Recommendations for reporting epidemiological parameters

Joshua W. Lambert[1*], Carmen Tamayo[1], Sangeeta Bhatia[2,3], Ruth McCabe[2], Gina Cuomo-Dannenburg[2], Adam Kucharski[1], Anne Cori[2]

1 London School of Hygiene and Tropical Medicine,
2 Imperial College London,
3 UK Health Security Agency,

❍These authors contributed equally to this work.

¤Current Address: Dept/Program/Center, Institution Name, City, State, Country

†Deceased

¶Membership list can be found in the Acknowledgments section.

* joshua.lambert@lshtm.ac.uk

Epidemiological parameters are a necessity in understanding the spread of infectious disease; from the rate of transmission, to severity, to serology, these parameters underline our ability to quantify and respond to disease outbreaks. The estimation of epidemiological parameters has a long history in epidemiology and the models and methods applied have become more complex; and with this complexity comes the numerous ways parameters can be reported in the literature. Here we provide guidance on how to clearly communicate estimated epidemiological parameters, to maximise their secondary use and minimise possible human errors that come with extracting parameters from the literature and applying them in their own epidemiological analysis. Our aim is for future work that reports epidemiological parameters to be consistent, reproducible and comparable.

Introduction

Epidemiological parameters are quantities that characterise the spread of infectious diseases, their epidemiological outcomes and temporal information on dynamics of disease progression and transmission [1]. They are critical to understand epidemic and pandemic dynamics and respond accordingly [2]. Most epidemiological parameters take the form of distributions because there is inherent variability in the epidemiological characteristics being measured. An illustration is the delay from infection to symptom onset. The variability of individuals in immune response and variability of the infectious agent in pathology are two ways, among many others, that lead to some individuals having shorter time delay between infection and onset of symptoms. Due to most epidemiological parameters being described by distributional forms they are estimated by fitting distributions to epidemiological data on cases or contacts.

It has become general practice that when epidemiological parameters are required, either for analyses of epidemiological case data or to make policy decisions like quarantine duration, that the literature is searched to find a suitable peer-reviewed publication reporting the parameter needed. However, this process has several limitations. The time requirement to search through papers to find the highest quality epidemiological parameter means that in time-limited scenarios, for example early in an outbreak when the situation is evolving rapidly and new data is continually gathered, a

suboptimal parameter set may be extracted and used. This has lead to previous *ad hoc* reviews of epidemiological parameters for specific pathogens (e.g. Ebola [3]). The choice of parameter is also likely to be somewhat subjective without a clear quality assurance framework to evaluate and compare different parameter estimates. The manual extraction of copying and pasting parameters out of the literature comes with the risk of discrepancies entering the calculations.

Epidemiological parameters have been reported for many diseases and the data used to infer parameter estimates and the methods of inference vary.

Efforts to compile a centralised database of epidemiological parameters have highlighted the variability and ambiguity in parameter reporting which can lead to uncertainty around what is being reported and how these epidemiological parameters can be applied in other epidemiological analyses [4, 5, 6].

This paper was motivated by several research groups independently attempting to compile a comprehensive library of epidemiological parameters which could serve as a public resource to easily search, filter and extract parameters. These groups gathered for a workshop convened by the World Health Organisation (WHO) Collaboratory in Spring 2024, in which a Global Repository of Epidemiological Parameters (GREP) was discussed, as well as ideas for guidance on reporting epidemiological parameters. The guidelines and examples of incorrect reporting and use were subsequently further developed and resulted in this paper.

Our focus is on guidance for reporting epidemiological parameters from a variety of study types and estimation methodologies. We do not cover or advice on best practises for parameter estimation methods. There are several papers that address avoiding biases and pitfalls [**?**]; brittonEstimationEmergingEpidemics2019a]. Specificall for guidance on methodologies when inferring delay distributions see **(author?)** [7] and **(author?)** [8], and when inferring the reproduction number see **?**] and abbottEstimatingTimevaryingReproduction2020. We focus on the reporting of epidemiological parameters post-inference and the benefits of reporting standardisation on the reuse of parameters by those involved in epidemic or humanitarian response.

We classify bad parameter reporting into two groups: 1) *information loss* and 2) *ambiguous reporting*. Information loss is defined as the presentation or sharing of less than the entirety of the parameter estimates, metadata and contextual information. For example, if a method to infer a case fatality rate outputs the uncertainty of the risk but this is not reported either in the text and the method cannot be reproduced then this information is lost when others extract the parameters, i.e. only a subset of the full inference is shared. Ambiguous reporting can be either the ambiguous reporting of metrics, such as a $X \pm Y$ where Y could be the standard deviation or standard error, or $X(Y_1 - Y_2)$ where the bounds Y_1 and Y_2 could be a confidence interval or credible interval when the inference framework is not reported. In both cases the secondary use of the parameters either is forced to make an assumption on what is reported, or does not utilise the information resulting in information loss.

Here we focus on the bias caused by badly reported epidemiological parameters on simple epidemic methods, sometimes referred to as outbreak analytics [sensu 2], to showcase the erroneous conclusions that can arise when using parameter estimates from the literature. The biases produce here will likely extrapolate to more complex epidemiologilogical modelling. Reporting guidelines can ensure standardised reporting becomes more commonplace, which can make it easier to review, summarise and aggregate epidemiological parameters. We hope that this paper, alongside other works on reporting best practises in epidemiology [**?** 8] enhance the interoperability of

research outputs and inputs.

Guidance

Parameter inference reporting

Parameterisation of distributions Many distributions have standard
parameterisations. In other words, they have one, two or in some cases three parameters
that are denoted by a name and often have a greek letter for shorthand. An example of
this is the Gamma distribution which has the parameterisation shape (α) and rate (β).
However, there are often alternative parameterisations, for the Gamma distribution this
is shape (k) and scale (θ). If left unspecified, the reported parameters may correspond
to different parameters depending on interpretation. Another example of ambiguous
reporting of distribution estimates is when the parameters and summary statistics have
similar names. This is the case for the lognormal distribution, whose common
parameterisation is meanlog (μ) and sdlog (σ) and common summary statistics reported
for a distribution are mean and standard deviation (sd), this is further confused as both
use the same greek letters. Therefore, it is possible to mistake the reporting of one set
of these for the other. Both types of misinterpretation outlined here can result in
substantial differences in the distributions (Figure 1).


Guidance

- Provide the formula for the Probability Density Function (PDF), or Probability Mass Function (PMF) if discrete, in the text or supplementary material.
- Clearly report which distribution parameterisation was used to estimate parameters and provide parameter names in the text.
- Share code used to estimate parameter(s) for others to reproduce and audit methods.

Gamma Distribution
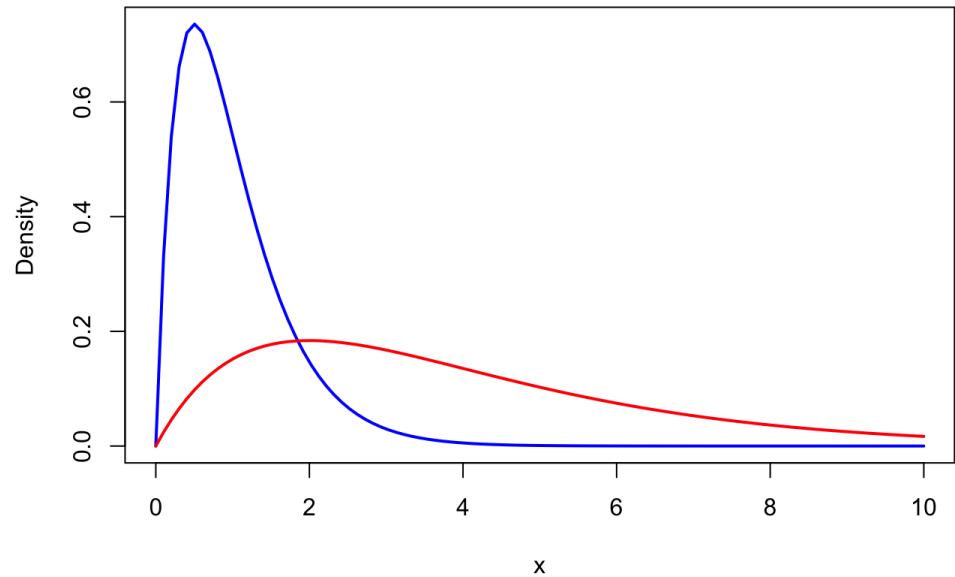
(a) Differences in distributions when misinterpreting the estimated parameters. (a) Two gamma distributions, the blue shows a gamma distribution with shape (α) and rate (β) of 2, the red shows a gamma distribution with shape (k) and scale (θ) of 2. The rate is the reciprocal of the scale ($\beta = 1/\theta$), but it is clear from the left hand plot that misinterpreting the parameter leads to a vastly different distribution density. (b) Two lognormal distributions, the orange shows the density of a lognormal distribution with meanlog (μ) and sdlog (σ) of 0.5, and the green with the meanlog and sdlog of 1.87 and 1.00, respectively (these values are the conversion from meanlog and sdlog into mean and standard deviation). Again showing how misinterpreting the parameters can lead to differences in epidemiological parameters.

Lognormal Distribution
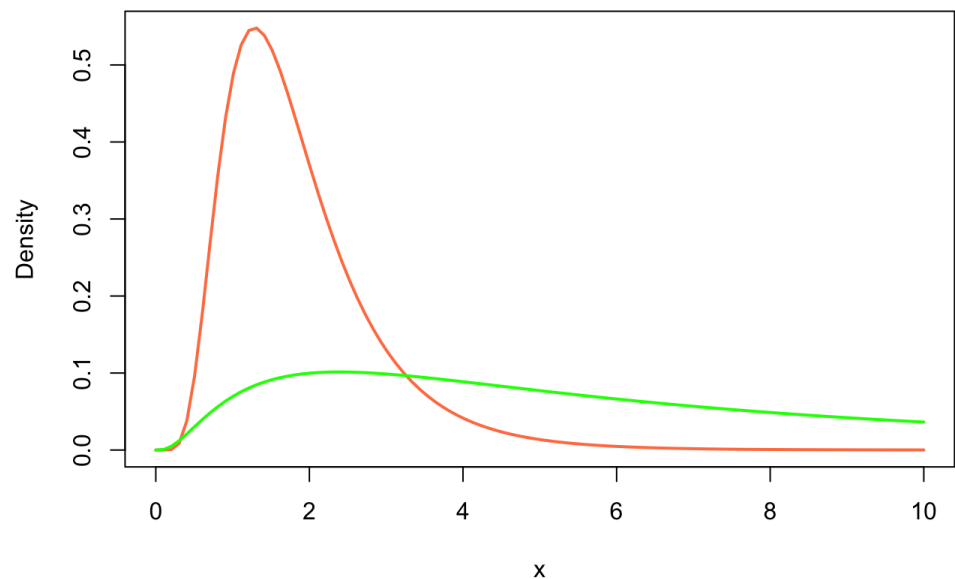
(b)

Figure 1

Parameter estimates vs summary statistics   Instead of reporting the parameter   99
estimates for a parametric distribution, summary statistics may be provided. In some   100
instances a set of summary statistics can be analytically converted into distribution   101
parameters (the specific summary statistics that can be converted into parameters   102
varies by distribution). In those cases where analytical conversion can be done there is   103
no loss in parameter estimate precision, i.e. summary statistics are sufficient statistics.   104
Commonly reported sufficient statistics are the mean and standard deviation or   105
variance of a distribution. However, it can also be the case that summary statistics that   106
cannot be analytically converted to distribution parameters are reported, for example   107
the mean or median and the 95th percentiles of the distribution. In these cases,   108
distribution parameters require a second estimation using a numerical conversion.   109
Numerical conversion can introduce more uncertainty and potentially return erroneous   110
estimates. Below we show an example of the bias and variance of distribution   111
parameters when numerically converted from summary statistics... (see {epiparameter}   112
R package article for full exploration of bias in numerical conversion).   113

> **Use case: Incubation period**
>
> **Example use case: estimating the upper quantile of the incubation period distribution**
> The incubation period, defined as the time between infection and the development of symptoms, is an important epidemiological quantity for establishing guidance on the duration of contact tracing of identified cases, as well as the duration of isolation or quarantine for subsequently identified contacts. It is critical that the duration of isolation mitigates the risk of onward transmission, but does not extend any longer than necessary due to the costs, both financial and social, associated with isolation.
> The incubation period is most typically presented as a probability distribution, but is often summarised by a mean and estimate of the variability (most commonly standard deviation), rather than the full distribution.
> Consider the following example. During a novel Ebola outbreak, the incubation period is estimated using a gamma distribution on just a small number of observations (due to the lack of testing and contact tracing). MCMC methods are used to obtain 100 realisations of a gamma distribution:

114
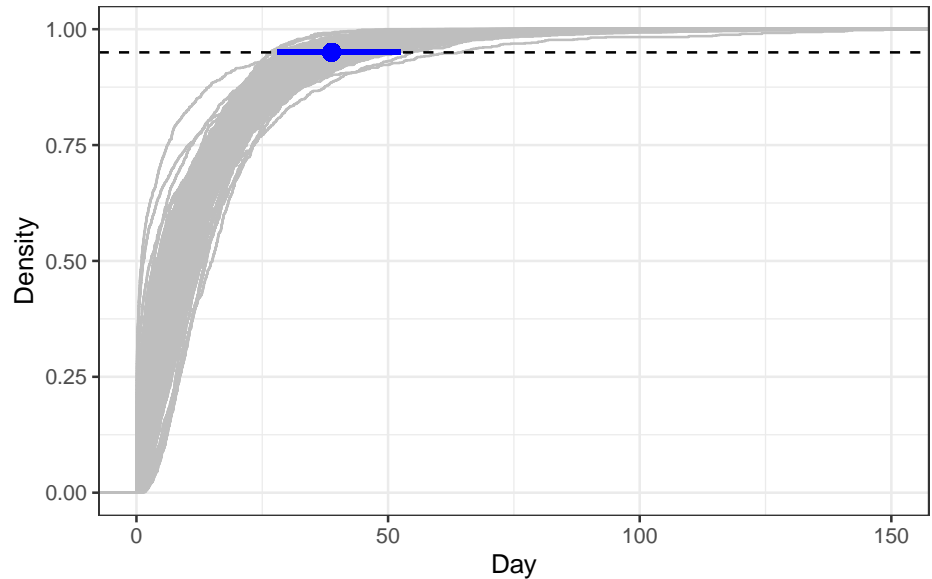
This analysis has been repeated several times for this paper, corresponding to different locations and time periods. For sake of space and simplicity, the results per fit are presented in the typical fashion, as noted above:

Mean: 12.56 (95% CI 9.69 - 16.18)

SD: 13.13 (95% CI 8.6 - 18.16)

However, this omits critical information about the upper 95% quantile of the distribution, which from the figure, is highly variable across realisations. Specifically, this is estimated as:

95% quantile: 38.8 (95% CI 28 - 52.4)

which has substantial uncertainty.

In a future outbreak of a viral hemorrhagic fever, the public health authority in the affected country looks to the previous analysis to guide their quarantine policy. Using only the provided information (the mean and SD), it is possible to use method of moments for the gamma distribution to obtain estimates of the shape and scale parameters (0.93 and 13.62, respectively). Parameterising a gamma distribution according to these parameters corresponds to a 95% quantile of 39 days. This recovers the central estimate above well. However, it is not possible to obtain any uncertainty around this estimate without the authors having provided the full set of samples used to generate these estimates. This may provide a false sense of certainty on which the quarantine policy is based, whereas there may be an argument to either extend or reduce this based on the considerable uncertainty estimated above.

Epidemiological parameters can be dimensionless quantities, for example R0 or secondary attack rate, while others have units. It is especially critical for the accurate reuse of with dimensions parameters that the units are reported. For parameters with a temporal dimension, such as delay distributions, the unit of time ensures that distributions fitted to data on days or weeks can be clearly understood. Another example is viral load data that can be reported as Ct or log10 RNA copies/ml. Many epidemiological parameters will have conventional units, for example incubation period and serial interval in days, or population density in individuals/km, but if readers have to assume units then misinterpretation can have consequences for others that apply the

findings in their own work. ₁₂₅

Guidance: ₁₂₆

- Report the estimated distribution parameters first and foremost before optionally ₁₂₇
  reporting summary statistics. This will avoid any secondary estimation step which ₁₂₈
  can introduce unwanted and unnecessary bias. ₁₂₉
- If a parameter has a unit, report this with the estimates, ensuring it matches the ₁₃₀
  data of input of the model. ₁₃₁

Use case: Severity

Use case: Ambiguous reporting of onset-to-death delay distribution and erroneous CFR estimates

In a scenario in which the case fatality risk (CFR) needs to be calculated for an ongoing, growing disease outbreak an onset-to-death delay distribution is required to calculate an unbiased CFR estimate, due to some individuals being infected but theiry outcome (i.e recovery or death) is unknown **(author?)** [9].

A line list of the current outbreak is available, but no estiates of the onset-to-death delay are available for this outbreak and there is not enough case data to reliably estimate it from the line list. Therefore a previously inferred onset-to-death distribution is searched and extracted from the literature for the same pathogen from a past outbreak.

The paper reporting the onset-to-death states:

> "... the average duration between the time when symptoms first appeared and death of the patients was estimated. The mean onset-to-death delay was of 14.5 days, with a standard deviation of 6.7."

The ambiguous reporting of the esimates means the onset-to-death delay can be (mis)interpreted in several ways. The paper is reporting the summary statistics mean and standard deviation for a lognormal distribution they fitted to the data. The estimates could be misinterpreted as meanlog and sdlog do the lognormal distribution, or could be misinterpreted as the summary statistics of the raw data (i.e. sample statistics). The CFR calculation for an unbiased estimate requires a parametric probability density/mass function. Therefore, given the ambiguity we demonstrate the correct interpretation and three misinterpretations of the reported onset-to-death and show how the CFR varies as a result. We use the {simulist} and {cfr} R packages to simulate line list data and calculate the CFR, respectively **(author?)** [10] and **(author?)** [11].

Source: Article Notebook
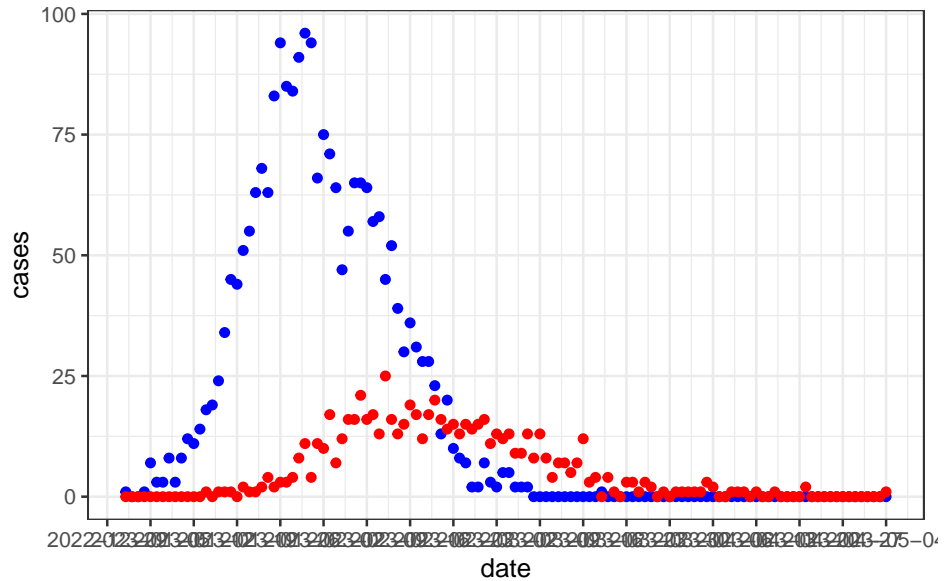
Source: Article Notebook

₁₃₂

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

The correct interpretation can analytically convert the mean and standard deviation to the lognormal distribution parameters ($\mu = 2.86$, $\sigma = 0.53$) and parameterise the onset-to-death, resulting in a CFR of 0.3042, or 30.42%. Misinterpreting the estimates to be the lognormal parameters results in an overestimated CFR of NA. Assuming that the reported estimates are sample summary statistics, the distribution can be assumed, here we test the assumption that it is a lognormal (correct assumption) and a gamma distribution (incorrect assumption). The assumed parametric distribution form can be used to simulate a sample and the same distribution can be fit to that sample to estimate the parameters. In the case of assuming a lognormal distribution the CFR is estimated as 0.2681, whereas assuming a gamma distribution results in a CFR of 0.2426. The estimated CFR is biased in both cases but more so when the distribution is assumed incorrectly.

Parameter uncertainty vs sample variability   The reporting of distributions is to encapsulate the variability of epidemiological delays, transmission, severity and others. However, there is also uncertainty around the parameters estimated. If it is not clearly stated that a distribution was fit to the data, it can be unclear whether the uncertainty around the mean corresponds to variation in the epidemiological case data, i.e. differences between individuals resulting in a distribution, or to the confidence or credible interval around the estimated mean.

Use case: Sample variability versus uncertainty

Issue: A measure of uncertainty is provided, but it is not clear whether this uncertainty is due to sample variability, or is an uncertainty around the estimator. Implication: Erronoeus attribution of uncertainty to sample variability or vice versa can bias the downstream analysis. This is particularly true when the underlying data are not shared.

Illustration: We illustrate this issue using simulated data from a gamma distribution D. We first simulate some data, get the observed mean and sd, and the uncertainty around these estimates. We then compare the observed data with that simulated with the erroneous interpretation of the uncertainty.
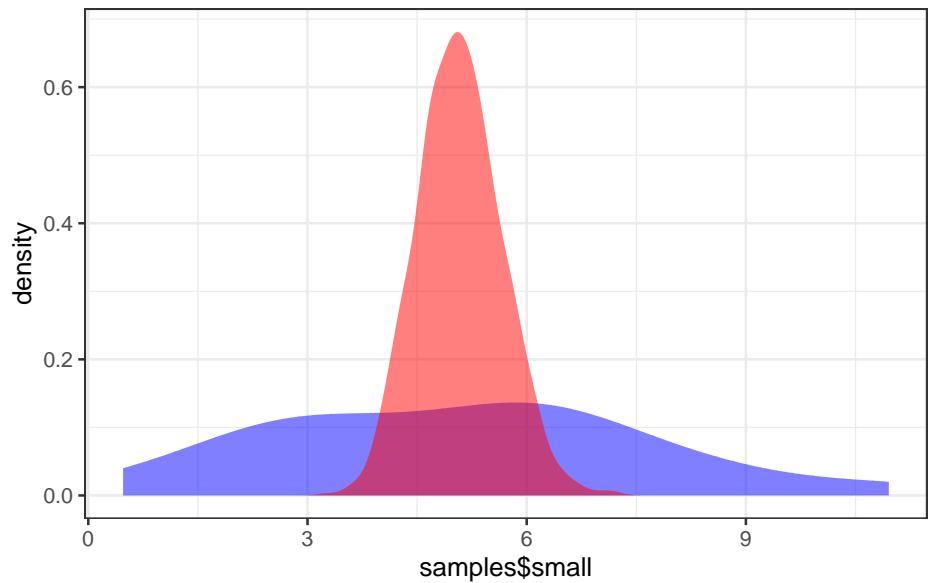
Source: Article Notebook

Reporting inference method The inference method used to infer epidemiological parameters and the uncertainty that is coupled with those estimates is also important to report precisely for it to be used in downstream analyses. One common distinction that can be made between inference method is whether it uses Maximum likelihood estimation (MLE) or Bayesian estimation. The resulting uncertainty of parameters, confidence intervals (CI) for MLE and credible intervals (CrI) for Bayesian fitting, cannot be interpreted or applied interchangeably [**?**]. Therefore if wanting to propagate uncertainty in parameter estimates, incorrectly treating CI as CrI or vice versa will lead to bias.

Reporting on Bayesian fitting also has several summary statistics to describe the central tendency of the inferred posterior sample, for example, mean, median, mode. It is beneficial for the specific central tendency statistic used to be explicitly stated.

Guidance:

- Report distribution parameters with uncertainty if available. This can be reporting the confidence interval (CI) or credible interval (CrI) making it clear
- When using Bayesian inference, specifying methods used for posterior distribution and making posterior sample openly available via data sharing platform, e.g., Zenodo, Data Dryad.

The types of distributions commonly fit to estimate delay distributions, such as serial interval, onset-to-event and incubation period, are Gamma, lognormal and Weibull. These are used as they are strictly positive (sometimes offsets or other distributions are used to account for negative serial intervals (Prete Jr. et al., 2021)) and are right-skewed, meaning that most of the distribution mass (i.e. area under the curve) is the left of the mean (Figure 1). It is best practice to fit multiple distributions

to the data and compare models using likelihoods, information criteria or likelihood ratio tests. By reporting these comparisons with each set of estimated parameters it allows others to use the distribution they choose while also being aware of the goodness-of-fit if choosing the non-best-fitting distribution.

Guidance:

- If multiple distributions are fit to the raw case data, report the goodness-of-fit (e.g. maximum likelihood, Akaike Information Criterion) and the parameter estimates of each distribution either in the main text or in the supplementary material.

Contextual information and metadata

Since pathogen transmission and spread is known to be affected by socioeconomic, demographic, and climatic factors, reporting relevant contextual information alongside parameter estimates is crucial to understand the circumstances in which these estimates were obtained. Doing so will allow external readers to make informed decisions about the generalisability of the reported parameters and usability in their own analyses.

An important contextual element is detailed information about the sample population from which parameters were estimated, including factors such as the geographic location, age distribution, and comorbidities. This is particularly relevant in those studies where only a specific subset of the population was sampled, such as health-care workers, pregnant individuals, or immunocompromised patients.

Other relevant details of the study, such as the type of design and sampling strategy, should not be overlooked, when reporting epidemiological parameters, as these provide relevant contextual information to assess the representativeness of the data and validity of the statistical methods applied. For instance, methods for estimating parameters like the serial interval require considering data collection methods, as different adjustments for biases are needed depending on whether data on transmission pairs was conducted prospectively or retrospectively (see section 1.5). The specific case definition used to estimate parameters should also be reported, where possible, given the range of clinical signs that many diseases exhibit at different stages of infection, which can have an impact on the estimation of parameters like the incubation period or delays from onset to outcome.

> **Use case: Seroprevalence**
>
> **Example use case: the importance of clear reporting of seroprevalence estimates**
> Estimates of seroprevalence provide critical insights into the level of susceptibility within a population, in turn informing the implementation of control measures, including vaccinations (e.g. through the critical fraction requiring vaccination to control spread).
> A novel coronavirus has been identified and is spreading throughout the population. A rapid seroprevalence study is undertaken to understand levels of immunity of the population. Initially, 5,000 tests are carried out via enzyme-linked immunoabsorbsent assay (ELISA), of which 250 are positive. This corresponds to a seroprevalence estimate of 5% (95% exact binomial confidence intervals 4.4% - 5.6%).
> However, it is known that ELISA assays can be prone to cross-reactivity with other coronaviruses, which are also in circulation in this population. Therefore, it is decided to undertake neutralisation tests, which are typically more sensitive, to

provide further confidence in the level of population-immunity. Due to a limited budget, only 500 tests can be re-tested with a neutralisation assay, and so all of the 250 positive tests and 250 of the 4,750 negatives are selected randomly for further testing.

Of the 500 samples sent for further testing, only 50 return as positive. For the neutralisation tests, in respect of the ELISA tests, this corresponds to a seroprevalence estimate of 10% (95% exact binomial confidence intervals 7.5% - 13%), which is notably higher than the seroprevalence obtained under ELISA only. However, discounting the results of the 4,500 tests incorrectly inflates estimated seroprevalence, implying a higher level of immunity in the population than that which is indicated by this study as a whole. Looking at the positive neutralisation tests out of the total tests (ELISA and neutralisation), seroprevalence is estimated as 1% (95% exact binomial confidence intervals 0.7% - 1.3%), again subtantially lower than suggested when using the neutralisation test denominator only.

Using the incorrect denominator could have important consequences for future planning and control strategies. Consider a population of 1,000,000 people. The basic reproduction number of this novel coronavirus is estimated at around 4, implying a herd immunity threshold (HIT) of 75%.

If it is (incorrectly) assumed that seroprevalence is 10%, this implies that 660,000 members of the population require immunity before herd immunity is reached. However, it is actually 740,000 members of the population requiring immunity to reach this threshold (an additional 90,000 people) under the correctly specified seroprevalence estimate of 1%.

If immunity were to be required through vaccination, then an additional 90,000 vaccines would be required. In the absence of a vaccine, e.g. if immunity were to be acquired via natural infection, assuming an overall case hospitalisation ratio (CHR) of 20% and case fatality ratio (CFR) of 1%, this suggests an additional 18,000 hospitalisations and 900 deaths than could be expected under the assumption of a seroprevalence of 10% respectfully. In at least the case of hospitalisation, this may require further preparation at hospital-level and the implementation of surge capacity protocols.

Source: Article Notebook

Where parameters are reported or inferred during an active outbreak, we recommend to provide information about the time into the outbreak since the first case was reported and epidemic phase at the time of the analysis, especially when inferring delay distributions (see section 1.5). Further contextual information is also relevant for a nuanced understanding of how parameter estimates may change throughout an outbreak, e.g., due to changes in containment measures, therapeutics and vaccination, or volume of testing. For instance, advancements in the therapeutic approach to critical care patients resulted in a significantly higher delay from onset to death for COVID-19 patients during the summer of 2020 (mean of 24.7 days), compared to the first wave of the pandemic (mean of 19.6 days) (Ward and Johnsen, 2021, PLoS).

Contextual information about the disease's causative agent should also be reported, including pathogen name, and, where applicable, its type, subtype and/or strain. This information is relevant, as the transmissibility, pathogenicity and severity of disease, and their resulting epidemiological parameters often vary across strains of the same pathogen. For instance, the incubation period for Influenza type A is reportedly longer on average than that of Influenza type B, with a median of 1.4 and 0.6 days, respectively (Lessler, 2009). If the causative agent is unknown to the authors, either due

to it being a novel pathogen, or simply because this information is not available, this ²¹⁷
should be explicitly stated in the publication. Beyond details about the causative agent, ²¹⁸
authors should also specify the transmission routes that have been considered when ²¹⁹
estimating parameters. This is particularly crucial for zoonotic and vector-borne ²²⁰
diseases, where explicit clarification is needed on whether the estimates account for ²²¹
human-to-human transmission only, or if animal-to-human or vector-to-human ²²²
transmission is also accounted for. ²²³

Guidance: ²²⁴

- Provide demographic information about the population and for the sample that ²²⁵
  was used to estimate epidemiological parameters. ²²⁶
- Specify whether analyses were stratified by certain groups, e.g., by age, or ²²⁷
  conducted using data for the whole population. ²²⁸
- Indicate whether reported parameters were obtained during an ongoing outbreak ²²⁹
  and, if so, provide information on the epidemic phase and time since outbreak was ²³⁰
  first declared. ²³¹
- Clearly state whether variant(s) of the pathogen of interest is known and, if so, ²³²
  report the name of the variant and how this was determined. ²³³
- Specify which transmission pathways of disease were considered, e.g., ²³⁴
  human-to-human only, or including animal-to-human transmission. ²³⁵

Open science and reproducibility to enhance reuse ²³⁶

The complexities involved in estimating and reporting epidemiological parameters mean ²³⁷
that it is unlikely that all methodological aspects and considerations can be documented ²³⁸
in the paper or even supplementary material. By sharing data and code it enables ²³⁹
reproducibility and auditing of the methods used. Sharing the code used to infer an ²⁴⁰
epidemiological parameter enables others to see which method, as well as any other ²⁴¹
packages that were used. There are several platforms that easily enable code sharing, ²⁴²
most common are GitHub, GitLab and BitBucket. To release the software used and ²⁴³
provide a unique identifier (e.g. DOI) services like Zenodo, Figshare and Dyrad, this ²⁴⁴
provides a single referenceable snapshot of the code, removing any issue if the code ²⁴⁵
changes on, for example GitHub. Following these and other good practices for code ²⁴⁶
sharing will help others navigate and review the code (Wilson et al., 2017). Openly ²⁴⁷
sharing code enables others to reproduce the estimates and verify the estimates. They ²⁴⁸
may also be able to assess the quality of the methods with respect to the available data ²⁴⁹
and possible bias-adjustments that may be required when working with real-time ²⁵⁰
outbreak data (citation needed). Sharing the analysis code can also resolve ambiguities ²⁵¹
in parameter reporting. If the parameterisation of the distribution is unclear from the ²⁵²
text (see Section 1) then by checking parameter arguments in the code clarifies their use. ²⁵³

Sharing the data is as important as sharing the code. By data we mean the input ²⁵⁴
data (i.e. outbreak case data) and output data (i.e. parameter estimates and fitting ²⁵⁵
metadata). If possible the raw data used to fit a model to estimate an epidemiological ²⁵⁶
parameter should be openly available. By sharing the raw data it enables ²⁵⁷
reproducibility of the analysis used to estimate the epidemiological parameters, but also ²⁵⁸
allows others to apply different models to the data. ²⁵⁹

Openly sharing epidemiological data can be restricted by personal identifiable ²⁶⁰
information (PII) and data usage restrictions. There are some methods available to ²⁶¹
enable reproducibility even when the raw data cannot be shared. Anonymisation, if the ²⁶²
personal identifiable information (PII) is not required by the method to infer the ²⁶³
epidemiological parameter then this information can be removed, de-identified or ²⁶⁴
anonymised prior to uploading the data (citations needed). Mock or synthetic data can ²⁶⁵

be generated which has the same characteristics as the empirical data. This enables the analysis to be reproduced while removing any risk of identification or leaking personal information.

The epidemiological parameter output should also be shared in full when possible. Often if the epidemiological parameter are distribution parameters these will be reported in the text. But the estimates correlation matrix, variance-covariance matrix, convergence metrics (e.g. ...) should be shared. For Bayesian analyses sharing the posterior distribution is most beneficial for reuse as it allows researchers to calculate whichever summary metric their use case requires (e.g. Highest Posterior Density (HPD) Interval).

Epidemiological parameter use and disjoint analysis pipelines

The aim of this paper has been to provide a set of reporting guidelines for epidemiological parameter, with the objective to make reusing them in other epidemiological analyses more reliable, with examples showcasing when analysis error can result from erroneous or ambiguous reporting. This argument is premised on the downstream epidemiological analysis being disjoint from the estimation of the epidemiological parameters, in other words the method that uses the parameters to estimate or infer another aspect of an outbreak does not estimate the parameters. An example of this is when an previously estimated generation time, or serial interval as a more commonly available replacement, is used to estimate the real-time reproduction number. If the data is available to jointly estimate the generation time or serial interval with the reproduction number, then this is the statistically optimal approach. However, for a variety of reasons, primarily model complexity of joint models leading to mathematical and computations simplification being required, the disjoint or 2-step analysis procedure is common (ref). Some models to offer joint estimation given sufficient data (ref). There have not been many studies exploring the statistical performance of joint versus disjoint estimation (check this sentence and find ref). There is another aspect to consider, whether a set of epidemiological parameters exists where the features of the data (e.g. sample size, collection procedure) make it more accurate than the available at hand. In this scenario even if a joint estimation framework is available and feasible, it might be better to choose estimated parameters. The contextual information of the data, such as demography, geography, and comorbidities of the sample, should also be considered in such a case as the two groups might not be epidemiologically equivalent. That is all to say that reporting guidelines are relevant due to the widespread use of disjoint estimation where clear, ambiguous reporting with coverage of key piece of statistical and contextual information are required.

Conclusion

References

References

1. Cori A, Kucharski A. Inference of epidemic dynamics in the COVID-19 era and beyond. Epidemics. 2024;48:100784. doi:10.1016/j.epidem.2024.100784.

2. Polonsky JA, Baidjoe A, Kamvar ZN, Cori A, Durski K, Edmunds WJ, et al. Outbreak analytics: a developing data science for informing the response to emerging pathogens. Philosophical Transactions of the Royal Society B: Biological Sciences. 2019;374(1776):20180276. doi:10.1098/rstb.2018.0276.

3. Van Kerkhove MD, Bento AI, Mills HL, Ferguson NM, Donnelly CA. A review of epidemiological parameters from Ebola outbreaks to inform early public health decision-making. Scientific Data. 2015;2(1):150019. doi:10.1038/sdata.2015.19.

4. Cuomo-Dannenburg G, McCain K, McCabe R, Unwin HJT, Doohan P, Nash RK, et al. Marburg virus disease outbreaks, mathematical models, and disease parameters: a systematic review. The Lancet Infectious Diseases. 2024;24(5):e307–e317. doi:10.1016/S1473-3099(23)00515-7.

5. Doohan P, Jorgensen D, Naidoo TM, McCain K, Hicks JT, McCabe R, et al.. Lassa fever outbreaks, mathematical models, and disease parameters: a systematic review and meta-analysis; 2024. Available from: http://medrxiv.org/lookup/doi/10.1101/2024.03.23.24304596.

6. Nash RK, Bhatia S, Morgenstern C, Doohan P, Jorgensen D, McCain K, et al. Ebola virus disease mathematical models and epidemiological parameters: a systematic review. The Lancet Infectious Diseases. 2024; p. S1473309924003748. doi:10.1016/S1473-3099(24)00374-8.

7. Park SW, Akhmetzhanov AR, Charniga K, Cori A, Davies NG, Dushoff J, et al.. Estimating epidemiological delay distributions for infectious diseases; 2024. Available from: http://medrxiv.org/lookup/doi/10.1101/2024.01.12.24301247.

8. Charniga K, Park SW, Akhmetzhanov AR, Cori A, Dushoff J, Funk S, et al.. Best practices for estimating and reporting epidemiological delay distributions of infectious diseases using public health surveillance and healthcare data; 2024. Available from: https://arxiv.org/abs/2405.08841.

9. Nishiura H, Klinkenberg D, Roberts M, Heesterbeek JAP. Early Epidemiological Assessment of the Virulence of Emerging Infectious Diseases: A Case Study of an Influenza Pandemic. PLoS ONE. 2009;4(8):e6852. doi:10.1371/journal.pone.0006852.

10. Lambert JW, Tamayo C. simulist: Simulate Disease Outbreak Line List and Contacts Data; 2024. Available from: https://zenodo.org/doi/10.5281/zenodo.10471458.

11. Gupte PR, Kucharski A, Russell T. cfr: Estimate disease severity and case ascertainment; 2024.