PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Estimating Time of Infection Using Prior Serological and Individual Information Can Greatly Improve Incidence Estimation of Human and Wildlife Infections

Benny Borremans[1]*, Niel Hens[2,3], Philippe Beutels[2], Herwig Leirs[1], Jonas Reijniers[1,4]

1 Evolutionary Ecology Group, University of Antwerp, Antwerp, Belgium, **2** Centre for Health Economics Research & Modelling Infectious Diseases (CHERMID), Vaccine & Infectious Disease Institute (VAXINFECTIO), University of Antwerp, Antwerp, Belgium, **3** Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BIOSTAT), Hasselt University, Diepenbeek, Belgium, **4** Department of Engineering Management, University of Antwerp, Antwerp, Belgium

* bennyborremans@gmail.com

CrossMark
click for updates

OPEN ACCESS

Citation: Borremans B, Hens N, Beutels P, Leirs H, Reijniers J (2016) Estimating Time of Infection Using Prior Serological and Individual Information Can Greatly Improve Incidence Estimation of Human and Wildlife Infections. PLoS Comput Biol 12(5): e1004882. doi:10.1371/journal.pcbi.1004882

Editor: Marcel Salathé, Ecole Polytechnique Federale de Lausanne, SWITZERLAND

Received: August 22, 2015

Accepted: March 24, 2016

Published: May 13, 2016

Copyright: © 2016 Borremans et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data can be found in the Supporting Information file S1 Data.

Funding: This work was supported by the University of Antwerp grant number GOA BOF FFB3567, Deutsche Forschungsgemeinschaft Focus Program 1596 and the Antwerp Study Centre for Infectious Diseases (ASCID). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Diseases of humans and wildlife are typically tracked and studied through incidence, the number of new infections per time unit. Estimating incidence is not without difficulties, as asymptomatic infections, low sampling intervals and low sample sizes can introduce large estimation errors. After infection, biomarkers such as antibodies or pathogens often change predictably over time, and this temporal pattern can contain information about the time since infection that could improve incidence estimation. Antibody level and avidity have been used to estimate time since infection and to recreate incidence, but the errors on these estimates using currently existing methods are generally large. Using a semi-parametric model in a Bayesian framework, we introduce a method that allows the use of multiple sources of information (such as antibody level, pathogen presence in different organs, individual age, season) for estimating individual time since infection. When sufficient background data are available, this method can greatly improve incidence estimation, which we show using arenavirus infection in multimammate mice as a test case. The method performs well, especially compared to the situation in which seroconversion events between sampling sessions are the main data source. The possibility to implement several sources of information allows the use of data that are in many cases already available, which means that existing incidence data can be improved without the need for additional sampling efforts or laboratory assays.

Author Summary

Human and wildlife diseases can be tracked by looking at incidence, which is the number of new infections per time unit (typically day, week or month). While theoretically this would only be a matter of counting the number of newly infected individuals, in reality these data are difficult to acquire due to limited sampling possibilities and undetectable

cases. This means that a method must be used to estimate the real incidence using a limited amount of data. For many infections, the concentration and quality of antibodies changes predictably over time, which means that one could use the antibody level at any point in time to back-calculate how much time passed since the infection entered the body. Other information, such as the age of the individual, or the presence of the pathogen, can also help to estimate when an individual became infected. Improving on existing methods, we developed a method that allows the use of a wide range of information sources for estimating individual time since infection. Using arenavirus infection in mice, we show that this method works well when sufficient background data are available, and that it can greatly improve the estimation of incidence patterns.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Infection incidence (the number of new infections per time unit) is a basic epidemiological measure that describes the transmission of an infection through time. Because the exact time at which an individual acquired an infection is difficult to assess, time of symptom onset is often used as a proxy (e.g. [1]). When the time between the moment of infection and symptom onset (the incubation period) is predictable, this proxy will not bias results, but incidence estimation does become problematic with asymptomatic infection or when incubation periods vary unpredictably [2].

Another common problem for measuring incidence is the time resolution of data, as the temporal precision of incidence is directly related to that of data "sampling". Ideally, each new infection is detected and recorded immediately, but in reality this is rarely possible and new cases are often recorded at irregular intervals and a low number of time points, resulting in sub-optimal resolution incidence data [3, 4]. Even more importantly, when sampling intervals are larger than the duration of symptoms, a proportion of cases will be missed. This problem is especially common in the case of wildlife diseases, as natural populations are often sampled incompletely and at relatively large intervals [5]. In such cases, indirect measures of incidence that rely on evidence of past infection are needed.

The presence of specific antibodies indicates whether an individual has previously been infected, and the distribution of different antibody (Ab) types (e.g. IgG, IgM, IgA) can give a rough indication of how recently the individual was infected [6–9]. If individuals in a population are sampled repeatedly, a seroconversion event in between two sampling events provides further information about the time since infection. Aside from being present or not, Abs vary over time in quantity (titer) and quality (avidity). On the condition that this temporal variation is sufficiently constant and predictable within and between individuals, these antibody dynamic properties can be used for a more accurate estimation of the time since infection.

Avidity (Ab-antigen bond strength) tends to increase with time since infection, which means that it can in some cases be used to back-calculate the time since infection. But although this method is used routinely, e.g. for human cytomegalovirus [10, 11], its sensitivity is low, and it can only differentiate between "recent" or "old" (e.g. less or more than 90 days since infection for cytomegalovirus) infection events [6, 12].

Temporal dynamics of Ab levels can be another source of information about time since infection. In such cases a model must be created that describes the course of Ab levels (titers) over time since infection using known serological response data. This model is then used to back-calculate, given an Ab titer, the time since infection, which in turn can be used for incidence estimation. This has been done for pertussis [13, 14], HIV [15, 16] and Salmonella [17, 18].

While this method is promising, significant improvements are still possible in two main ways. A common, important limitation for developing good time since infection models is the lack of detailed information about individual Ab dynamics, which limits the explanatory power of such models as they must in that case be estimated using cross-sectional instead of individual data (e.g. [18]). Experimental challenge studies, in which the exact time since infection is known, would be needed to describe and model the within-individual Ab dynamics needed to calculate time since infection, but these are notoriously difficult to conduct [19]. A perhaps more feasible approach to improving time since infection models would be to make optimal use of all available sources of information on the course of infection. While changes in Ab presence/titer over time can contain much information on time since infection and are the most obvious input data, additional information is contained in parameters such as the presence/quantity of the pathogen (or of other immune response markers), individual age (e.g. for typical childhood infections, young individuals are more likely to have been infected recently than older ones) or season (e.g. for seasonal infections, individuals are more likely to have been infected recently during or short after the peak transmission season).

Here, we present a novel method that allows the integration of multiple serological biomarkers (Ab presence/absence/titer, pathogen presence/absence) as well as additional prior knowledge (e.g. age, season, capture probability) to inform a semi-parametric mixed model that back-calculates the time since infection of each individual, in a Bayesian framework. The integration of multiple sources of information ensures the optimal use of data that are often already available but not yet taken into account.

We apply this method to estimate the incidence of Morogoro virus (MORV) infection in Natal multimammate mice (*Mastomys natalensis*). This model system is used because the epidemiological and demographic parameters necessary for testing this method are well known for this infection. MORV is a member of the arenaviruses, a family of zoonotic viruses that includes viruses able to cause hemorrhagic fever in humans after acquiring infection from wild rodents (e.g. Lassa virus (LASV), Junin virus, Machupo virus) [20]. It is restricted to East-Africa, and while it does not seem to cause disease in humans it is closely related to Lassa virus which causes Lassa hemorrhagic fever in West-Africa, and with which it shares the same host species. Because both the population ecology of the rodent host *M. natalensis* and the infection ecology of MORV have been studied thoroughly (driven by the host's status as an agricultural pest species and the virus' close resemblance to LASV) [21, 22], MORV infection provides a good model system for testing the current method.

As is the case for other time since infection methods, two types of datasets are needed to estimate incidence. A first dataset, consisting of any type of data that contains information on the temporal course of infection (e.g. Ab titer dynamics in an infected individual), is used once in order to create an integrated model of individual time since infection. Once created, this model can be used to estimate incidence from cross-sectional sampling data that ideally (but not necessarily) includes repeated measures of individuals.

We use a wildlife disease model system to develop and test the method because detailed individual-level infection/antibody dynamics are available, but also to show that the method is applicable to both human and wildlife infections. Because it is usually difficult to monitor infections at a high time-resolution, this method can provide a way to improve the quality of longitudinal data without having to increase sampling efforts.

Methods

In the following, we show how different types of data (e.g. levels, presence/absence) can be used to estimate the time of infection, and as a proof of principle we apply the method to MORV transmission in the multimammate mouse *M. natalensis*. For each type of data we present a generalised method and immediately apply it to MORV, and we show how to use individual estimates of the time of infection to estimate incidence in the population. Finally, through the use of simulated MORV transmission data we investigate method performance under different conditions.

MORV Ab level dynamics and virus presence in blood and excretions (urine, feces, saliva) have been quantified previously in a challenge study, described in [23], where multimammate mice from a breeding colony were injected with cultured MORV and sampled frequently for 210 days, which is more than their average lifetime in natural conditions (Fig 1 and [23]).

Back-Calculation Model

Bayes' rule. In the following, we assume that an individual can be encountered at different times, at which it can be tested for different types of information: Ab level, pathogen presence, age, body weight, sex, etc. For each measurement type k, the experimental information for a single individual can be represented by a vector $\boldsymbol{X}^k = [x_1^k, x_2^k, \ldots, x_n^k]$, of which the different coordinates represent the responses that have been measured at times $\boldsymbol{T} = [t_1, t_2, \ldots, t_n]$ for a particular individual.

Decoding the information about the individual time of infection θ from these experimental data $\boldsymbol{X^k}$ essentially comes down to the calculation of $P(\theta|\boldsymbol{X^k}, \boldsymbol{T})$, which is the probability that, given the information $\boldsymbol{X^k}$ measured at times $\boldsymbol{T}$, the tested individual was infected at time θ. In order to calculate $P(\theta|\boldsymbol{X^k}(\boldsymbol{T}), \boldsymbol{T})$, we make use of Bayes' Rule to arrive at

$$P(\theta|\boldsymbol{X}^k, \boldsymbol{T}) = \frac{P(\boldsymbol{T})}{P(\boldsymbol{X}^k, \boldsymbol{T})} P(\boldsymbol{X}^k|\boldsymbol{T}, \theta) P(\theta|\boldsymbol{T}). \quad (1)$$

Both the numerator and denominator of the first factor are independent of θ, and consequently this fraction can be inferred from the fact that $\int P(\theta|\boldsymbol{X^k}, \boldsymbol{T})dt = 1$. Calculating the posterior probability $P(\theta|\boldsymbol{X^k}, \boldsymbol{T})$ is then reduced to the calculation of $P(\boldsymbol{X^k}|\boldsymbol{T}, \theta)$, i.e. the likelihood that a time of infection θ produces the information $\boldsymbol{X^k}$ at times $\boldsymbol{T}$, and $P(\theta|\boldsymbol{T})$, i.e. the prior for θ if we assume that the individual was encountered at times $\boldsymbol{T}$. In the following, we describe how to model $P(\boldsymbol{X^k}|\boldsymbol{T}, \theta)$ and $P(\theta|\boldsymbol{T})$ using different sources of information.

Modeling $P(X^k|T, \theta)$

The estimation of the time of infection θ can be based on different dimensions of the immune response that each require a slightly different approach. In the following we consider two different sources of information.

Using level information. In a situation where the level of a measured biomarker (e.g. Ab or pathogen levels in blood) exhibits predictable temporal variation we can extract information on the time since infection from the measured level [18]. For example, in the particular case of MORV, Fig 1 clearly shows that the Ab-level contains information about the time since infection.

First, let us consider the case of a single level x_i^k where we have to determine $P(x_i^k|t_i, \theta)$, i.e. the conditional probability of measuring level x_i^k if the individual was infected at time θ and tested at time t_i. As is clear from the data shown in Fig 1, a particular value of the time since infection $t_i - \theta$ does not necessarily result in a single possible biomarker level due to variation

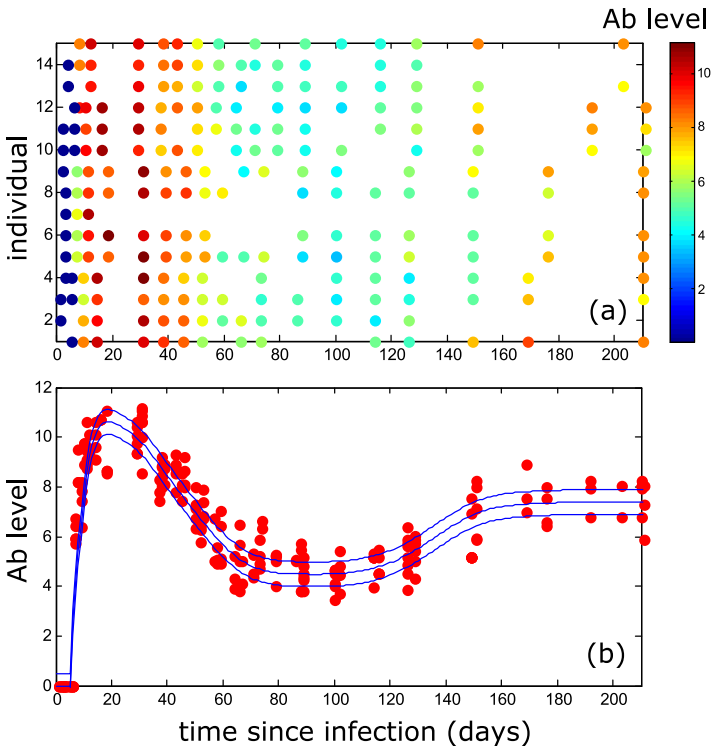

Fig 1. Temporal variation of antibody levels obtained from experimental data [23] for 15 different individuals (a) and for all individuals combined (red dots) with fitted function mean and standard deviation (blue lines) (b).

doi:10.1371/journal.pcbi.1004882.g001

caused by inherent measurement errors, temporal variation and/or individual differences. The measured level x_i^k at time t_i can be written as

$$x_i^k = L(t_i - \theta) + \delta_i, \tag{2}$$

$$\delta_i \sim \mathcal{N}(0, \sigma(t_i - \theta)), \tag{3}$$

i.e. the mean level corresponding to a time since infection, $L(t_i - \theta)$, plus an 'error' δ_i. This model and the error distribution are system-specific, and can take any empirical form as long as it adequately describes the course of the biomarker over time. It is typically derived from experimental infection data. Here, we assume that the error is normally distributed, with a variance σ that may be dependent on the time since infection $t_i - \theta$, because this is probably a common situation. Using these approximations, we arrive at the following conditional probability for a single level measurement:

$$P(x_i^k | t_i, \theta) = \frac{1}{\sqrt{2\pi}\sigma(t_i - \theta)} \exp\left\{ -\frac{1}{2[\sigma(t_i - \theta)]^2} \left[x_i^k - L(t_i - \theta)\right]^2 \right\}, \tag{4}$$

with $t_i - \theta$ the time since infection.

This model describes the conditional probability based on a single measurement, but one often has more information on the evolution of the levels, since an individual may be encountered and tested at different times. In this case, the temporal level information is contained within a vector $\boldsymbol{X^k}$ of which the different coordinates represent the responses measured at times $\boldsymbol{T}$. If we again consider the individual to have been infected at time t, Eqs 2 and 3 can be

generalized to

$$\boldsymbol{X}^k = L(\boldsymbol{T}-\theta)+\boldsymbol{\delta}, \quad (5)$$

$$\boldsymbol{\delta} \sim \mathcal{N}(0,\boldsymbol{\Sigma}), \quad (6)$$

with the covariance matrix Σ over all n times the individual was tested, and δ a $n-$ dimensional vector drawn from a multivariate normal distribution. Finally, this results in

$$P(\boldsymbol{X}^k|\boldsymbol{T},\theta) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \times \exp\left\{-\frac{1}{2}[\boldsymbol{X}^k - L(\boldsymbol{T}-\theta)]^T\boldsymbol{\Sigma}^{-1}[\boldsymbol{X}^k - L(\boldsymbol{T}-\theta)]\right\}.$$

The covariance matrix would typically be inferred from experimental data and accounts for the possible interdependence of level responses at different times. Indeed, the error $\boldsymbol{\delta}$ of different measurements may not be independent over time. Also, it is possible that part of the variance is caused by individual differences, i.e., $\boldsymbol{\delta} = \boldsymbol{\delta}_{\text{ind}} + \boldsymbol{\delta}_{\text{noise}}$, as some individuals may have a stronger immune response (higher overall levels) than others.

Applied to MORV: Ab level. We apply this to MORV by considering information about one Ab (IgG) measurement, shown in Fig 1. First, in order to arrive at errors that can be adequately described by a normal distribution, we take the logarithm of the Ab level. Next, we estimate $L(t)$ by fitting a smooth spline to the data to arrive at the curve shown in Fig 1b.

Then, we subtract the corresponding L-value from each datapoint and calculate to what extent individual variation and temporal variation account for the variance observed in the residual errors, as this would then have to be taken into account in the covariance matrix. Using an ANOVA, we found no significant effect of individual ($p = 0.085$) or time ($p = 0.089$) on the variation of the residual errors. Based on the sum of squares, the relative contributions to the total variance were estimated to be 1.3% for time and 9.6% for individual. From this analysis, we find that the effects of individual and time can be ignored, compared to the residual variance, and consequently we consider the covariance matrix to be proportional to the unitary matrix, $\sigma^2\,\boldsymbol{I}$, independent of t. All off-diagonal elements are assumed zero. The residual standard deviation was measured to be $\sigma = 0.99$ and approximated to 1.

Note that although we here estimate $L(t)$ using a spline method and with the assumption that there is no individual or temporal effect on variation, $P(\boldsymbol{X}^k|\boldsymbol{T},\theta)$ can be modeled using any method, as long as the model adequately describes the data. Indeed, an alternative to using a spline method is to use a mechanistic model, and an alternative to determine the appropriate covariance structure is to use a hierarchical modelling approach in which likelihood theory is used to test the contribution of the different sources of variability (see e.g. [17, 18, 24]).

Using presence/absence information. Often, information on presence/absence of biomarkers is more easily available than level data. This can be due to biomarker assay limitations, because level variability of the measured biomarker is too high and unpredictable, or because the levels do not change sufficiently over time. In such situations, it may be possible to use presence ($x_i^k = 1$) or absence ($x_i^k = 0$) of a biomarker (e.g. IgG, IgM, virus), often measured using assays that result in values above or below a detection threshold. Given that an individual was infected at time θ, the probability of biomarker presence or absence x_i^k at time t_i is given by

$$P(x_i^k|t_i,\theta) = x_i^k[2p(t_i-\theta)-1]+[1-p(t_i-\theta)],$$

which would typically be derived from experimental infection data.

In the case of multiple (n) measurements, presence/absence data are contained in a vector $\boldsymbol{X}^{\boldsymbol{k}}$, with n measurements $[x_1^k, x_2^k, \ldots, x_n^k]$, where x_n^k is the n-th measurement indicating presence (1) or absence (0). Assuming that measurements at different times are independent, we can write

$$P(\boldsymbol{X}^k|\boldsymbol{T},\theta) = \prod_{i=1}^{n} P(x_i^k|t_i,\theta). \tag{7}$$

Applied to MORV: Ab presence. Usually in epidemiology only information about Ab presence or absence (seroconversion events) is used to estimate the time since infection, resulting in incidence estimates with low temporal resolution [25, 26]. Here, we use that situation as a reference, in order to evaluate the improvements offered by using Ab level instead of only presence/absence data.

When only considering Ab presence, the measurement x_i^{ab} is a binary variable of which the value depends on whether Ab was present (1) or absent (0) at time t_i. The probability $p^{ab}(t)$ of detecting Ab in blood if an animal was infected at $t = 0$ is then given by

$$p^{ab}(t \leq 6) = 0$$

$$p^{ab}(t > 6) = 1,$$

as it was found that Ab are never present before day 7 after infection [23]. After this initial period, we assume the test to be sensitive enough to detect Ab presence with a probability of 1 (Fig 1).

Applied to MORV: Virus presence in blood. Based on experimental data, the probability $p^{vb}(t)$ to detect virus in blood (Vb) if an animal was infected at $t = 0$ can be adequately modeled by

$$p^{vb}(t \leq 1) = 0$$
$$p^{vb}(1 < t \leq 8) = 1$$
$$p^{vb}(t > 8) = \exp[-0.3(t-8)],$$

as shown in Fig 2a.

Applied to MORV: Virus presence in excretions. Similar to using information on Vb, another source of information is the presence/absence of virus in excretions (Ve; urine, saliva or feces). Based on the experimental data shown in Fig 2b, we model the probability $p^{ve}(t)$ to detect Ve if an animal was infected at $t = 0$ as

$$p^{ve}(t \leq 2) = 0$$
$$p^{ve}(2 < t \leq 12) = (t-2)/20$$
$$p^{ve}(12 < t \leq 45) = (45-t)/66$$
$$p^{ve}(t > 45) = 0.$$

Combined biomarker information. After modeling all biomarkers of interest, the separate models can easily be combined into one conditional probability of the time of infection that incorporates information about different biomarkers, including levels (or presence/absence) of different antibodies (e.g. IgG, IgM, . . .), pathogen (e.g. virus, bacteria) concentration (or presence/absence), and in different tissues (blood, excretions, organs, . . .). One should keep in mind that the errors, levels or presence of some of the different sources can be correlated, which should be taken into account in the covariance matrix.

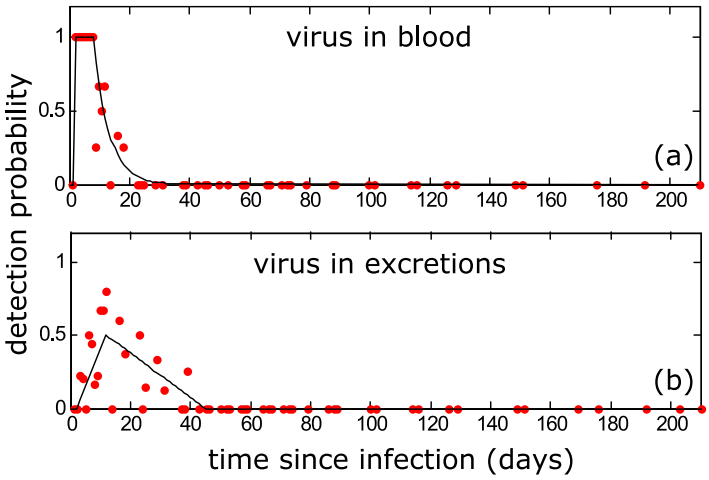

Fig 2. Probability of virus presence in blood (a) and excretions (b), estimated from experimental data [23]. Detection probability is given by the proportion of tested individuals that was RNA-positive on a given sampling day.

doi:10.1371/journal.pcbi.1004882.g002

If we assume N independent sources of information, we can combine these by simple multiplication of their respective conditional probabilities to arrive at

$$P(\boldsymbol{X}|\boldsymbol{T},\theta) = \prod_{k=1}^{N} P(\boldsymbol{X}^k|\boldsymbol{T},\theta), \tag{8}$$

where k runs over N different sources of information. The resulting conditional probability can then be inserted into Eq 1.

Modeling $P(\theta|T)$

Because an individual can of course only have been infected when it was alive and present in the population, the estimation of θ can be improved by incorporating prior information about the probability of an individual being alive/present, i.e. by modeling $P(\theta|\boldsymbol{T})$. Here, we show how to implement information on mortality rate and maximum life span, age at the time of sampling, and encounter probability, but note that any source of information can be used in a similar way as long as it results in a realistic prior distribution.

Knowledge about the maximum life span can be informative because it sets an upper boundary to the possible time since infection, and is especially useful in situations where the maximum life span is short relative to the possible time since infection. If an individual was last tested at time t_n and the maximum life span is known, then the prior distribution $P(\theta|\boldsymbol{T})$ can be reduced to

$$P(\theta|\boldsymbol{T}) \sim \frac{1}{\text{life span}}[\theta > (t_n - \text{life span})][\theta < t_n],$$

with $[. < .]$ is a boolean operator that returns 1 or 0 when the equality is true or false, as shown in Fig 3a.

Similarly, one could make use of the mortality rate, as this is directly associated with the possible age of an individual. If an individual was first encountered at time $\boldsymbol{t}_1$ and we assume a mortality rate γ as inferred from data, we arrive at prior distribution

$$P(\theta|\boldsymbol{T}) \sim \max[\exp(\gamma(\theta - t_1)), 1][\theta < t_n],$$

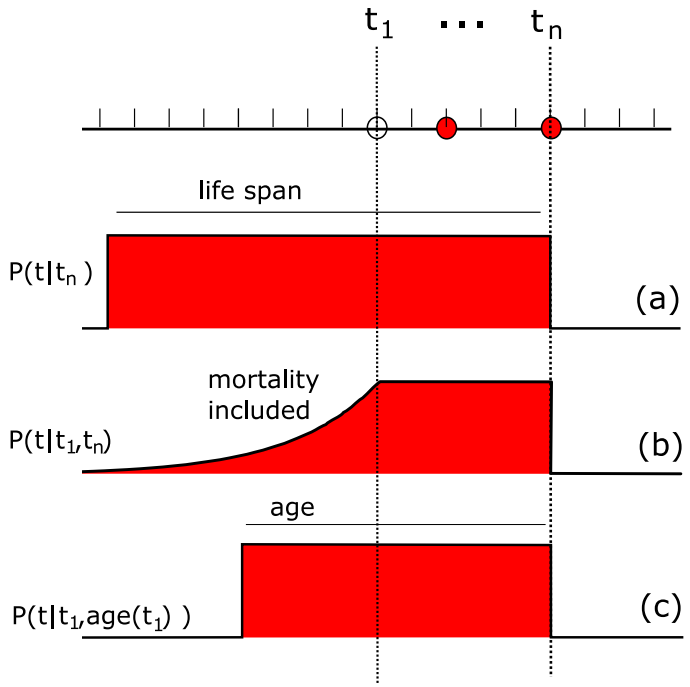

Fig 3. Example of the possible use of information about maximum life span (a), mortality rate (b) and individual age (c). The three dots on the time axis indicate the different times at which a hypothetical individual was sampled. The red blocks indicate the probability of being alive at a certain point back in time, which can be included as prior information on the estimated time of infection.

doi:10.1371/journal.pcbi.1004882.g003

as shown in Fig 3b. This figure clearly shows that, due to mortality, it becomes increasingly unlikely for individuals to have been alive, and therefore infected, further in the past.

When more precise information exists on the age of an individual focus individual (which is trivial for humans, while for wild animals this can be based on physiological or morphological features such as weight), this can be taken into account explicitly by including

$$P(\theta|\boldsymbol{T}) \sim [\theta > (t_1 - \text{age}(t_1))][\theta < t_n],$$

if the individual was first encountered at time t_1, see Fig 3c.

More applicable to wildlife infections is the use of encounter probability (typically termed trapping or capture probability, but for consistency and human application we will here refer to it as encounter probability). In a typical capture-mark-recapture study, only a proportion of individual is captured during each session, and well-developed methods exist for estimating encounter probability [27, 28]. This encounter probability can be used to estimate the likelihood of an individual being alive at a certain point in time, assuming a closed population during that time (no migration). If an individual is first encountered at time t_1, the probability of it being born at time θ decreases with $t_1 - \theta$, as it becomes increasingly unlikely that it was not encountered during $(t_1 - \theta) / \Delta t$ trapping sessions.

If we estimate encounter probability p_{enc} for every trapping session, this information can be used to further improve the prior time distribution:

$$\begin{aligned} P(\theta|\boldsymbol{T}) &\sim \max[(1 - p_{enc})^{(t_1-\theta)/\Delta t}, 1][\theta < t_n] \\ &\approx \max\left\{\exp\left[p_{enc}\frac{(\theta - t_1)}{\Delta t}\right], 1\right\}[\theta < t_n], \end{aligned}$$

where Δt is the sampling or trapping interval time, and with the latter approximation valid only when $p_{enc} << 1$. This approach only holds if one can assume a closed population where every individual was in the population during its lifetime and the effects of migration are negligible.

One could also use seasonal information or cross-sectional data to inform the prior $P(\theta|\boldsymbol{T})$, or in fact any other data source that contains any type of information about the time since infection.

Decision Criterion

Given the resulting posterior probability $P(\theta|\boldsymbol{X}, \boldsymbol{T})$, the observer still has to use a decision criterion to decide which time of infection θ is most likely. Probably the most obvious decision criterion is the mean squared error (MSE) of the time since infection by selecting the $\widehat{\theta}_i$ for which

$$MSE = \frac{1}{N_{ind}} \sum_{1=1}^{N_{ind}} \left[\widehat{\theta}(i) - \theta(i)\right]^2,$$

with i running over a population of N_{ind} individuals, is minimal. It can be shown that this is the case for $\widehat{\theta} = \int d\theta P(\theta|\boldsymbol{T}, \boldsymbol{X})\, \theta$ [29].

In order to assess the quality of the estimates, the remaining uncertainty on the time since infection can be inspected conditional on the observed data ($\boldsymbol{X}$, $\boldsymbol{T}$), which can be quantified using the conditional entropy $E(\theta|\boldsymbol{X}, \boldsymbol{T})$ [29], i.e.,

$$E(\theta|\boldsymbol{X}, \boldsymbol{T}) = \int_{\theta'} d\theta' p(\theta'|\boldsymbol{X}, \boldsymbol{T}) \log_2 p(\theta'|\boldsymbol{X}, \boldsymbol{T}),$$

where θ' runs over all possible time since infection values. Conditional entropy is a commonly used measure in information theory that quantifies (in *bits*) the remaining amount of uncertainty about the actual value of the quantity of interest (here: time since infection). The highest entropy is attained for a uniform posterior probability distribution (maximum uncertainty), whereas the minimum (zero) entropy is obtained when there is no uncertainty left about the actual value [29]. In an epidemiological context, the entropy value can be used to improve the reliability of estimated incidence (see next paragraph) by removing all estimates of θ for which the entropy value is larger than a threshold value. The choice of this threshold value will mostly depend on the trade-off between sample size and estimation error: a low threshold value will generally result in a higher quality of the remaining θ estimates, but at the cost of reducing the final size of the dataset, and will therefore be dataset-specific.

Estimating Incidence

One of the main purposes of knowing the time of infection of an individual is to analyse and model infection incidence on a population level. To this end, we need to estimate the time of infection θ_i for all sampled individuals i in the population and count the number of newly infecteds on a regular (usually daily) basis. Because in most situations only a proportion of individuals will be encountered and sampled, the "real" proportion of new infections needs to be estimated. This can be done by dividing the number of infecteds by an estimate of the proportion of encountered individuals. Given a certain sampling interval Δt and an encounter probability at each session (p_{enc}), this proportion can be approximated by

$$\text{proportion encountered} = \gamma \int dt \exp(-\gamma t)\left[1 - (1 - p_{enc})^{t/\Delta t}\right],$$

where the integral runs over all the survival times t following an exponential distribution with $1/\gamma$ (the average lifespan of an individual in our simulation), $t / \Delta t$ is the approximate number of sampling sessions during lifetime t, and $(1 - p_{enc})^{t/\Delta t}$ is the approximate probability that an individual is never encountered during these sessions.

Application to MORV Infection in *M. natalensis*

Next, in order to test the back-calculation scheme, we need a dataset of individuals in a population, with full knowledge of their infectious status at each moment. Also, to test the efficacy of the method as a function of sample size (with regard to intervals between sampling sessions as well as the sampling effort), we need datasets collected under different trapping regimes. We therefore simulate MORV transmission in a population of multimammate mice, "sampled" in different trapping sessions, with each individual given simulated infection attribute data based on the experimentally-derived [23] course of Ab levels and probability of virus presence in blood and excretions. These simulated data are equivalent to epidemiological data obtained through surveys with repeated sampling, but now of course with the difference that our simulated data are completely known for testing purposes. All simulated data, as well as the Matlab code used to apply the time of infection estimation method, can be found in S1 Data.

As input for the model, we use simulated data from an existing individual-based spatially-explicit SEIR model, which models the population dynamics and the transmission of Morogoro virus in *M. natalensis* [30]. In this model, individuals are born in the susceptible (S) state and can become infected through contact with infectious (I—infectious state) individuals. When infected, they enter a latent stage (E—exposed state) during which they cannot transmit the virus, until they become infectious (I) after around 6 days. After around 45 days they stop being infectious, recover from the infection (R—recovered state) and remain immune against re-infection for the remainder of their life. Latent and infectious periods were simulated assuming an exponential distribution. The simulation is run over a total area of 10ha, but in order to recreate a realistic situation in which individuals can move freely in and out of the study site, only the individuals that are encountered within a central 5ha area were available for "trapping". Realistic population densities and fluctuations are used, ranging between around 10 and 150 per ha. After a simulation burn-in period, two years of data are considered (from day 1000 until 1730).

Throughout the simulation we keep track of each individual's age, time since infection t, and we simulate trapping sessions with a time interval Δt, in which every individual present in the 5ha area has a probability p_{trap} to be trapped. Whether an individual is trapped or not is determined using pseudo random numbers. This way, for every individual we can generate an artificial set of measurements $(\boldsymbol{T}, \boldsymbol{X}^{\boldsymbol{k}})$ that we can then use to estimate the time of infection $\widehat{\theta}$. $\boldsymbol{X}^{ab}$ are random realisations according to the multivariate distribution shown in Eq 8 at times $\boldsymbol{T}$. $\boldsymbol{X}^{vb}$ and $\boldsymbol{X}^{ve}$ are random draws with respective probabilities p^{vb} and p^{ve} at times $\boldsymbol{T}$. We vary the time intervals between capture sessions using Δt = 1, 7, 14, 28, 56 days, as well as the probability for each of the individuals to be captured using $p_{trap} \in (0, 1)$.

We implement a maximum life span of *M. natalensis* of 450 days based on [31]. The average mortality rate (averaged across the year) is calculated from the simulation data, and estimated to be μ = 0.008537 mice/day (average life span of 117 days). Both maximum and average life span are used as prior information for all time of infection estimates.

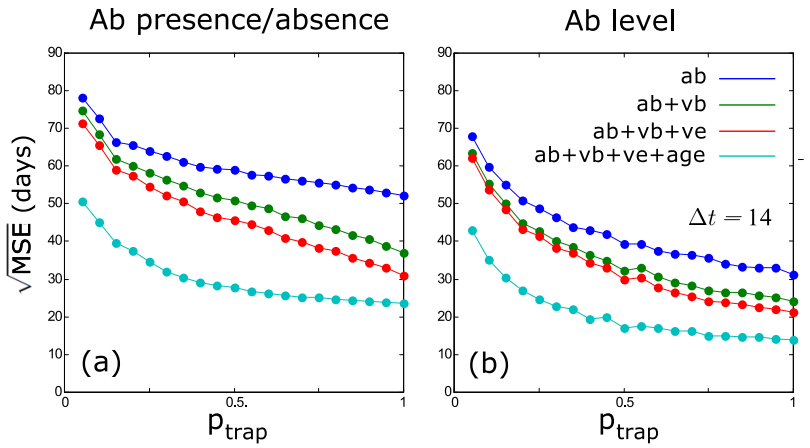

Fig 4. Estimation error ($\sqrt{MSE}$) on the time of infection for different encounter probabilities (p_{trap}) and for different levels of included prior information (ab: antibody, vb: virus in blood, ve: virus in excretions, age: individual age); (a) is based on antibody presence/absence, while (b) is based on antibody levels. The trapping interval was 14 days for all situations.

doi:10.1371/journal.pcbi.1004882.g004

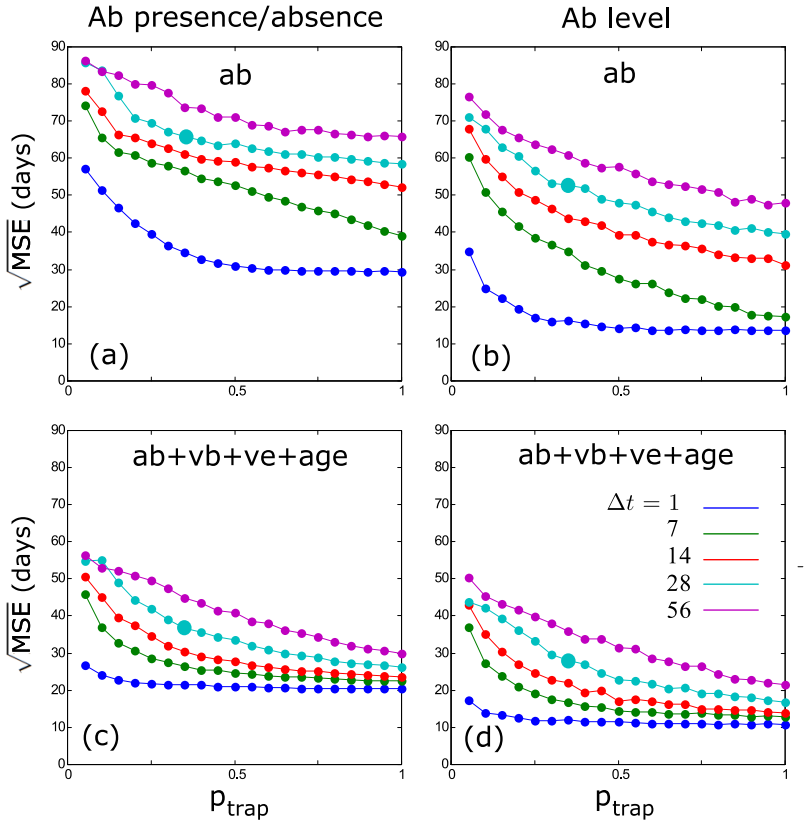

Fig 5. Estimation error ($\sqrt{MSE}$) on the time of infection for different encounter probabilities (p_{trap}) and different sampling intervals (Δt) (ab: antibody, vb: virus in blood, ve: virus in excretions, age: individual age); (a) and (c) are based on antibody presence/absence, while (b) and (d) are based on antibody levels; (a) and (b) only include antibody information, while (c) and (d) include all available information. The larger dots on the 28 day line indicate the situation for which incidence plots are shown in Fig 6.

doi:10.1371/journal.pcbi.1004882.g005

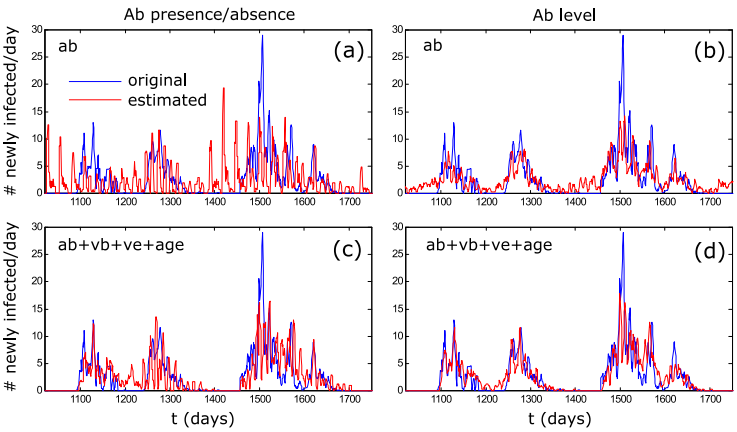

Fig 6. Simulated (blue) and estimated (red) incidence using different sources of information; (a) and (c) are based on antibody presence/absence, while (b) and (d) are based on antibody levels; (a) and (b) only include antibody information, while (c) and (d) include all available information; (a) represents the situation that is mostly used in existing studies. Larger dots on Fig 5 (28 day line) indicate the situation for which incidence plots are shown.

doi:10.1371/journal.pcbi.1004882.g006

Results and Discussion

Ab Level vs Presence/Absence, without Additional Information

The estimation of the time since infection is much improved by the use of Ab levels, as opposed to when only using Ab presence/absence data (Figs 4 and 5). The use of Ab levels also results in a much better reconstruction of incidence dynamics, even without including additional information such as virus presence or individual age (Fig 6). When using Ab presence/absence data, incidence can only be estimated with a low temporal resolution, the main consequence being that the peaks and troughs of the incidence dynamics were estimated badly (Fig 6). Although the incidence peaks are estimated quite well when using Ab levels, the periods of low incidence are still often over-estimated (Fig 6).

Including Additional Information

The inclusion of additional information (Vb, Ve, individual age) greatly improves the estimation of time since infection and incidence (Figs 4–6). Interestingly, this effect is more pronounced when using Ab presence/absence than when using Ab levels. The combined use of Ab levels and other available information results in the highest quality reconstruction of incidence dynamics, where the inclusion of additional information mainly reduces the previously observed over-estimation of low incidence levels between peaks.

Nevertheless, even when using Ab presence/absence instead of Ab level data, incidence can be reconstructed well when including Vb, Ve and individual age. This is encouraging, given the fact that many datasets, especially for wildlife infections, already contain some or all of this information; it means that by applying the back-calculation method, many existing incidence estimations can be improved significantly without additional laboratory or sampling efforts.

Sampling Frequency and Encounter Probability

The quality of the estimates strongly depends on sampling frequency (or trapping interval) and the proportion of individuals that is encountered (or trapped) and sampled. While more additional prior information always results in a better estimation of the time since infection, we see that, at low (realistic) encounter probabilities, this effect is strongest (Fig 4). We also observe

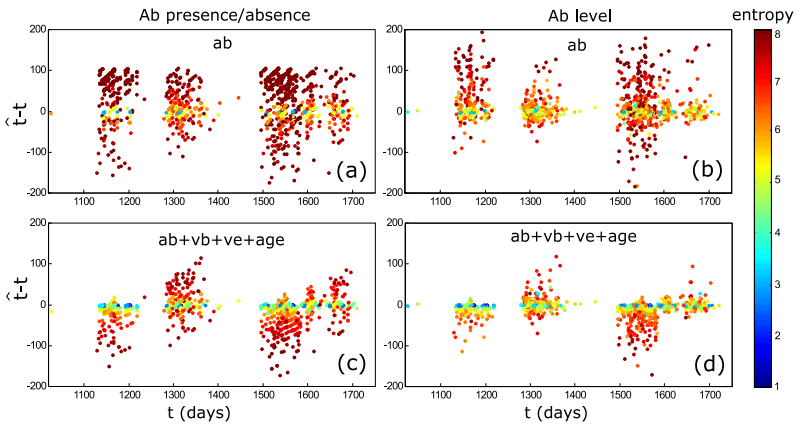

Fig 7. Difference between the estimated and real time since infection in relation to the entropy level (bits). Each datapoint is a "sampled" individual.

doi:10.1371/journal.pcbi.1004882.g007

that a higher sampling frequency results in better estimates (Fig 5), and this is largely an effect of increased sample sizes: when adjusting the trapping probability to equalise sample sizes of different sampling frequencies, this effect mostly disappears (S1 Fig). This means that, in theory, similar results can be reached for any sampling frequency or trapping interval, but only if the sampling effort is increased so that a sufficient number of individuals can be sampled. Nevertheless, we observe that long sampling intervals (28–56 days) generally result in lower quality estimates (S1 Fig), indicating that a shorter interval would still be preferred.

Entropy Threshold

In the model, we introduce the use of entropy (which is inversely related to information) as an indicator of the amount of uncertainty contained by an estimate. Fig 7 shows how estimates of the time since infection with a higher deviation from the real time since infection generally also contain less information (i.e. have a higher entropy). Similarly, we observe a strongly positive correlation between the MSE of the estimated time of infection and the entropy level (S2 Fig). Therefore, by removing estimates above a critical entropy value, the MSE can be lowered, albeit at the cost of a lower sample size. Because of this trade-off it is not possible to suggest an optimal critical entropy cut-off value, which should rather be chosen depending on the specific situation, sample size and quality of available information.

Model Limitations

Although the model performs well and seems promising for a wide range of situations, there are a number of important assumptions and prerequisites that must be met before it is possible to apply the model to data. First, of course, empirical data on the dynamics of biomarkers (e.g. antibodies, viral RNA, etc) within individuals must be available. These can be relatively straightforward data such as knowledge about when after infection individuals seroconvert and how long antibodies remain detectable, or more elaborate information such as the temporal variation of antibody and virus levels after infection.

Then, these data can only be used if they are sufficiently consistent across individuals. If there is too much inter-individual variation in the shape of biomarker dynamics, it will not be possible to predict individual patterns. This does not however mean that there can not be individual variation in the magnitude of the response, as this would in fact be easy to implement into the model.

Further care must be taken if biomarker data have been obtained through laboratory experiments. Because laboratory conditions are often controlled and limited, natural variation in factors such as individual differences in immune response, stress, secondary infection, initial dose, boosting, etc. may result in different biomarker dynamics that could invalidate a time since infection model if they can not be incorporated into the model [32]. Ideally this is tested through a comparative study between laboratory and field patterns, but if such a study has not been done we must assume that the patterns observed in laboratory conditions apply to the natural situation.

Other factors that could render the use of a time since infection model difficult are the existence of maternal antibodies and the simultaneous presence of chronically and acutely infected individuals, as these factors would be difficult (but not necessarily impossible) to disentangle and take into account. On the other hand, under certain conditions these factors may even improve the model, as they provide additional information; for example, if maternal antibodies only occur for a certain period in newborn individuals, and if maternal antibodies can be distinguished from other antibodies (e.g. because of lower levels or using a different assay), this information can likely improve the estimation of the time since infection when incorporated into the model.

Model Novelty and Applicability

Under the conditions described here, the model is a significant improvement on existing models (e.g. [14, 17, 18, 33]). It provides a relatively simple probabilistic framework for the incorporation of any data source that can inform the estimation of time since infection, such as biomarker level/presence, age, season, sex, weight, etc., and thus allows for the use of individual-level data to interpret cross-sectional survey data and estimate population-level incidence. An important strength of the method is that it does not assume a certain form for the underlying models, which makes it possible to use a general spline method but also a more specific ordinary differential equation (ODE) method when a good ODE can be found (e.g. [17]).

More specifically for wildlife infections, the method has the potential to enhance existing long-term data. Often, large logistical efforts are necessary to collect longitudinal data on wildlife infections, and even the best datasets have a relatively low temporal resolution, typically consisting of monthly (but often less frequent) capture sessions [5, 34–37]. Prevalence or incidence patterns resulting from such data are usually also limited to this capture frequency, and to our knowledge the only efforts for improving these data have been the rough estimation of seroconversion events between two capture sessions (e.g. [38, 39]). We have shown however that by integrating multiple sources of information (that have often already been collected or analysed), the quality of incidence data can be greatly improved, especially (but not uniquely) when predictable antibody level dynamics are available.

Conclusion

Due to its flexibility, the model presented here allows the integration of multiple sources of information, thus making optimal use of all available data for estimating individual times of infection and population incidence. It provides a conceptually simple, flexible framework for estimating the time since infection and incidence of human as well as wildlife infections, and can potentially be used to significantly improve incidence estimation based on already existing data.

Supporting Information

S1 Fig. Estimation error ($\sqrt{MSE}$) for different sampling intervals (Δt) and different sample size corrected encounter probabilities (m) (ab: antibody, vb: virus in blood, ve: virus in excretions, age: individual age); (a) and (c) are based on ab presence/absence, while (b) and (d) are based on ab levels; (a) and (b) only include ab information, while (c) and (d) include all available information. In order to adjust the trapping probability so that the number of individuals captured during a month is more or less teh same, a constant m was used such that $p_{trap} = 28^{*} m / \Delta t$, with $m \in [1, 20]$. Smaller m-values correspond with a lower p_{trap} but a similar number of individuals.
(EPS)

S2 Fig. (a): Frequency distribution of entropy values for different levels of additional information; (b) Correlation between entropy and the mean absolute difference between estimated and real time since infection. Different lines (a, b, c, d) correspond with the respective situations in Fig 7 in the main text.
(EPS)

S1 Data. Matlab code and transmission model simulation results. This file contains the Matlab code used to generate the results in this article, as well as the data. The datafile consists of 3 matrices, where the columns are the daily model situations (increasing in time from left to right, starting after a 500-day burn-in period and selected within a 5 ha grid as described in the methods) and the rows represent all individuals present in the simulation. Empty cells (no individuals) are indicated by a negative number. The id matrix gives the unique identifier of each individual, and the corresponding age (in days) and time since infection (in days) values are given in the two other matrices.
(ZIP)

Author Contributions

Conceived and designed the experiments: BB NH HL JR. Analyzed the data: BB JR NH. Wrote code and performed simulations: JR. Wrote the paper: BB NH PB HL JR.

References

1. Khan AS, Tshioko FK, Heymann DL, Le Guenno B, Nabeth P, et al. (1999) The reemergence of Ebola hemorrhagic fever, Democratic Republic of the Congo, 1995. Commission de Lutte contre les Epidémies à Kikwit. J Infect Dis 179: S76–S86. PMID: 9988168
2. Brookmeyer R (2010) Measuring the HIV/AIDS epidemic: approaches and challenges. Epidemiol Rev 32: 26–37. doi: 10.1093/epirev/mxq002 PMID: 20203104
3. Richardson Ba, Hughes JP (2000) Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty. Biostatistics 1: 341–354. doi: 10.1093/biostatistics/1.3.341 PMID: 12933514
4. Sal y Rosas VG, Hughes JP (2011) Nonparametric and semiparametric analysis of current status data subject to outcome misclassification. Stat Commun Infect Dis 3: 7.
5. Begon M, Feore S, Bown KJ, Chantrey J, Jones T, et al. (1998) Population and transmission dynamics of cowpox in bank voles: testing fundamental assumptions. Ecol Lett 1: 82–86. doi: 10.1046/j.1461-0248.1998.00018.x
6. Hangartner L, Zinkernagel RM, Hengartner H (2006) Antiviral antibody responses: the two extremes of a wide spectrum. Nat Rev Immunol 6: 231–43. doi: 10.1038/nri1783 PMID: 16498452
7. Best JM, Banatvala JE, Watson D (1969) Serum IgM and IgG responses in postnatally acquired rubella. Lancet 2: 65–68. doi: 10.1016/S0140-6736(69)92386-1 PMID: 4182759

8. Uhr JW, Finkelstein MS (1963) Antibody formation. IV. Formation of rapidly and slowly sedimenting antibodies and immunological memory to bacteriophage phi-X 174. J Exp Med 117: 457–477. doi: 10.1084/jem.117.3.457 PMID: 13995245
9. Schoppel K, Kropff B, Schmidt C, Vornhagen R, Mach M (1997) The humoral immune response against human cytomegalovirus is characterized by a delayed synthesis of glycoprotein-specific antibodies. J Infect Dis 175: 533–544. doi: 10.1093/infdis/175.3.533 PMID: 9041323
10. Baccard-Longere M, Freymuth F, Cointe D, Seigneurin JM, Grangeot-keros L (2001) Multicenter evaluation of a rapid and convenient method for determination of Cytomegalovirus immunoglobulin G avidity. Clin Diagn Lab Immunol 8: 429–431. doi: 10.1128/CDLI.8.2.429-431.2001 PMID: 11238233
11. Gray JJ, Cohen BJ, Desselberger U (1993) Detection of human parvovirus BI9-specific IgM and IgG antibodies using a recombinant viral VP1 antigen expressed in insect cells and estimation of time of infection by testing for antibody avidity. J Virol Methods 44: 11–23. doi: 10.1016/0166-0934(93)90003-A PMID: 8227275
12. Revello MG, Genini E, Gorini G, Klersy C, Piralla A, et al. (2010) Comparative evaluation of eight commercial human cytomegalovirus IgG avidity assays. J Clin Virol 48: 255–259. doi: 10.1016/j.jcv.2010.05.004 PMID: 20561816
13. Versteegh FGA, Mertens PLJM, de Melker HE, Roord JJ, Schellekens JFP, et al. (2005) Age-specific long-term course of IgG antibodies to pertussis toxin after symptomatic infection with *Bordetella pertussis*. Epidemiol Infect 133: 737–748. doi: 10.1017/S0950268805003833 PMID: 16050521
14. de Melker HE, Versteegh FGA, Schellekens JFP, Teunis PFM, Kretzschmar M (2006) The incidence of *Bordetella pertussis* infections estimated in the population from a combination of serological surveys. J Infect 53: 106–113. doi: 10.1016/j.jinf.2005.10.020 PMID: 16352342
15. de Angelis D, Gilks W, Day N (1998) Bayesian projection of the acquired immune deficiency syndrome epidemic. J R Stat Soc Ser C (Applied Stat 47: 449–498. doi: 10.1111/1467-9876.00123
16. Heisterkamp SH, de Vries R, Sprenger HG, Hubben GAA, Postma MJ (2008) Estimation and prediction of the HIV-AIDS-epidemic under conditions of HAART using mixtures of incubation time distributions. Stat Med 27: 781–794. doi: 10.1002/sim.2974 PMID: 17597471
17. Simonsen J, Mølbak K, Falkenhorst G, Krogfelt K, Linneberg A, et al. (2009) Estimation of incidences of infectious diseases based on antibody measurements. Stat Med 28: 1882–1895. doi: 10.1002/sim.3592 PMID: 19387977
18. Teunis PFM, van Eijkeren JCH, Ang CW, van Duynhoven YTHP, Simonsen JB, et al. (2012) Biomarker dynamics: estimating infection rates from serological data. Stat Med 31: 2240–2248. doi: 10.1002/sim.5322 PMID: 22419564
19. Miller FG, Grady C (2001) The ethical challenge of infection-inducing challenge experiments. Clin Infect Dis 33: 1028–1033. doi: 10.1086/322664 PMID: 11528576
20. Marty AM, Jahrling PB, Geisbert TW (2006) Viral hemorrhagic fevers. Clin Lab Med 26: 345–386. doi: 10.1016/j.cll.2006.05.001 PMID: 16815457
21. Borremans B, Leirs H, Gryseels S, Günther S, Makundi R, et al. (2011) Presence of Mopeia virus, an African arenavirus, related to biotope and individual rodent host characteristics: implications for virus transmission. Vector-Borne Zoonotic Dis 11: 1125–1131. doi: 10.1089/vbz.2010.0010 PMID: 21142956
22. Leirs H, Stenseth NC, Nichols JD, Hines JE, Verhagen R, et al. (1997) Stochastic seasonality and nonlinear density-dependent factors regulate population size in an African rodent. Nature 389: 176–180. doi: 10.1038/38271 PMID: 9296494
23. Borremans B, Vossen R, Becker-Ziaja B, Gryseels S, Hughes N, et al. (2015) Shedding dynamics of Morogoro virus, an African arenavirus closely related to Lassa virus, in its natural reservoir host *Mastomys natalensis*. Sci Rep 5: 10445. doi: 10.1038/srep10445 PMID: 26022445
24. Andraud M, Lejeune O, Musoro JZ, Ogunjuni B, Beutels P, et al. (2012) Living on three time scales: the dynamics of plasma cell and antibody populations illustrated for hepatitis A virus. PLoS Comput Biol 8: e1002418. doi: 10.1371/journal.pcbi.1002418 PMID: 22396639
25. Hens N, Shkedy Z, Aerts M, Faes C, Van Damme P, et al. (2012) Modeling infectious disease parameters based on serological and social contact data. A modern statistical perspective. Springer.
26. Bollaerts K, Aerts M, Shkedy Z, Faes C, Van der Stede Y, et al. (2012) Estimating the population prevalence and force of infection directly from antibody titres. Stat Modelling 12: 441–462. doi: 10.1177/1471082X12457495
27. Pollock KH, Nichols JD, Brownie C, Hines JE (1990) Statistical inference for capture-recapture experiments. Wildl Monogr 107: 3–97.
28. White GC, Anderson DR, Burnham KP, Otis DL (1982) Capture-recapture and removal methods for sampling closed populations. Los Alamos, New Mexico: Los Alamos National Laboratory.

29. Cover TM, Thomas JA (1991) Elements of information theory. New York: John Wiley & Sons.

30. Goyens J, Reijniers J, Borremans B, Leirs H (2013) Density thresholds for Mopeia virus invasion and persistence in its host *Mastomys natalensis*. J Theor Biol 317: 55–61. doi: 10.1016/j.jtbi.2012.09.039 PMID: 23041432

31. Leirs H, Verhagen R, Verheyen W (1993) Productivity of different generations in a population of *Mastomys natalensis* rats in Tanzania. Oikos 68: 53–60. doi: 10.2307/3545308

32. Childs JE, Peters CJ (1993) Ecology and epidemiology of arenaviruses and their hosts. In: Salvato MS, editor, The Arenaviridae, New York: Springer US. pp. 331–384. doi: 10.1007/978-1-4615-3028-2_19

33. Simonsen J, Strid M, Mølbak K, Krogfelt K, Linneberg A, et al. (2008) Sero-epidemiology as a tool to study the incidence of Salmonella infections in humans. Epidemiol Infect 136: 895–902. doi: 10.1017/S0950268807009314 PMID: 17678562

34. Smith MJ, Telfer S, Kallio ER, Burthe S, Cook AR, et al. (2009) Host-pathogen time series data in wildlife support a transmission function between density and frequency dependence. Proc Natl Acad Sci 106: 7905–7909. doi: 10.1073/pnas.0809145106 PMID: 19416827

35. Tersago K, Verhagen R, Leirs H (2010) Temporal variation in individual factors associated with Hantavirus infection in bank voles during an epizootic: implications for Puumala virus transmission dynamics. Vector-Borne Zoonotic Dis 11: 715–721. doi: 10.1089/vbz.2010.0007 PMID: 21142469

36. Sluydts V, Davis S, Mercelis S, Leirs H (2009) Comparison of multimammate mouse (*Mastomys natalensis*) demography in monoculture and mosaic agricultural habitat: Implications for pest management. Crop Prot 28: 647–654. doi: 10.1016/j.cropro.2009.03.018

37. Borremans B, Hughes NK, Reijniers J, Sluydts V, Katakweba AAS, et al. (2014) Happily together forever: temporal variation in spatial patterns and complete lack of territoriality in a promiscuous rodent. Popul Ecol 56: 109–118. doi: 10.1007/s10144-013-0393-2

38. Bernshtein AD, Apekina NS, Mikhailova TV, Myasnikov Y, Khlyap L, et al. (1999) Dynamics of Puumala hantavirus infection in naturally infected bank voles (*Clethrinomys glareolus*). Arch Virol 144: 2415–2428. doi: 10.1007/s007050050654 PMID: 10664394

39. Begon M, Hazel S, Telfer S, Bown K, Carslake D, et al. (2003) Rodents, cowpox virus and islands: densities, numbers and thresholds. J Anim Ecol 72: 343–355. doi: 10.1046/j.1365-2656.2003.00705.x