is calculated as

$$\mu^* = \sum_{j=1}^{\infty} j \cdot r_j' = \frac{p_{obs} \cdot \mu}{1-s_0'}. \quad (S8)$$

Results



402

403 ***Figure S1. Effect of misdiagnosis and imperfect observation on the expected true pathogen***

404 ***burden assuming perfect diagnosis.*** *The mean true transmission chain size (dots) and 95% CI*

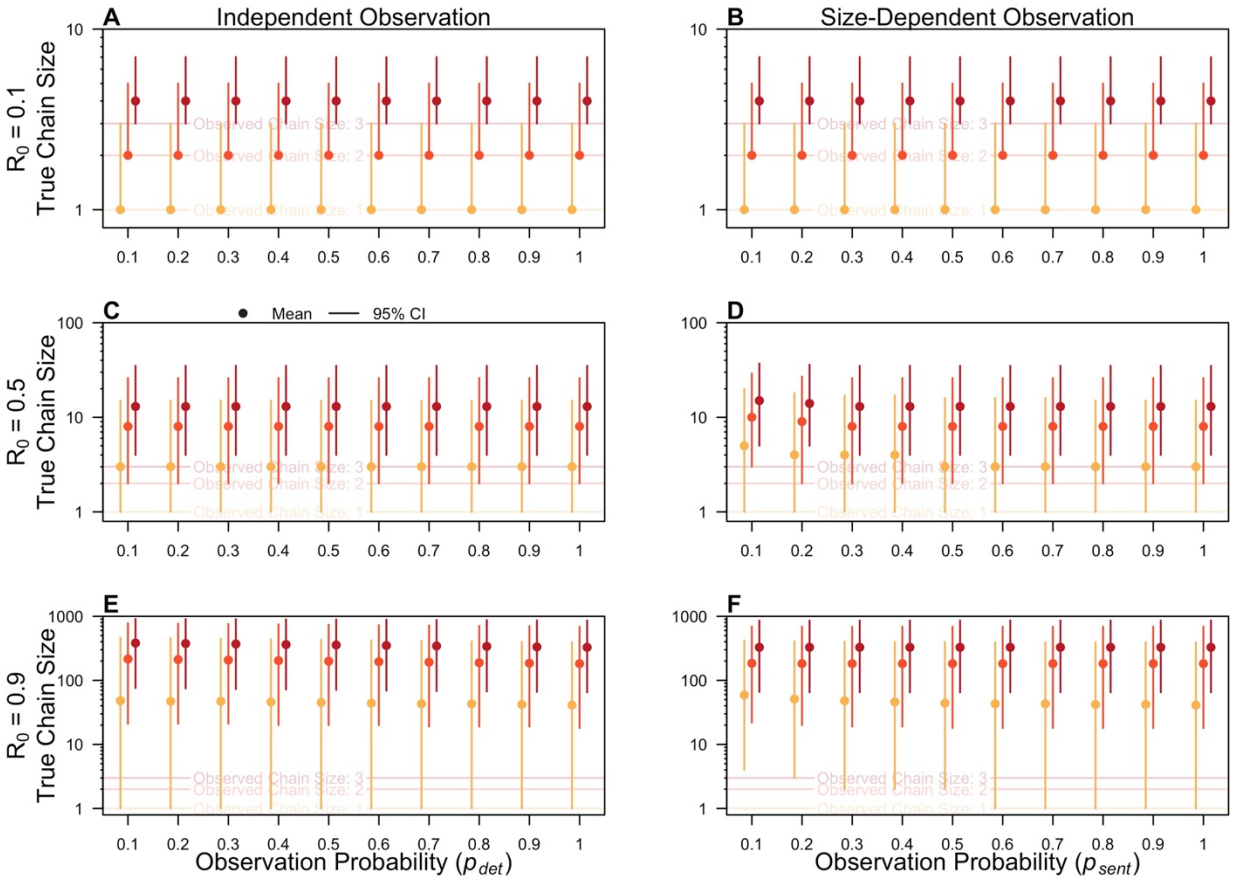
405 *(segments) are shown conditional upon on an observed transmission chain size of one (yellow),*

406 *two (orange), or three (red) cases and an* R_0 *of 0.1 (A,B), 0.5 (C,D), and 0.9 (E,F). The*

407 *horizontal axis is the observation probability, representing* p_{det} *for the Model of Independent*

408 *Observation (A, C, E) and* p_{sent} *for the Model of Size-Dependent Observation (B, D, F).*

409



410

411 ***Figure S2. Effect of misdiagnosis and imperfect observation on the expected true pathogen***

412 ***burden when κ → ∞.*** *The mean true transmission chain size (dots) and 95% CI (segments) are*

413 *shown conditional upon on an observed transmission chain size of one (yellow), two (orange), or*

414 *three (red) cases and an R_0 of 0.1 (A,B), 0.5 (C,D), and 0.9 (E,F). The horizontal axis is the*

415 *observation probability, representing p_{det} for the Model of Independent Observation (A, C, E)*

416 *and p_{sent} for the Model of Size-Dependent Observation (B, D, F).*

417

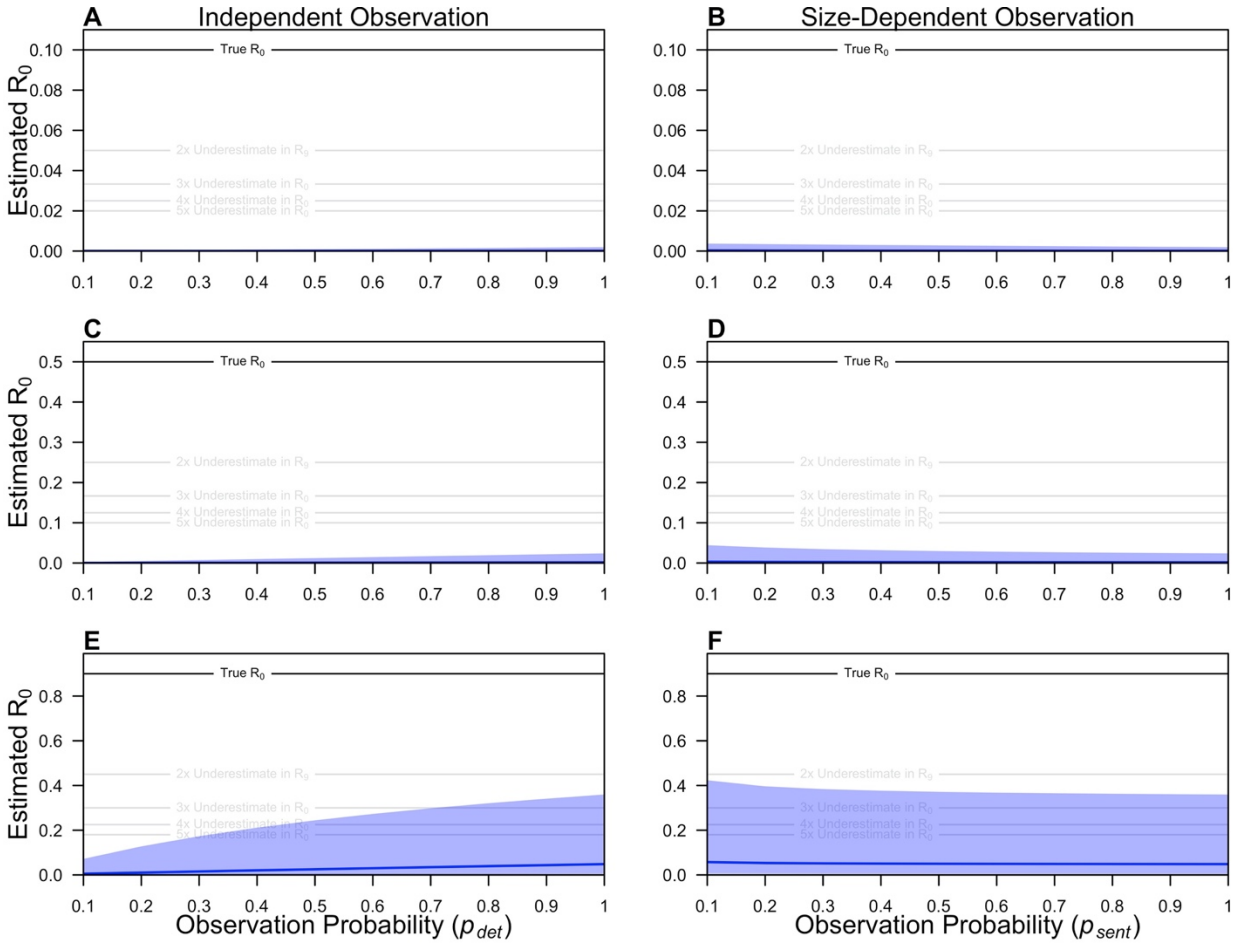


Figure S3. Effect of misdiagnosis and imperfect observation on estimates of transmission when κ → ∞. *The posterior median (blue line) and 95% posterior prediction interval (blue shaded region) of maximum-level estimates of R_0 are shown as a function of the observation probability. The observation probability represents p_{det} for the Model of Independent Observation (A, C, E) and p_{sent} for the Model of Size-Dependent Observation (B, D, F). The solid black denotes the true R_0 in each panel, and the grey lines denote two-to-five-fold underestimates of R_0 in each panel.*

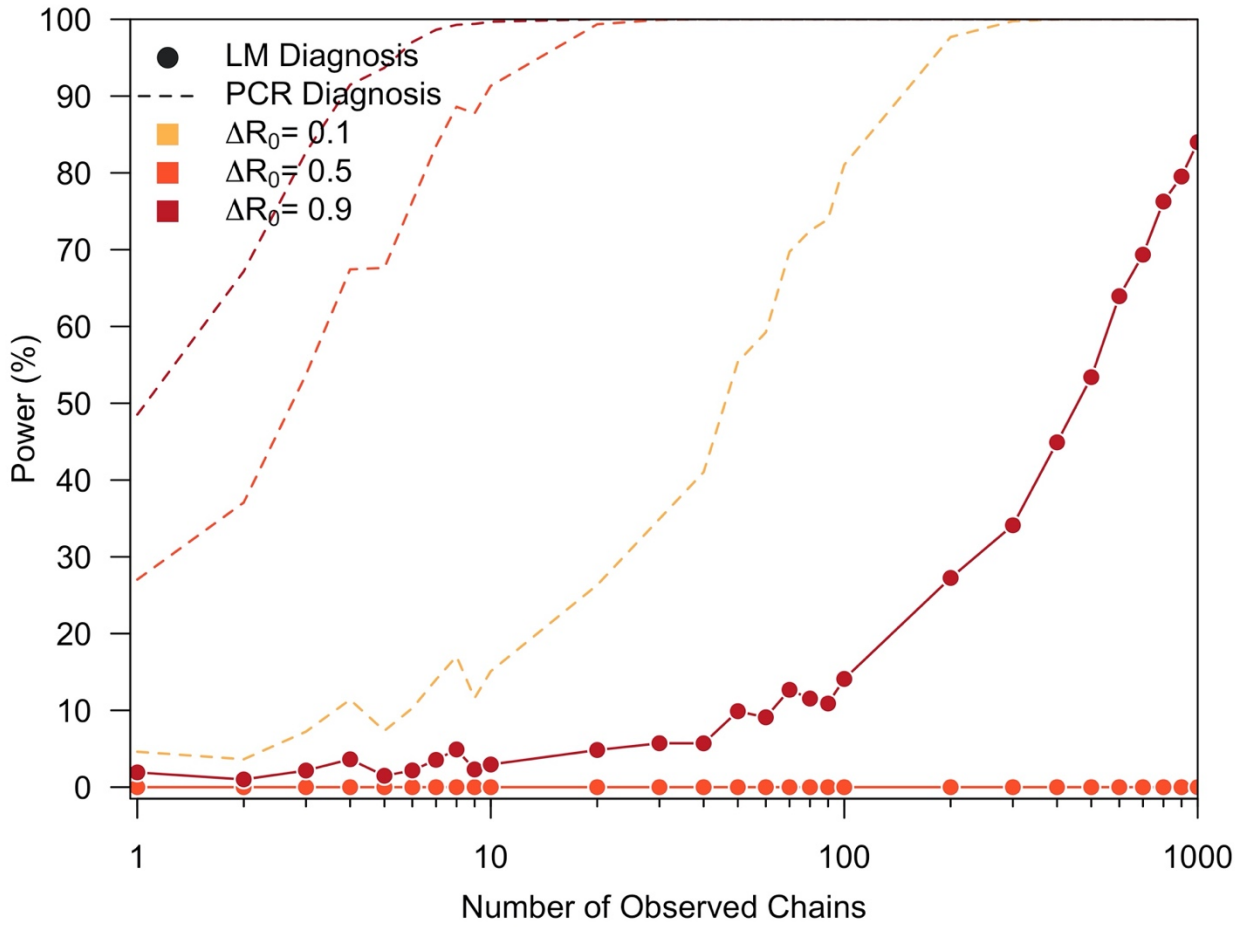


Figure S4. Effect of misdiagnosis on the statistical power to detect changes in transmission when κ → ∞. *The statistical power (%) to detect an increase in transmission is shown as a function of the number of observed chains for a transmission increase (ΔR_0) of 0.1 (yellow), 0.5 (orange), and 0.9 (red). Solid lines and points represent an imperfect diagnostic method (i.e., LM) and the dotted lines represent perfect diagnosis (i.e., PCR).*

REFERENCES

1. Blumberg S, Lloyd-Smith JO. Inference of R(0) and transmission heterogeneity from the size distribution of stuttering chains. PLoS Comput Biol. 2013;9(5):e1002993.

2. Blumberg S, Funk S, Pulliam JRC. Detecting Differential Transmissibilities That Affect the Size of Self-Limited Outbreaks. Wilke CO, editor. PLoS Pathog. 2014 Oct 30;10(10):e1004452.

3. Blumberg S, Worden L, Enanoria W, Ackley S, Deiner M, Liu F, et al. Assessing Measles Transmission in the United States Following a Large Outbreak in California. PLoS Curr [Internet]. 2015 [cited 2021 Jul 27]; Available from: https://currents.plos.org/outbreaks/article/assessing-measles-transmission-in-the-united-states-following-a-large-outbreak-in-california/

4. Farrington CP. Branching process models for surveillance of infectious diseases controlled by mass vaccination. Biostatistics. 2003 Apr 1;4(2):279–95.

5. Churcher TS, Cohen JM, Novotny J, Ntshalintshali N, Kunene S, Cauchemez S. Measuring the path toward malaria elimination. Science. 2014 Jun 13;344(6189):1230–2.

6. Plowright RK, Parrish CR, McCallum H, Hudson PJ, Ko AI, Graham AL, et al. Pathways to zoonotic spillover. Nat Rev Microbiol. 2017 Aug;15(8):502–10.

7. Ferguson NM. PUBLIC HEALTH: Enhanced: Public Health Risk from the Avian H5N1 Influenza Epidemic. Science. 2004 May 14;304(5673):968–9.

8. Ambrose MR, Kucharski AJ, Formenty P, Muyembe-Tamfum J-J, Rimoin AW, Lloyd-Smith JO. Quantifying transmission of emerging zoonoses: Using mathematical models to maximize the value of surveillance data [Internet]. Epidemiology; 2019 Jun [cited 2021 Jul 27]. Available from: http://biorxiv.org/lookup/doi/10.1101/677021

9. Cauchemez S, Nouvellet P, Cori A, Jombart T, Garske T, Clapham H, et al. Unraveling the drivers of MERS-CoV transmission. Proc Natl Acad Sci USA. 2016 Aug 9;113(32):9081–6.

10. De Serres G, Gay NJ, Farrington CP. Epidemiology of Transmissible Diseases after Elimination. American Journal of Epidemiology. 2000 Jun 1;151(11):1039–48.

11. Blumberg S, Lloyd-Smith JO. Comparing methods for estimating R0 from the size distribution of subcritical transmission chains. Epidemics. 2013 Sep;5(3):131–45.

12. Oidtman RJ, España G, Perkins TA. Co-circulation and misdiagnosis led to underestimation of the 2015–2017 Zika epidemic in the Americas. Lacerda MVG, editor. PLoS Negl Trop Dis. 2021 Mar 1;15(3):e0009208.

13. Glennon EE, Jephcott FL, Oti A, Carlson CJ, Bustos Carillo FA, Hranac CR, et al. Syndromic detectability of haemorrhagic fever outbreaks [Internet]. Epidemiology; 2020 Mar [cited 2021 Jul 27]. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.03.28.20019463

14. Shearer FM, Huang Z, Weiss DJ, Wiebe A, Gibson HS, Battle KE, et al. Estimating Geographical Variation in the Risk of Zoonotic Plasmodium knowlesi Infection in Countries

Eliminating Malaria. Churcher TS, editor. PLoS Negl Trop Dis. 2016 Aug 5;10(8):e0004915.

15. Collins WE. *Plasmodium knowlesi* : A Malaria Parasite of Monkeys and Humans. Annu Rev Entomol. 2012 Jan 7;57(1):107–21.

16. Kantele A, Jokiranta TS. Review of cases with the emerging fifth human malaria parasite, Plasmodium knowlesi. Clin Infect Dis. 2011 Jun;52(11):1356–62.

17. Feachem RGA, Chen I, Akbari O, Bertozzi-Villa A, Bhatt S, Binka F, et al. Malaria eradication within a generation: ambitious, achievable, and necessary. The Lancet. 2019 Sep;394(10203):1056–112.

18. Barber BE, William T, Grigg MJ, Yeo TW, Anstey NM. Limitations of microscopy to differentiate Plasmodium species in a region co-endemic for Plasmodium falciparum, Plasmodium vivax and Plasmodium knowlesi. Malar J. 2013 Jan 8;12:8.

19. Huber JH, Elliott M, Koepfli C, Perkins A. The impact of emerging Plasmodium knowlesi on accurate diagnosis by light microscopy: a systematic review and modelling analysis [Internet]. Epidemiology; 2021 Sep [cited 2021 Sep 13]. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.09.08.21263294

20. Imai N, White MT, Ghani AC, Drakeley CJ. Transmission and Control of Plasmodium knowlesi: A Mathematical Modelling Study. Churcher TS, editor. PLoS Negl Trop Dis. 2014 Jul 24;8(7):e2978.

21. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr. 1974 Dec;19(6):716–23.

22. Lee K-S, Divis PCS, Zakaria SK, Matusop A, Julin RA, Conway DJ, et al. Plasmodium knowlesi: Reservoir Hosts and Tracking the Emergence in Humans and Macaques. Kazura JW, editor. PLoS Pathog. 2011 Apr 7;7(4):e1002015.

23. Alpert E, Collins WE, Jeter MH, Chin W, Contacos PG. Experimental Mosquito-Transmission of Plasmodium Knowlesi to Man and Monkey. The American Journal of Tropical Medicine and Hygiene. 1968 May 1;17(3):355–8.

24. Geoghegan JL, Senior AM, Di Giallonardo F, Holmes EC. Virological factors that increase the transmissibility of emerging human viruses. Proc Natl Acad Sci USA. 2016 Apr 12;113(15):4170–5.

25. Faust CL, McCallum HI, Bloomfield LSP, Gottdenker NL, Gillespie TR, Torney CJ, et al. Pathogen spillover during land conversion. Ostfeld R, editor. Ecol Lett. 2018 Apr;21(4):471–83.

26. Baeza A, Santos-Vega M, Dobson AP, Pascual M. The rise and fall of malaria under land-use change in frontier regions. Nat Ecol Evol. 2017 May;1(5):0108.

27. Morse SS. Factors in the Emergence of Infectious Diseases. Emerg Infect Dis. 1995 Mar;1(1):7–15.