INFECTIOUS DISEASES

Getting the most out of noisy surveillance data

A recent study proposes a metric to quantify how much information different types of epidemiological surveillance data, such as case counts and death counts, convey about the real-time transmission of an epidemic.

Lauren McGough

HIV, Ebola, H1N1 influenza, SARS-CoV-1, SARS-CoV-2 and monkeypox. These are just a few examples of pathogen outbreaks that have generated widespread public attention over recent decades. When a pathogen outbreak enters the public awareness, everybody wants to know: what are the chances that I will get infected? Individuals need to know how much transmission is occurring in their communities so they can decide whether to adjust their behavior to decrease their risk. The COVID-19 pandemic has driven home an essential limitation to transmission estimation: our estimates are only as good as the data on which they are based. Because we never directly observe transmission events at the moment they occur, we must infer underlying transmission from imperfect proxies, such as case reports, hospitalization counts and death counts. Different sources of data can paint different pictures of the state of an epidemic; how closely these pictures reflect reality depends on the changing landscape of our measurement capacities and our evolving understanding of the pathogen of concern. Ideally, researchers would know which data sources to use to make the most accurate and precise inferences. However, it is difficult to choose among several data streams when each is subject to its own trade-offs. As reported in *Nature Computational Science*, Kris Parag and colleagues use analytic calculations informed by information theory to derive a calculable, interpretable and pathogen-agnostic metric that quantifies how much information different data streams convey about real-time transmission during an epidemic[1].

The study measures transmission using the instantaneous reproductive number R_t (ref. [2]). This number has an easily interpretable threshold at $R_t = 1$ — above this value the epidemic grows, whereas below this value the epidemic shrinks — and it has been frequently cited in public

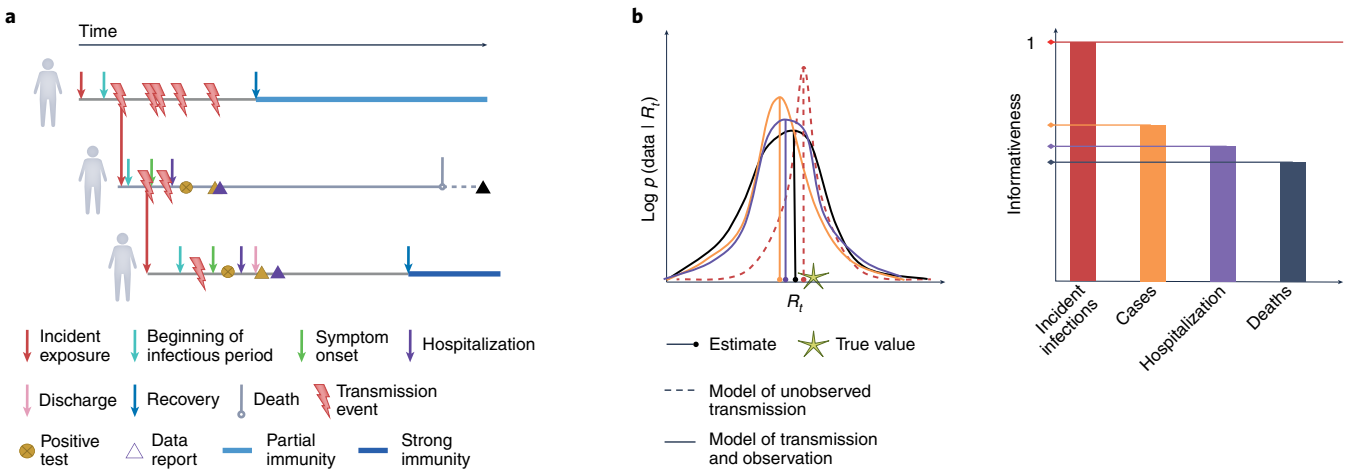

Fig. 1 | Information loss depends on the type of data collected. a, Schematic of three courses of infection that are represented differently across data sources. Lag times are measured from the red exposure arrow. The first individual has an asymptomatic infection, never tests positive, infects others, and is never ascertained by cases, deaths or hospitalizations. This individual maintains a low level of susceptibility by developing only partial immunity. The second individual is infected by the first individual, quickly develops symptoms, tests positive upon being hospitalized, and dies from the disease. The case and the hospitalization are reported near-simultaneously; the death report is significantly more lagged. The third individual is infected by the second individual, tests positive early, is briefly hospitalized, avoids transmitting upon discharge (for example, by isolating), and recovers. The positive test report is less lagged than the hospitalization report, and the infection is never reflected in death counts. This individual develops strong immunity and is subsequently well protected from future infection. The inter-individual differences in infection timelines lead to different amounts of information lost in the data collection, and the differential outcomes have the potential to complicate calculations of future transmission. **b**, Schematic relating the curvature of the log likelihood of different data sources to their informativeness. In the first panel, none of the maximum likelihood estimates exactly equals the true value, including the estimate derived from the incident infection model, as neither the transmission model nor any of the observation models perfectly reflects reality. The data-derived log likelihoods are distinct from one another and broader than that of the incident infection model. The second panel demonstrates that the lower-curvature likelihoods have lower informativeness. Informativeness of the incident infection curve is set to one, as Parag et al. normalize the informativeness of a data source against that of the incident infection model.

discourse as a measure of SARS-CoV-2 transmission. Despite its popularity, R_t is not straightforward to calculate[3]. The value of R_t equals the number of new (incident) infections at time t divided by the current infectiousness of the individuals that generated those infections, taking into account variability based on how long ago each individual was infected. Neither the total number of infectious individuals nor the number of incident infections at a specific time is directly observable; these must be estimated using measurements reported at later times. One strategy that researchers (including Parag et al.) use to infer R_t from data is to define two probabilistic processes: the transmission process by which incident infections arise and an observation process by which incident infections appear in the data. They manipulate these probabilities to compute the time series of R_t that most likely generated the observed data (the maximum likelihood estimate)[4].

Different forms of data are not equally informative because the distinct observation processes that produce them lose inequivalent information about infections (Fig. 1a). For example, the lags and imperfect ascertainment that prevent case counts from exactly encoding incident infections can be highly variable, as test availability changes and people take tests at different points in their infections. Death counts tend to be more consistently reported, with lags that depend more on biological factors (such as illness length) than external factors (such as test availability) compared to case counts, but the fraction of infections that death counts reflect is intrinsically limited by the infection fatality rate. The information lost in a measurement process directly corresponds to the uncertainty of the R_t estimate obtained from the data. Because we want our estimates of transmission to be as precise as possible, Parag et al. propose ranking different data sources according to the information they retain.

Parag and colleagues' major contribution is an analytic derivation of a simple, conceptually interpretable metric for this information loss. They begin with the standard maximum likelihood estimate of the time series of R_t values, and then they go beyond the estimate by computing its uncertainty in terms of the shape of the likelihood function. The authors creatively import techniques from information theory; more specifically, they define informativeness in terms of the curvature of the (log) likelihood function (Fig. 1b), a well-known quantity also known as the Fisher information[5]. Higher Fisher information values indicate less uncertainty. The expression they obtain is conceptually enlightening: it is expressed in terms of two quantities — the fraction of infections reported over time and the lag distribution — which contribute independently via single numbers computed from their respective distributions.

The authors demonstrate the importance of quantifying uncertainty to inform choices of data sources through an example: they show that when the infection fatality ratio is low, as for SARS-CoV-2, case data can be preferable to death data, whereas the opposite can hold for a pathogen with a higher infection fatality ratio, such as Ebola. This result runs counter to previous claims that death data is generally preferred for SARS-CoV-2 because deaths are more consistently measured and reported than cases[6]. Specifically, the authors calculate a bound relating the informativeness of death counts to that of case counts, showing that even under idealized assumptions about death data, when the infection fatality rate is smaller than the geometric mean of the case under-ascertainment rates over a period of time, deaths are less informative than cases.

One major challenge for directly using the Fisher information metric to evaluate different data sources as it stands is that measurement noise is often poorly characterized and non-stationary. Although Parag and colleagues demonstrate that their conclusions are robust to this form of uncertainty, the issue remains that highly uncertain or mis-specified noise can bias the R_t point estimates, and this effect could dominate informativeness considerations. Additional information theoretic analyses could address this caveat by quantifying the relative importance of these two effects; this would close the gap between being able to estimate the informativeness of a data source given a model of how the data were generated, and evaluating the usefulness of different data streams when we lack trustworthy estimates of their lag distributions and reporting fractions over time.

R_t will remain an important transmission metric as epidemics caused by diverse pathogens spread through diverse host populations. As such, Parag et al. have paved the way for going forward with more informative estimates of pathogen transmission when faced with imperfect data. ❐

Lauren McGough ✉

Department of Ecology and Evolution, The University of Chicago, Chicago, IL, USA.

✉*e-mail: mcgough@uchicago.edu*

Published online: 26 September 2022
https://doi.org/10.1038/s43588-022-00319-9

References

1. Parag, K. V., Donnelly, C. A. & Zarebski, A. E. *Nat. Comput. Sci.* https://doi.org/10.1038/s43588-022-00313-1 (2022).
2. Fraser, C. *PLoS ONE* **2**, e758 (2007).
3. Gostic, K. M. et al. *PLoS Comput. Biol.* **16**, e1008409 (2020).
4. Goldstein, E. et al. *Proc. Natl Acad. Sci. USA* **106**, 21825–21829 (2009).
5. Fisher, R. A. *Phil. Trans R. Soc. Lond. A* **222**, 309–368 (1922).
6. Flaxman, S. et al. *Nature* **584**, 257–261 (2020).

Competing interests
The author declares no competing interests.