ARTICLES
https://doi.org/10.1038/s43588-022-00313-1

nature computational science

Quantifying the information in noisy epidemic curves

Kris V. Parag[1,2] ✉, Christl A. Donnelly[2,3] and Alexander E. Zarebski[4]

Reliably estimating the dynamics of transmissible diseases from noisy surveillance data is an enduring problem in modern epidemiology. Key parameters are often inferred from incident time series, with the aim of informing policy-makers on the growth rate of outbreaks or testing hypotheses about the effectiveness of public health interventions. However, the reliability of these inferences depends critically on reporting errors and latencies innate to the time series. Here, we develop an analytical framework to quantify the uncertainty induced by under-reporting and delays in reporting infections, as well as a metric for ranking surveillance data informativeness. We apply this metric to two primary data sources for inferring the instantaneous reproduction number: epidemic case and death curves. We find that the assumption of death curves as more reliable, commonly made for acute infectious diseases such as COVID-19 and influenza, is not obvious and possibly untrue in many settings. Our framework clarifies and quantifies how actionable information about pathogen transmissibility is lost due to surveillance limitations.

The instantaneous reproduction number, denoted R_t at time t, is an important and popular temporal measure of the transmissibility of an unfolding infectious disease epidemic[1]. This parameter defines the average number of secondary infections generated by a primary one at t, providing a critical threshold for delineating growing epidemics ($R_t > 1$) from those likely to become controlled ($R_t < 1$). Estimates of R_t derived from surveillance data are widely used to evaluate the efficacies of interventions[1,2] (for example, lockdowns), forecast upcoming disease burden[3,4] (for example, hospitalizations), inform policy-making[5] and improve public health awareness[6].

The reliability of these estimates depends fundamentally on the quality and timeliness of the surveillance data available. Practical epidemic monitoring is subject to various errors or imperfections that can obscure or bias inferred transmission dynamics[7]. Prime among these are under-reporting and reporting delays, which can scale and smear R_t estimates, potentially misinforming public health authorities[8,9]. Under-reporting causes some fraction of infections to never be reported, while delays redistribute reports of infections incorrectly across time. The ideal data source for estimating R_t is the time series of new or incident infections, I_t.

Unfortunately, infections are difficult to observe directly and proxies such as reported cases, deaths, hospitalizations, prevalence and viral surveys from wastewater must be used to gauge epidemic transmissibility[5,10]. Each of these data streams provides a noisy approximation to the unknown I_t but with distinct and important relative advantages. We focus on the most popular ones: the epidemic curve of reported cases C_t at t, and that of death counts D_t, and investigate how their innate noise sources differentially limit R_t inference quality.

C_t records the most routinely available data, that is, counts of new cases[11], but is limited by delays and under-reporting. Ascertainment delays smear or reorder the case incidence and may emerge from fixed surveillance capacities, weekend effects and lags in diagnosing symptomatic patients (for example, the time from infection to a positive test)[8,12]. Delays may be classed as occurred but not yet reported (OBNR), when source times of delayed cases eventually become known (there delays cause right censoring of the case counts), or what we term never reported (NEVR), when source times of past cases are never uncovered[13–15].

Case under-reporting or underascertainment strongly distorts the true, but unknown, infection incidence curve, altering its size and shape[9,16]. Temporal fluctuations in testing intensity, behavior-based reporting (for example, by severity of symptoms)[17], undetected asymptomatic carriers and other surveillance bottlenecks can cause underascertainment or inconsistent reporting[18,19]. Constant reporting (CONR) describes the situation when the case detection fraction or probability is stable. We term the more realistic scenario in which this probability varies appreciably with time variable reporting (VARR).

D_t counts newly reported deaths attributable to the pathogen being studied and is also subject to under-reporting and reporting delays, but with two main differences[10]. First, death reporting delays incorporate an extra lag for the intrinsic time it takes an infection to culminate in mortality (this also subsumes hospitalization periods). Second, apart from the under-reporting fraction of deaths, there is another scaling factor known as the infection–fatality ratio (ifr), which defines the proportion of infections that result in mortality[1,20]. We visualize how the noise types underlying case and death curves distort infection incidence in Fig. 1.

Although the influences of surveillance latencies and underascertainment fractions on key parameters, such as R_t, are known[8,19,21,22] and much ongoing work attempts to compensate for these noise sources[10,23–25], there exists no formal framework for assessing and exposing how they inherently limit information available for estimating epidemic dynamics. Most studies utilize simulation-based approaches (with some exceptions, for example, refs. [9,22]) to characterize surveillance defects. While invaluable, these preclude generalizable insights into how epidemic monitoring shapes parameter inference.

Here we develop one such analytic framework for quantifying the information within epidemic data. Using Fisher information theory we derive a measure of how much usable information an

[1]NIHR Health Protection Research Unit in Behavioural Science and Evaluation, University of Bristol, Bristol, UK. [2]MRC Centre for Global Infectious Disease Analysis, Imperial College London, London, UK. [3]Department of Statistics, University of Oxford, Oxford, UK. [4]Department of Zoology, University of Oxford, Oxford, UK. ✉e-mail: kris.parag@bristol.ac.uk

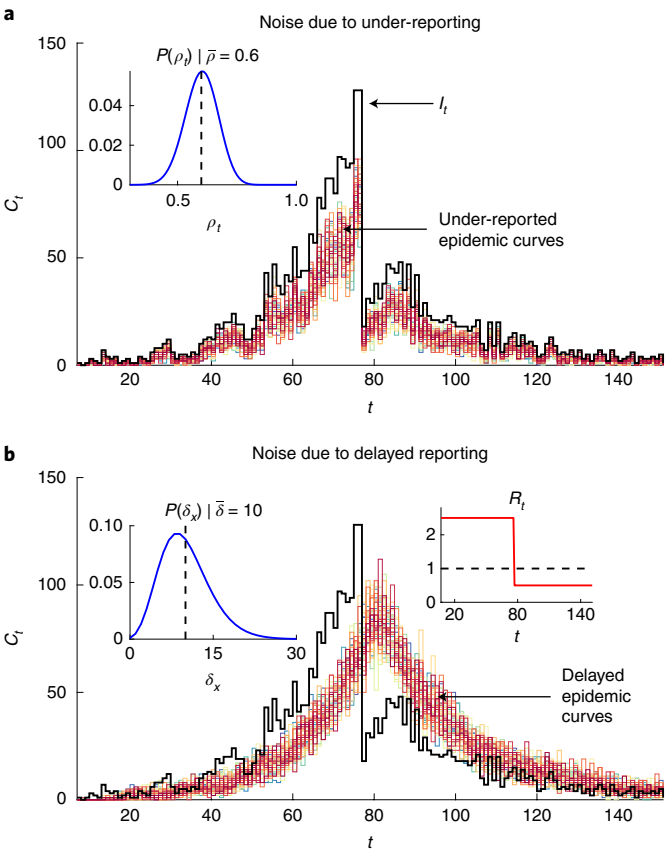

Fig. 1 | Under-reporting and delayed reporting noise. We simulate true infection incidence I_t over t (in days) from a renewal model (equation (9)) with Ebola virus dynamics) with reproduction number R_t that switches from supercritical to subcritical spread due to an intervention. **a**, Under-reported case curves (50 realizations, various colors) with reporting fractions sampled from the distribution $P(\rho_t)$ for sample fraction ρ_t plotted in the inset (mean sample fraction is $\bar{\rho}$). **b**, Delays in case reports (50 realizations, various colors) from the distribution $P(\delta_x)$ for delay δ_x (in days) plotted in the left-hand inset (mean delay is $\bar{\delta}$). We also provide the true R_t as the right-hand inset (red). The main question of this study is how we quantify which of scenarios **a** or **b** incurs the larger loss of the information originally available from I_t, ideally without simulation.

epidemic time series contains for inferring R_t at every time. This yields metrics for cross-comparing different types of surveillance time series, as we are able to explicitly quantify how under-reporting (both CONR and VARR) and reporting delays (exactly for OBNR with a tight upper bound for NEVR) degrade available information. As this metric only depends on the properties of surveillance (and not R_t or I_t), we extract simulation-agnostic insights into what are the least and most detrimental types of surveillance noise.

We prove for constrained mean reporting fractions and mean delays that it is preferable to minimize variability among reporting fractions but to maximize the variability of the reporting delay distribution such that a minority of infections face large delays but the majority possess short lags to notification. This proceeds from standard experimental design theory applied to our metric, which shows that the information embedded within an epidemic curve depends on the product of the geometric means of the reporting fractions and cumulative delay probabilities corrupting this curve. This central result also provides a non-dimensional score for summarizing and ranking the reliability of (or uncertainty within) different surveillance data for inferring pathogen transmissibility.

Finally, we apply this framework to explore and critique a common claim in the literature, which asserts that death curves are more robust for inferring transmissibility than case curves. This claim is usually made for acute infectious diseases such as COVID-19 and pandemic influenza[1,20], where cases are severely under-reported, with symptom-based fluctuations in reporting. In such settings it seems plausible to reason that deaths are less likely undercounted and more reliable for R_t inference. However, we compute our metrics using COVID-19 reporting rate estimates[18,26] and discover few instances in which death curves are definitively more informative or reliable than case counts.

While this may not rule out the possibility of having a more reliable death time series, it elucidates and exposes how the different noise terms within the two data sources corrupt information and presents a methodology for exploring these types of questions more precisely. We illustrate how to compute our metrics practically using empirical COVID-19 and Ebola virus disease (EVD) noise distributions and outline how other common data such as hospitalizations, prevalence and wastewater virus surveys conform to our framework. Hopefully the tools we develop here will improve quantification of noise and information and highlight key areas where enhanced surveillance strategies can maximize impact.

Results

Methods overview. We summarize the salient points from Methods and outline the main arguments that underpin all subsequent Results sections. Our analysis is centered on the renewal model[27], which is widely applied to describe the dynamics of epidemics of COVID-19, EVD, influenza, dengue, measles and many others[21]. This model posits that I_t are Poisson (Pois) distributed with a mean that is the product of R_t and total infectiousness (Λ_t). Here Λ_t defines how past infections engender new ones on the basis of $\mathbf{w}$, the generation time distribution of the pathogen. In equation (9) and Supplementary Table 1 we provide precise definitions of these variables (as well as others later in the text).

An important problem in infectious disease epidemiology is the estimation of R_t across the duration of an epidemic[5]. However, as infections cannot be observed, we commonly have to infer R_t from noisy proxies such as the time series of reported cases or deaths. These can be described by generalized renewal models that include terms for practical noise sources such as under-reporting and delays in reporting[28]. We define these models in equations (1) and (2) and detail the properties of various noise sources in Methods. Our aim is to understand and quantify how much information for inferring R_t, as a fraction of what would be available if infections were observable, can be extracted from these proxies.

We pursue this aim by adapting concepts from statistical theory and information geometry. We first construct the log-likelihood function of the parameter time series $R_1^\tau := \{R_t : 1 \le t \le \tau\}$, with τ as the present or last observation time and t scaled in units (for example, weeks) so that each R_t can be assumed independent. This function is $\ell(R_1^\tau) = \sum_{t=1}^{\tau} \ell(R_t)$ with $\ell(R_t) := \log \mathbb{P}(I_t \mid R_t)$ computed from the Poisson distribution of the renewal model. Equation (10) results and admits the maximum likelihood estimates (MLEs) $\hat{R}_t$ for all t as its maxima. The reliability of these MLEs is characterized by the Fisher information (FI) of R_1^τ from the time series or curve of incident infections $I_1^\tau := \{I_t : 1 \le t \le \tau\}$. We also often compute FI values under the robust transform $\mathcal{R}_t = 2\sqrt{R_t}$, which has useful statistical properties[29].

Larger FI values imply smaller asymptotic uncertainty around the MLEs[30]. We obtain $\mathbb{F}_I(R_t)$, the FI of R_t, in equation (11) by evaluating the average curvature of the log-likelihood function. We then formulate the total information, $\mathbb{T}(I_1^\tau)$, as a product of $\mathbb{F}_I(R_t)$ terms across t as equation (12). This follows from the independence of the R_t variables and is a measure of the reliability of the infection time series. It is also delimits the maximum possible precision around the

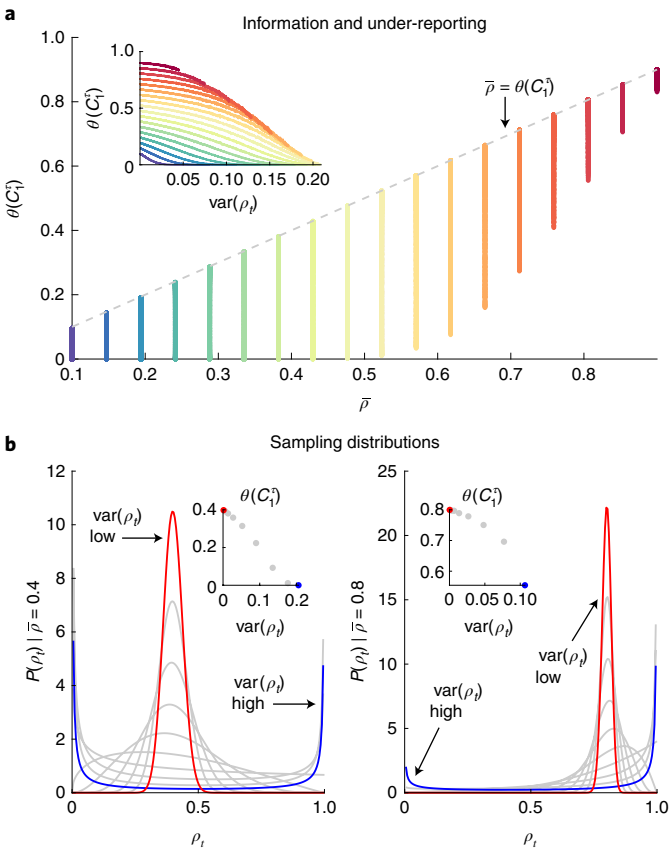

Fig. 2 | The information loss in under-reporting. We investigate the effective information metric ($\theta(C_1^\tau)$) for VARR strategies with reporting fraction ρ_t drawn from Beta distributions with different shapes. Smaller values of $\theta(C_1^\tau)$ indicate more substantial information losses. **a**, Changes in $\theta(C_1^\tau)$ with the mean reporting probability ($\bar{\rho}$) and its variance var(ρ_t) (inset, where each color indicates the various schemes with a given $\bar{\rho}$). The gray line (dashed) is the optimal CONR protocol. **b**, The Beta sampling distributions and their resulting var(ρ_t) and $\theta(C_1^\tau)$ (inset). The most and least variable reporting strategies for a given $\bar{\rho}$ are in blue and red, respectively.

MLEs of R_t for any time series. Since I_1^τ is often unobservable, $\mathbb{T}(I_1^\tau)$ is generally not computable and is a theoretical maximum. However, our subsequent results circumvent this issue.

Using the models of equations (1) and (2), we employ this same recipe of constructing a log-likelihood and computing MLEs and FI values, but now for practical time series or data streams that are corrupted by under-reporting and delays. This yields equations (13)–(18), which contain the ingredients for deriving the total information in case, death and any other incidence data that is related to I_1^τ via a generalized renewal model (this includes prevalence, hospitalizations and virus abundance found in wastewater). We derive a key result for $\mathbb{T}(C_1^\tau)$ in equation (3), showing exactly how case data $C_1^\tau := \{C_t : 1 \le t \le \tau\}$ cause a loss in R_t estimate reliability.

Building on this expression we develop metrics $\eta(C_1^\tau)$ and $\theta(C_1^\tau)$ in equations (4) and (5), which effectively quantify $\frac{\mathbb{T}(C_1^\tau)}{\mathbb{T}(I_1^\tau)}$ that is, the level of informativeness of case data relative to true infections. The smaller these metrics are, the more information is lost due to surveillance noise. Importantly, these metrics are analytic, require no knowledge of I_1^τ or the generation time distribution (both are difficult to observe) and are interpretable, since each noise type contributes a separate geometric mean term. Further, they play an integral role in defining the statistical complexity of the generalized renewal model describing that time series, as we find in equation (6).

Repeating the above recipe we derive analogous metrics for death data $D_1^\tau := \{D_t : 1 \le t \le \tau\}$, by characterizing the ratio $\frac{\mathbb{T}(D_1^\tau)}{\mathbb{T}(I_1^\tau)}$ in equations (7) and (8). We can similarly compute ratios for hospitalizations, prevalence and viral wastewater data by inputting appropriate delay and under-reporting terms. We complete our results by including empirical estimates of case and death noise sources within our framework to compare $\frac{\mathbb{T}(C_1^\tau)}{\mathbb{T}(I_1^\tau)}$ and $\frac{\mathbb{T}(D_1^\tau)}{\mathbb{T}(I_1^\tau)}$ for COVID-19 and EVD and hence determine whether case or death data are likely more reliable for inferring R_1^τ.

Renewal models with noisy observations. We denote the empirically observed or reported number of cases at time step or unit t, subject to noise from both under-reporting and reporting delays, as C_t with C_1^τ as the epidemic case curve. This curve is obtained from routine outbreak surveillance and is a corrupted version of the true incidence I_1^τ (ref. [10]), modeled by equation (9). These noise sources (see Methods for statistical descriptions) are parametrized by reporting fractions $\rho_1^\tau := \{\rho_t : 1 \le t \le \tau\}$ and a delay distribution $\boldsymbol{\delta} := \{\delta_x : x \ge 0\}$. Here ρ_t is the fraction of infections reported as cases at t and δ_x the probability of a lag from infection time to case report of x units.

We assume that these noise sources are estimated from auxiliary line-list or contact tracing data[12,31]. As a result, we can construct equation (1) as in ref. [25] (Methods). Note that if noise source estimates are unavailable then R_1^τ becomes statistically non-identifiable or ill-defined.

$$C_t \sim \text{Pois}\left(\sum_{x=1}^{t} \delta_{t-x}\rho_x \Lambda_x R_x\right). \quad (1)$$

This noisy renewal model suggests that C_t (unlike I_t) contains partial information about the entire time series of reproduction numbers for $x \le t$ as mediated by the delay and reporting probabilities. Perfect reporting corresponds to $\rho_x = 1$ for all x, $\delta_0 = 1$ (with all other $\delta_x = 0$) and means $C_t \to I_t$. Models (1) and (2) in Methods are obtained by individually removing noise sources from equation (1).

Other practical epidemic surveillance data such as the time series of new deaths or hospitalizations conform to the framework in equation (1) either directly or with additional effective delay and under-reporting stages[20]. The main one we investigate here is the count of new deaths (due to infections) across time, which we denote D_1^τ. The death curve involves a reporting delay that includes the intrinsic lag from infection to death. We let $\boldsymbol{\gamma} := \{\gamma_x : x \ge 0\}$ represent the distribution of lag from infection to observed death and $\sigma_1^\tau := \{\sigma_t : 1 \le t \le \tau\}$ be the fraction of deaths that is reported.

An important additional component when describing the chain from I_1^τ to D_1^τ is ifr_t, which is the probability at t that an infection culminates in a death event[10]. Fusing these components yields as a model for D_t

$$D_t \sim \text{Pois}\left(\sum_{x=1}^{t} \text{ifr}_x \gamma_{t-x}\sigma_x \Lambda_x R_x\right). \quad (2)$$

In a later section we explain how analogs of equations (1) and (2) also fit other data streams such as hospitalizations and prevalence. Some studies[1,32] replace this Pois formulation with a negative binomial (NB) distribution to model extra variance in these data. In Supplementary Notes we show that this does not disrupt our subsequent results on the relative informativeness of surveillance data, but the NB formulation is less tractable and unsuitable for extracting generalizable, simulation-free insights.

Reliability measures for surveillance data. We analyze the information in the generalized renewal models of equations (1) and (2) by computing the FI for each R_t or its transform $\mathcal{R}_t = 2\sqrt{R_t}$

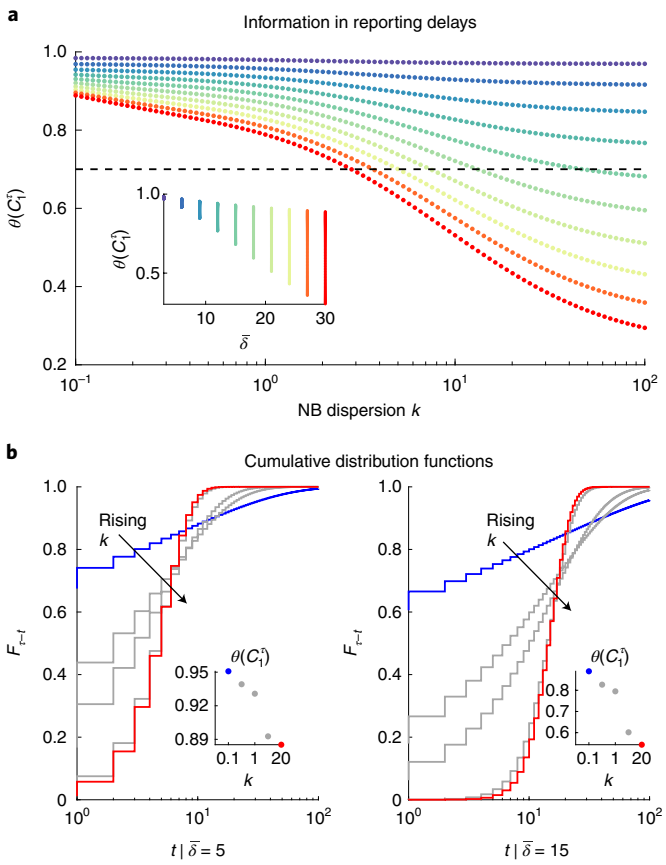

Fig. 3 | The information loss from delays. We compute $\theta(C_1^\tau)$ for delay distributions that are NB with various dispersions k. Smaller values of $\theta(C_1^\tau)$ indicate more substantial information losses, while smaller k indicates a more dispersed or variable reporting delay distribution. **a**, Influence of mean delay ($\bar{\delta}$) and k on $\theta(C_1^\tau)$, with colored curves representing different $\bar{\delta}$ and the black dashed line as a reference value. Inset: variations in $\theta(C_1^\tau)$ at a given $\bar{\delta}$ (matching colors) due to the dispersion k of the reporting delay distribution. **b**, Influence of the shape of the cumulative delay distributions, $F_{\tau-t}$, at different k (increasing from blue to red) on our metric $\theta(C_1^\tau)$ (inset with corresponding colors).

(see Methods for details). This is denoted $\mathbb{F}_C(.)$ for case data, and as derived in Methods (equation (15)) allows us to obtain the total information, $\mathbb{T}(C_1^\tau)$, contained in those data about the parameter time series R_1^τ or $\mathcal{R}_1^\tau$ as a product across t of $\mathbb{F}_C(.)$ terms. This total information, $\mathbb{T}(C_1^\tau)$, relates inversely to the smallest joint uncertainty around unbiased estimates of all our parameters[33]. As larger $\mathbb{T}(C_1^\tau)$ implies reduced overall uncertainty, this is a rigorous measure of the statistical reliability of noisy data sources for inferring pathogen transmissibility.

We first consider the OBNR delay case under VARR reporting rates. Since the FI matrix under OBNR delays is diagonal, with each element given by equation (15), we can adapt equation (12) to derive

$$\mathbb{T}(C_1^\tau) = \prod_{t=1}^{\tau} \sqrt{\mathbb{F}_C(\mathcal{R}_t)} = \prod_{t=1}^{\tau} \sqrt{\rho_t F_{\tau-t} \Lambda_t}. \tag{3}$$

Here we have applied the $\mathcal{R}_t = 2\sqrt{R_t}$ transformation to show that the total information in this noisy stream can be obtained without knowing R_t. In the absence of this transform we would have $\prod_{t=1}^{\tau} \sqrt{\rho_t F_{\tau-t} \Lambda_t R_t^{-1}}$.

Since C_1^τ is a distortion of the true infection incidence I_1^τ we normalize equation (3) by equation (12) to develop a reliability metric, $\eta(C_1^\tau) := \mathbb{T}(C_1^\tau)\mathbb{T}(I_1^\tau)^{-1}$. This is given in the following equation and valid under both R_t and $\mathcal{R}_t$.

$$0 \le \eta(C_1^\tau) = \prod_{t=1}^{\tau} \sqrt{\rho_t F_{\tau-t}} \le 1. \tag{4}$$

We can relate this reliability measure to a fixed, effective reporting fraction, $\theta(C_1^\tau)$, which causes an equivalent information loss. Applying equation (4), we obtain $\eta(C_1^\tau) = \sqrt{\theta(C_1^\tau)^\tau}$, which yields the following equation. Here, $\mathbb{G}(.)$ indicates the geometric mean of its arguments over $1 \le t \le \tau$.

$$\theta(C_1^\tau) = \prod_{t=1}^{\tau} \sqrt[\tau]{\rho_t F_{\tau-t}} = \mathbb{G}(\rho_t)\mathbb{G}(F_{\tau-t}). \tag{5}$$

Equation (5) is a central result of this work. It states that the total information content of a noisy epidemic curve is independently modulated by the geometric mean of its reporting fractions, $\mathbb{G}(\rho_t)$, and that of its cumulative delay probabilities, $\mathbb{G}(F_{\tau-t})$. Moreover, equation (5) provides a framework for gaining analytic insights into the separate influences of both noise sources from different surveillance data and for ranking the overall quality of these diverse data. For example, from the properties of geometric means, we know that $\mathbb{G}(.)$ is bounded by the smallest and largest noise terms across t. Importantly, equation (5) has no dependence on Λ_t, which is generally unknown and sensitive to difficult-to-infer changes in the generation time distribution[34].

Equation (5) applies to OBNR delays exactly and upper bounds the reliability of data streams with NEVR delays (see Methods for derivation). Tractable results for NEVR delays are not possible and would necessitate numerical computation of Hessian matrices of $-\log \mathbb{P}(C_1^\tau \mid R_1^\tau)$ (we outline the log-likelihoods and other equations for a $\tau = 3$ example in Supplementary Notes). However, we find that the equation (5) upper bound is tight for two elementary settings. The first is under a constant or deterministic delay of d, that is, $\delta_{x=d} = 1$. Equation (1) reduces to $C_t \sim \text{Pois}(\rho_{t-d}\Lambda_{t-d}R_{t-d})$. As each C_t only informs on R_{t-d}, OBNR and NEVR delays are the same and are corrected by truncation. Degenerate delays such as these can serve as useful elements for constructing complex distributions[35].

The second occurs when transmissibility is constant or stable, that is, $R_t = R$ for all t. This applies to inferring the basic reproduction number (R_0) during initial phases of an outbreak[5]. We can sum equation (15) to obtain $\mathbb{F}_C(R) = \sum_{t=1}^{\tau} \rho_t F_{\tau-t} \Lambda_t R^{-1}$ for OBNR delays. We can calculate the FI for NEVR delays from equation (16), which admits a derivative $\frac{\partial \ell(R)}{\partial R} = \sum_{t=1}^{\tau} C_t R^{-1} - F_{\tau-t}\rho_t\Lambda_t$ and hence an FI and MLE that are precisely equal to those for OBNR delays (derived in Methods). This proves a convergence in the impact of two fundamentally different delay noise sources and emphasizes that noise has to be contextualized with the complexity of the signal to be inferred. Simpler signals, such as a stationary R that remains robust to the shifts and reordering of I_1^τ caused by delays, may be notably less susceptible to fluctuations in noise probabilities.

Ranking noise sources by their information loss. The metric proposed in equation (5) provides a general framework for scoring proxies of incidence (for example, epidemic case curves, death counts, hospitalizations and others) using only their noise probabilities and without the need for simulations. We explore the implications of equation (5) for both understanding noise and ranking these proxies. The geometric mean decomposition allows us to separately dissect the influences of under-reporting and delays. We start by

applying experimental design theory[36,37] to characterize the best and worst noise types for inferring effective reproduction numbers.

We consider $\mathbb{G}(\rho_t)$, the geometric mean of the reporting probabilities across time. If we assume the mean sampling fraction $\bar{\rho} = \frac{1}{\tau}\sum_{t=1}^{\tau}\rho_t$ is fixed (for example, by some overall surveillance capacity) then we immediately know from design theory that $\bar{\rho} = \arg\max_{\boldsymbol{\rho}} \mathbb{G}(\rho_t)$. This means that of all the possible distributions of sampling fractions fitting that constraint, $\boldsymbol{\rho}$, CONR with probability $\bar{\rho}$ is the most informative[38]. This result supports earlier studies recognizing that CONR is preferred to VARR, although they investigate estimator bias and not information loss[9,21].

Accordingly, we also discover that the worst sampling distribution is maximally variable. This involves setting $\rho_t \sim 1$ for some time subset $\mathbb{S}$ such that $\sum_{t\in\mathbb{S}}\rho_t = \tau\bar{\rho}$ with all other $\rho_t \sim 0$ (we use approximate signs as we assume non-zero sampling probabilities). Relaxing this constraint, equation (5) presents a framework for comparing different reporting protocols. We demonstrate these ideas in Fig. 2, where $\rho_t \sim \text{Beta}(a, b)$, that is, each reporting fraction is a sample from a Beta distribution. Reporting protocols differ in (a, b) choices. We select 10^4 ρ_t samples from each of 2,000 distributions with $10^{-1} \le b \le 10^2$ and a computed to fulfill the mean constraint $\bar{\rho}$. Variations in the resulting $\theta(C_1^\tau)$ metrics indicate the influence of reporting fraction uncertainties under this mean.

Figure 2a shows that $\theta(C_1^\tau)$ generally increases with the mean reporting probability $\bar{\rho}$. However, this improvement can be denatured by the variance, $\text{var}(\rho_t)$, of the reporting scheme (inset, where each color indicates the various schemes with a given $\bar{\rho}$). The CONR scheme is outlined with a grey line (dashed), and as derived is the most informative. Figure 2b confirms our theoretical intuition on how $\text{var}(\rho_t)$ reduces total information, with the extreme (worst) sampling scheme outlined above in blue and the most stable protocol in red. There are many ways to construct ρ_t protocols. We chose Beta distributions because they can express diverse reporting probability shapes using only two parameters.

Similarly, we investigate reporting delays via $\mathbb{G}(F_{\tau-t})$, the geometric mean of the cumulative delay or latency distribution across time. Applying a mean delay constraint $\bar{\delta} = \sum_{x\ge 0} x\delta_x = \sum_{t=1}^{\tau}(1 - F_{\tau-t})$ (for example, reflecting operational limits on the speed of case notification), we adapt experimental design principles. As we effectively maximize an FI determinant (see derivation of equation (5)) our results are termed D optimal[38]. These suggest that $\max_{\boldsymbol{\delta}} \mathbb{G}(F_{\tau-t})$ is achieved by cumulative distributions with the most uniform shape. These possess the largest δ_0 within this constraint. Delay distributions with substantial dispersion (for example, heavy tails) attain this optimum while fixed delays (where $\delta_{x\sim\bar{\delta}} = 1$ and 0 otherwise) lead to the largest information loss under this constraint.

This may seem counterintuitive, as deterministic delays best preserve information outside of that delay and can be treated by truncating the observed epidemic time series: for example, for a fixed weekly lag we can ignore the last week of data. However, this causes a bottleneck. No information is available for that truncated week, eliminating any possibility of timely inference (and making epidemic control difficult[39]). In contrast, a maximally dispersed delay distribution slightly lags the majority of cases, achieving the mean constraint with large latencies on a few cases. This ensures that, overall, we gain more actionable information about the time series.

We illustrate this point in Fig. 3, where we verify the usefulness of equation (5) as a framework for comparing the information loss induced by delay distributions of various shapes and forms. We model $\boldsymbol{\delta}$ as $\text{NB}(k, \frac{\bar{\delta}}{\delta+k})$, with k describing the dispersion of the delay. Figure 3a demonstrates how our $\theta(C_1^\tau)$ metric varies with k (30 values taken between 10^{-1} and 10^2) at various fixed mean constraints ($3 \le \bar{\delta} \le 30$, each given as a separate color). In line with the theory, we find that decreasing k (increasing dispersion of the delay distribution) improves information at any given $\bar{\delta}$.

The importance of both the shape and mean of reporting delays is indicated in the inset as well as by the number of distributions (seen as intersects of the dashed black line) that result in the same $\theta(C_1^\tau)$. Figure 3b plots corresponding cumulative delay probability distributions, validating our assertion from design theory that the best delays (blue, with metric in inset) are dispersed, forcing the cumulative probability of reporting delays up to $\tau - t$ time units ($F_{\tau-t}$) high very early on (maximizing δ_0 and leading to the most uniform shape). In contrast, the worst delay distributions are more deterministic (red, larger k). These curves are for OBNR delays and upper bound the performance expected from NEVR delays except for the settings described in the previous section, where the two types coincide.

Comparing different epidemic data streams. Our metric (equation (5)) not only allows the comparison of different under-reporting schemes and reporting delay protocols (Ranking noise sources by their information loss) but also provides a common score for assessing the reliability or informativeness of diverse data streams for inferring R_1^τ. The best stream, from this information theoretic viewpoint, maximizes the product of the geometric means $\mathbb{G}(.)$ of the $F_{\tau-t}$ and ρ_t. Many common surveillance data types used for inferring pathogen transmissibility have been modeled within the framework of equation (1) and therefore admit related $\theta(.)$ metrics. Examples include time series of deaths, hospitalizations, the prevalence of infections and incidence proxies generated from viral surveys of wastewater.

We detail death count data in the next section but note that its model, given in equation (2), is a simple extension of equation (1). Hospitalizations may be described similarly with the ifr term replaced by the proportion of infections hospitalized and the intrinsic delay distribution now defining the lag from infection to hospital admission[5]. The infection prevalence conforms to equation (1) because it can be represented as a convolution of the infections with a duration of infectiousness distribution, which essentially contributes a reporting delay[40]. Viral surveys also fit equation (1). They offer a downsampled proxy of incidence, which is delayed by a shedding load distribution defining the lag before infections are detected in wastewater[41]. Consequently, our metrics are widely applicable.

While in this study we focus on developing methodology for estimating and contrasting the information from the above surveillance data, we find that our metric is also important for defining the complexity of a noisy renewal epidemic model. Specifically, we rederive equation (5) as a key term of its description length (L). Description length theory evaluates the complexity of a model from how succinctly it describes its data (for example, in bits)[33,42]. This measure accounts for model structure and data quality and admits the approximation $L_C \sim -\ell(\hat{R}_1^\tau) + \frac{p}{2}\log\frac{m}{2\pi} + \log\int \det\left[\frac{1}{m}\mathbb{F}_C(R_1^\tau)\right] \mathrm{d}R_1^\tau$. Here the first term indicates model fit by assessing the log-likelihood at our MLEs $\hat{R}_1^\tau$. The second term includes data quality through the number of parameters (p) and data size (m). The final term defines how model structure shapes complexity with the integral across the parameter space of R_1^τ.

This formulation was adapted for renewal model selection problems in ref. [43] assuming perfect reporting. We extend this and show that our proposed total information $\mathbb{T}(C_1^\tau)$ plays a central role. Given some epidemic curve C_1^τ we can rewrite the previous integral as $-\frac{p}{2}\log m + \log\prod_{t=1}^{\tau}\int\sqrt{\mathbb{F}_C(R_t)}\,\mathrm{d}R_t$ and observe that $m = p = \tau$. It is known that under a robust transform such as $\mathcal{R}_t = 2\sqrt{R_t}$ this integral is conserved[33,37]. Consequently, $\int\sqrt{\mathbb{F}_C(R_t)}\,\mathrm{d}R_t = \sqrt{\mathbb{F}_C(\mathcal{R}_t)}\int_0^{2\sqrt{R_{\max}}} 1\,\mathrm{d}\mathcal{R}_t$ with $R_{\max}$ as some maximum value that every R_t can take. Combining these expressions we obtain the following equation, highlighting the importance of our total information metric.

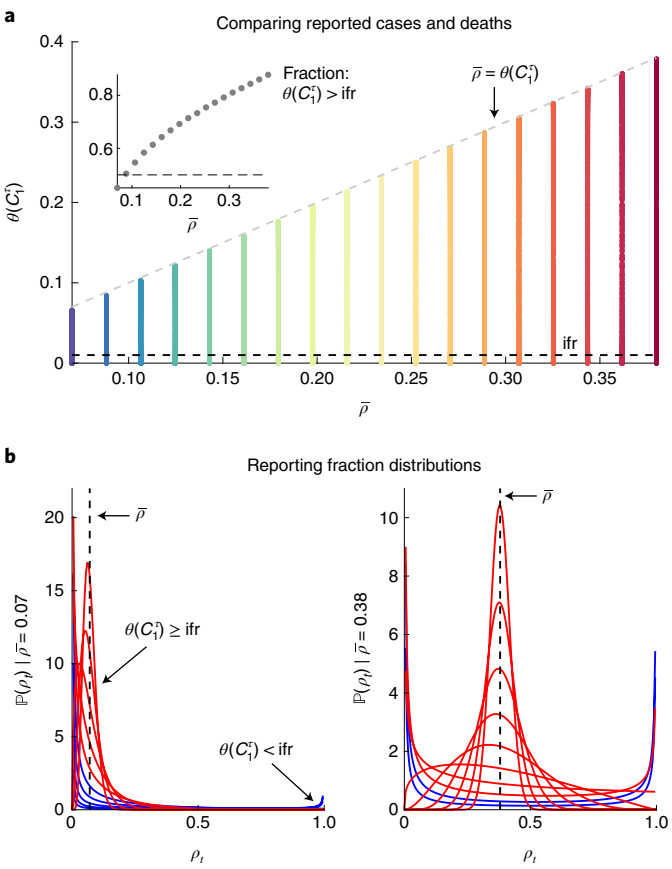

Fig. 4 | The relative information in epidemic case data and death counts. Using $\theta(C_1^\tau)$ we compare the information in case curves C_1^τ and death counts D_1^τ under assumptions that lead to equation (8). We examine various case reporting strategies parametrized as Beta distributions with $\bar{\rho}$ from 0.07 to 0.38 (ref. [18]) and compare the resulting $\theta(C_1^\tau)$ against the equivalent from deaths (which reduces to just the infection–fatality ratio, $\theta(D_1^\tau) = \text{ifr}$). **a**, $\theta(C_1^\tau)$ for reported case data at different $\bar{\rho}$ (each color represents a fixed $\bar{\rho}$) as compared with ifr (black dashed threshold). The best reporting strategy is in gray. Inset: proportion of case reporting distributions from the main plot for which $\theta(C_1^\tau) > \text{ifr}$. **b**, Those distributions $\mathbb{P}(\rho_t)$ at the ends of the empirical $\bar{\rho}$ range with red indicating when $\theta(C_1^\tau) > \text{ifr}$.

$$L_C \sim -\ell(\hat{R}_1^\tau) + \frac{\tau}{2}\log\frac{2R_{\max}}{\pi} + \log\mathbb{T}(C_1^\tau). \tag{6}$$

If we have two potential data sources for inferring R_1^τ then we should select the one with the smaller L_C value. Since the middle term in equation (6) remains unchanged in this comparison, the key points when comparing model complexity relate to the level of fit to the data and the total FI of the model given those data[42]. Our metrics therefore play a central role when comparing different data streams.

Are COVID-19 deaths or cases more informative? In the above sections we developed a framework for comparing the information within diverse but noisy data streams. We now apply these results to better understand the relative reliabilities of two popular sources of information about transmissibility R_1^τ: the time series of new cases C_1^τ and of new death counts D_1^τ. Both data streams have been extensively used across the ongoing COVID-19 pandemic to better characterize pathogen spread[5]. Known issues stemming from fluctuations in the ascertainment of COVID-19 cases[18,19] have motivated some studies to assert D_1^τ as the more informative and hence trustworthy data for estimating R_1^τ (refs. [1,20]).

These works have reasonably assumed that deaths are more likely to be reliably ascertained. Case reporting can be substantially biased by testing policy inconsistencies and behavioral changes (for example, symptom-based healthcare seeking). In contrast, given their severity, deaths should be less likely to be underascertained[5]. However, no analysis, as far as we are aware, has explicitly tested this assumption. Here we make some progress towards better comprehending the relative merits of the two data streams. We start by computing ratios of our metric in equation (5) for both C_1^τ and D_1^τ via equations (1) and (2).

This results in $\theta(C_1^\tau) = \mathbb{G}(\rho_t)\mathbb{G}(F_{\tau-t})$ for cases and, by analogy, $\theta(D_1^\tau) = \mathbb{G}(\sigma_t\text{ifr}_t)\mathbb{G}(H_{\tau-t})$ for deaths. In the same way that ρ_t defines the proportion of infections reported as cases, the product $\sigma_t\text{ifr}_t$ defines the proportion of infections that are reported as deaths. This follows because ifr_t is the fraction of infections that engender deaths and σ_t is the proportion of those deaths that are reported. $H_{\tau-t} := \sum_{x=0}^{\tau-t}\gamma_x$ describes the cumulative probability of delays from infection to death up to $\tau - t$ time units in duration.

Using shorthand $C_1^\tau \succcurlyeq D_1^\tau$ for when $\theta(C_1^\tau) \geq \theta(D_1^\tau)$ that is, $\succcurlyeq$ indicates greater than or equal to with respect to total information, we obtain the following equation. We rearrange terms to obtain reporting fractions and delays on different sides by decomposing the geometric mean of a product into products of the geometric means in each term.

$$C_1^\tau \succcurlyeq D_1^\tau : \mathbb{G}\left(\frac{\rho_t}{\sigma_t\text{ifr}_t}\right) \geq \mathbb{G}\left(\frac{H_{\tau-t}}{F_{\tau-t}}\right). \tag{7}$$

Equation (7) states that cases are more informative when the geometric mean of the case to death reporting fractions is at least as large as that of the death and case cumulative delays. Studies preferring death data effectively claim that the variation in ρ_t (which we proved in a previous section always decreases the geometric mean for a given mean constraint) is sufficiently strong to mask the influences of ifr_t, σ_t and any expected variations in those quantities.

Proponents of using death data to infer R_1^τ recognize that the infection-to-death delay (with cumulative distribution $H_{\tau-t}$) is appreciably larger in mean than corresponding reporting lags from infection ($F_{\tau-t}$) and therefore unsuitable for real-time estimation (where this extra lag denatures recent information as we showed in earlier sections). We allow for all of these adjustments. We assume that the ifr is constant (maximizing $\mathbb{G}(\text{ifr}_t)$) and that death ascertainment is perfect ($\sigma_t = 1$). Even for purely retrospective estimation with correction for delays we expect $\mathbb{G}\left(\frac{H_{\tau-t}}{F_{\tau-t}}\right) \leq 1$. We set this to 1, maximizing the informativeness of D_1^τ.

Combining these assumptions we reduce equation (7) to the following equation. This presents a sufficient condition for case data to be more reliable than the death time series.

$$C_1^\tau \succcurlyeq D_1^\tau : \mathbb{G}(\rho_t)_{\bar{\rho}\in[0.07,0.38]} \geq \text{ifr} \sim 0.01. \tag{8}$$

Here we choose a relatively large ifr for COVID-19 of 1% (ref. [44]). Case reporting fraction estimates range from about 7% to 38% (ref. [18]), which we apply to constrain $\bar{\rho}$, the mean ρ_t. Inputting these estimates, we examine possible ρ_t sampling distributions under the Beta(a, b) formulation from earlier sections. Our main results are in Fig. 4. We take 10^4 samples of ρ_t from each of 2,000 distributions parametrized over $10^{-1} \leq b \leq 10^2$ with a set to satisfy our $\bar{\rho}$ reporting constraints.

Figure 4a plots our metric against these constraints and the ifr threshold. Whenever $\theta(C_1^\tau) \geq \text{ifr}$ we find that case data are more reliable. This appears to occur for many possible combinations of ρ_t. The inset charts the proportion of Beta distributions that cross

the threshold. This varies from about 45% at $\bar{\rho} = 0.07$ to 90% at $\bar{\rho} = 0.38$. While these figures will differ depending on how likely a given level of variability is, they offer robust evidence that death counts are not necessarily more reliable. Even when deaths are perfectly ascertained ($\sigma_t = 1$) the small ifr term in D_1^τ means that 99% of the original incidence data are lost, contributing appreciable uncertainty.

These points are reinforced by the design choices we have made, which inflate the relative information in the death time series. In reality, $\sigma_t < 1$, $\text{ifr} < 0.01$, neither is constant[44,45] and the uncertainty we include around $\bar{\rho}_t$ is wider than that inferred in ref. [18]. Our results are therefore resilient to uncertainties in noise source estimates. Figure 4b displays the distributions of our sampling fractions, with red (blue) indicating which shapes provide more (less) information than death data (equation (8)). Our results also hold for both real-time and retrospective analyses, as we ignored the noise induced by the additional delays that death data contain (relative to case reports) when we maximized $\mathbb{G}\left(\frac{H_{\tau-t}}{F_{\tau-t}}\right)$.

Consequently, death data cannot be assumed, without rigorous and context-specific examination, to be generally more epidemiologically meaningful. For example, while D_1^τ is unlikely to be more reliable in well mixed populations, it may be in high-risk settings (for example, care homes) where the local ifr is notably larger. Vaccines and improved healthcare, which substantially reduce ifr values in most contexts, will make death time series less informative about R_1^τ. However, pathogens such as Ebola virus, which induce large ifr parameters, might result in death data that are more reliable than their case counts. We explore these points and demonstrate the practical applicability of our metrics in the next section.

Practical applications of information metrics. Our metrics provide an interpretable, simulation-agnostic and easily computable approach to quantifying the relative reliability of different epidemic time series. Because $\theta(.)$ is independent of usually unknown R_t and Λ_t terms, it is robust to generation time misspecification and only requires estimates of noise terms for its calculation (hence no epidemic curve simulations are needed). Moreover, it depends purely on the geometric means of noise variables, which can be decomposed such that the influence of any noise source is clearly interpreted from the magnitude of its specific mean (see equation (5)).

These properties make $\theta(.)$ of practical use and we illustrate the benefits of our methodology using COVID-19 and EVD examples. In contrast to Fig. 4 where we maximized the information in deaths and minimized that from cases to bolster our rejection of the assertion that death data are definitively more informative, here we focus on inputting empirical noise distributions derived from real data. When distributions are unavailable we describe noise uncertainties via maximum entropy distributions based on what estimates are available (these are geometric, Geo, if a mean is given and uniform, Unif, over 95% credible intervals).

For COVID-19 we once again examine if death data are more reliable. From equation (7) we conclude $C_1^\tau \succcurlyeq D_1^\tau$ if $\frac{\mathbb{G}(\rho_t)}{\mathbb{G}(\sigma_t)\mathbb{G}(\text{ifr}_t)} \geq \frac{\mathbb{G}(H_{\tau-t})}{F_0}$. This follows as $F_0 = \delta_0 = \min \mathbb{G}(F_{\tau-t})$ and ensures (if we are using NEVR delays) that we do not take ratios of upper bounds, as $\mathbb{G}(H_{\tau-t})$ already bounds the information in the infection-to-death delay. If delays are OBNR then equation (7) will be exact. We model $\rho_t \sim \text{Unif}(0.06, 0.08)$ (ref. [18]), $\delta_x \sim \text{Geo}(\frac{1}{1+10.8})$ (ref. [46]), $\sigma_t \sim \text{Unif}(\frac{1}{1.34}, \frac{1}{1.29})$ (ref. [45]), $\text{ifr}_t \sim \text{Unif}(\frac{0.53}{100}, \frac{0.82}{100})$ (ref. [44]) and $\gamma_x \sim \text{NB}(\frac{1}{1+1.1}, \frac{21}{21+\frac{1}{1+1.1}})$ (ref. [47]) and sample from these distributions 10^4 times. We compute the terms in the inequality above and represent the relative information as $\log\theta(C_1^\tau) - \log\theta(D_1^\tau)$ for easy visualization.

This leads to the top panel of Fig. 5. Despite our use of the smallest reporting proportions from ref. [18] we find that death data are less reliable. For EVD, we test the alternative hypothesis that case data are less reliable in the bottom panel of Fig. 5. We decide $D_1^\tau \succcurlyeq C_1^\tau$ if $\frac{\mathbb{G}(\rho_t)}{\mathbb{G}(\sigma_t)\mathbb{G}(\text{ifr}_t)} \geq H_0$, as we know $H_0 = \gamma_0 = \min \mathbb{G}(H_{\tau-t})$ and $\max \mathbb{G}(F_{\tau-t}) = 1$. We let $\sigma_t = 1$ (no estimates were easily available) and model $\rho_t \sim \text{Unif}(0.33, 0.83)$ (ref. [16]), $\text{ifr}_t \sim \text{Unif}(0.69, 0.73)$ (ref. [48]) and $\gamma_x \sim \text{NB}(1.5, \frac{21.4}{21.4+1.5})$ (roughly from ref. [48]). The negative values of $\log\theta(C_1^\tau) - \log\theta(D_1^\tau)$ in Fig. 5 suggest EVD death data as the more informative source. However, this can easily change if $\sigma_t \ll 1$, as the difference is not as strong as for COVID-19.

While we tried to keep estimates as realistic as possible, the point of Fig. 5 is to demonstrate how our metrics may be practically applied given noise estimates. Sampling from appropriate distributions means that we can propagate the uncertainty on these estimates into our metrics. We provide open source code for modifying this template analysis to include any user-defined distributions at https://github.com/kpzoo/information-in-epidemic-curves. As high-resolution outbreak data collection initiatives such as Global.health[49] and REACT[7] progress, enhancing surveillance and our quantification of noise sources, we expect our framework to grow in practical utility.

Discussion

Public health policy-making is becoming progressively data driven. Key infectious disease parameters[6] such as instantaneous reproduction numbers and growth rates, fitted to heterogeneous outbreak data sources (for example, case, death and hospitalization incidence curves), are increasingly contributing to the evidence base for understanding pathogen spread, projecting epidemic burden and designing effective interventions[4,6,50]. However, the validity and value of such parameters depends substantially on the quality of the available surveillance data[1,7]. Although many studies have made important advances in underscoring and correcting errors in these data[12,31], a framework to directly and generally quantify epidemic data quality is needed.

We made progress towards such a framework by finding the total information, $\mathbb{T}(.)$, available from epidemic curves corrupted by reporting delays and under-reporting. These are predominant noise sources that limit surveillance quality and apply to common outbreak data for inferring pathogen transmissibility such as cases, death counts, hospitalizations, infection prevalence and wastewater virus surveys. By maximizing $\mathbb{T}(.)$, we minimize the overall uncertainty of our transmissibility estimates, hence measuring the reliability of this data stream. This approach yielded a non-dimensional metric $\theta(.)$ that allows analytic and generalizable insights into how noisy surveillance data degrade estimate precision.

Our framework provided insight into the nuances of noise, elucidating how the mean and variability of delay and under-reporting schemes both matter. For example, fluctuating reporting protocols with larger mean may outperform more stable ones at lower mean. Moreover, under mean surveillance constraints, our metrics revealed that constant under-reporting of cases minimizes loss of information, while constant delays in reporting maximize this loss. The first result bolsters conventional thinking[9], while the second highlights the need for timely data[39].

Because the reporting of cases can vary substantially when tracking acute diseases such as COVID-19, various studies have assumed death data to be more reliable[5]. Using our metrics, we were able to qualify this claim. We found that ifr acts as a reporting fraction with very small mean. Only the most severely varying case reporting protocols can cause larger information loss, suggesting that in many instances this assertion may not hold. Note that this analysis does not even consider the additional advantages that case data bring in terms of timeliness, and shows the ability of our framework to rank the quality of different data streams.

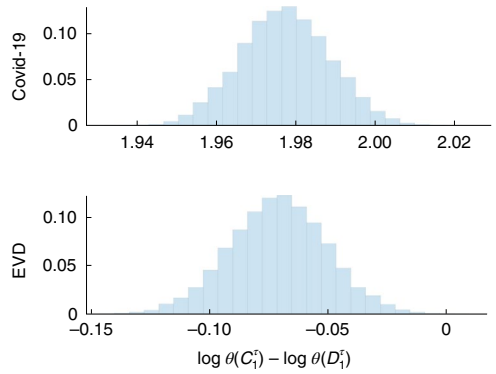

Fig. 5 | The relative information in case and death data for EVD and COVID-19 case studies. We compute information metrics $\theta(.)$ for case (C_1^τ) and death (D_1^τ) time series using empirically derived under-reporting and delay noise distributions for COVID-19 (top panel) and EVD (bottom panel). See the main text for the specific distributions used, which account for the uncertainty in noise estimates (in the absence of knowledge of this uncertainty, maximum entropy distributions are applied). We take 10^4 samples from each distribution and calculate the logarithmic difference $\log\theta(C_1^\tau) - \log\theta(D_1^\tau)$, with positive or negative values indicating when case or death data have the higher information content, respectively.

However, there may be other crucial reasons for preferring to estimate pathogen spread from death data. For example, if extremely little is known about the level of reporting (very limited surveillance capacity might cause insurmountable case reporting fraction uncertainties) or if a death-based reproduction number is itself of interest as a severity indicator[20]. Our framework can also help inform these discussions by improving the precision of our reasoning about noise. This is exemplified by our EVD analysis, where we could show that the large ifr of the disease translated into death counts being the better data, provided their under-reporting is not large.

As hospitalization curves generally interpolate among the types of noise in case and death data, this might be the best a priori choice of data for inferring transmissibility. Some studies also propose to circumvent these ranking issues by concurrently analyzing multiple data streams[32,50]. This then opens questions about how each data stream should be weighed in the ensuing estimates. Our framework may also help by quantifying the most informative parts of each contributing stream. A common way of deriving consensus weights individual estimates by their inverse variance[51]. As the FI defines the best possible inverse variance of estimates, our metrics naturally apply.

While our framework can enhance understanding and quantification of surveillance noise, it has several limitations. First, it depends on renewal model descriptions of epidemics[27]. These models assume homogeneous mixing and that the generation time distribution of the disease is known. While the inclusion of more realistic network-based mixing may not improve transmissibility estimates[52] (and this extra complexity may occlude insights), the generation time assumption may only be ameliorated through the provision of updated, high-quality line-list data[34,49]. However, our relative metrics in equations (4) and (5) and equations (7) and (8) are mostly robust to generation time distribution misspecifications (and even changes) as they do not depend on the Λ_t terms (these cancel out).

Further, our analysis is contingent on having estimates of the delays, underascertainment rates and other noise sources within data streams. These may be unavailable or themselves highly unreliable. If at least some information on their uncertainties is available we can propagate these into our metrics by replicating the Monte Carlo approach underlying our case studies. If no estimates are available then we cannot perform any analyses as R_t will not be identifiable. However, our framework can still be of use as a rigorous testbed for examining hypotheses on potential noise sources without extensive simulation.

Recent initiatives have aimed at improving the resolution and completeness of outbreak data[7,49]. Concurrently, estimating noise sources from both existing and novel data streams is a growing research area[18,53]. As a result, we expect that our metrics will only increase in practical utility and that concerns around the availability of noise estimates will diminish. We also assume that the time scale t chosen ensures that R_t parameters are independent. This may be invalid but in such instances we can append non-diagonal terms to FI matrices or use our metric as an upper bound.

Finally, we defined the reliability or informativeness of a data stream in terms of minimizing the joint uncertainty of the entire sequence of reproduction numbers R_1^τ. This is known as a D-optimal design[36]. However, we may instead want to minimize the worst uncertainty among the R_1^τ (which may better compensate known asymmetries in inferring transmissibility[54]). Our framework can be reconfigured to tackle such problems by appealing to other design laws. We can solve this specific problem by deriving an E-optimal design, which maximizes the smallest eigenvalue of our FI matrix.

Methods

Renewal models and Fisher information theory. The renewal model[27,35] is a popular approach for describing how infections dynamically propagate during the course of an epidemic. The number of new infections at t, I_t, depends on R_t, which counts the new infections generated per infected individual (on average), and Λ_t, which measures how many past infections (up to $t-1$) will effectively produce new ones. This measurement weighs past infections by $\mathbf{w}$. We define w_s as the probability that it takes s time units for a primary infection to generate a secondary one. The distribution is then $\mathbf{w} = w_1^\infty := \{w_1, w_2, \ldots, w_\infty\}$.

The statistical relationship between these quantities is commonly modeled as in the following equation, with Pois specifying a Poisson distribution[21]. This relationship only strictly holds if I_t is perfectly recorded both in size (no under-reporting) and in time (no delays in reporting).

$$I_t \sim \text{Pois}\left(\Lambda_t R_t\right), \quad \Lambda_t := \sum_{x=1}^{t-1} w_{t-x} I_x. \tag{9}$$

However, as infections are rarely observed, I_t is often approximated by proxies such as reported cases and $\mathbf{w}$ replaced with the serial interval distribution, describing the times between the onset of symptoms of those cases. Equation (9) has been widely used to model transmission dynamics of many infectious diseases, including COVID-19[5], influenza[55] and EVD[48].

A common and important problem in infectious disease epidemiology is the estimation of the latent variable R_t from the incidence curve of infections or some more easily observed proxy. If this time series persists during $1 \le t \le \tau$, then we aim to infer the vector of parameters $R_1^\tau := \{R_t : 1 \le t \le \tau\}$ from time series $I_1^\tau := \{I_t : 1 \le t \le \tau\}$ or its proxy (see Results for this more practical inference problem). We assume that time is scaled in units such that R_t can be expected to change (independently) at every t. This may be weekly for COVID-19 or malaria[21,56] but monthly for rabies[57]. Note that $\mathbf{w}$ and I_t must be aggregated, as needed, to match these units. Related branching[58] and moving-average models[59] feature similar aggregation.

Following the development in refs. [43,55], we solve this inference problem by constructing the incidence log-likelihood function $\ell(R_1^\tau) = \log\mathbb{P}(I_1^\tau \mid R_1^\tau)$ as in the following equation with K_τ as some constant that does not depend on any R_t. This involves combining Poisson likelihoods from equation (9) across time units $1 \le t \le \tau$ as in ref. [21].

$$\ell(R_1^\tau) = \sum_{t=1}^{\tau} I_t \log R_t - \Lambda_t R_t + K_\tau. \tag{10}$$

We compute the MLE of R_t as $\hat{R}_t$, which is the maximal solution of $\frac{\partial\ell(R_1^\tau)}{\partial R_t} = 0$. From equation (10) this gives $\hat{R}_t = I_t\Lambda_t^{-1}$ (ref. [27]). Repeating this for all t we obtain estimates of the complete vector of transmissibility parameters R_1^τ underlying I_1^τ.

To quantify the precision (the inverse of the variance, var) around these MLEs or any unbiased estimator of R_t we calculate the FI that I_1^τ contains about R_t. This is $\mathbb{F}_I(R_t) := \mathbb{E}\left[-\frac{\partial^2\ell(R_1^\tau)}{\partial R_t^2}\right]$, where expectation $\mathbb{E}[.]$ is taken across the data I_1^τ (hence the subscript I). The FI defines the best (smallest) possible uncertainty asymptotically achievable by any unbiased estimate, $\tilde{R}_t$. This follows from the Cramér–Rao bound[30], which states that $\text{var}(\tilde{R}_t) \ge \mathbb{F}_I(R_t)^{-1}$. The confidence

intervals around $\tilde{R}_t$ converge to $\tilde{R}_t \pm 1.96\,\mathbb{F}_I(R_t)^{-\frac{1}{2}}$. The FI also links to the Shannon mutual information that I_1^τ contains about R_t (these measures are bijective under Gaussian approximations)[60,61] and is pivotal to describing both model identifiability and complexity[30,33].

Using the Poisson renewal log-likelihood in equation (10) we obtain the FI as the left-hand equality in the following equation. Observe that this depends on the unknown 'true' R_t.

$$\mathbb{F}_I(R_t) = \Lambda_t R_t^{-1}, \quad \mathbb{F}_I(2\sqrt{R_t}) = \Lambda_t. \tag{11}$$

This reflects the heteroscedasticity of Poisson models, where the estimate mean and variance are co-dependent. We construct a square root transform that uncouples this dependence[43], yielding the right-hand formula in equation (11). We can evaluate $\mathbb{F}_I(2\sqrt{R_t})$ purely from I_1^τ. The result follows from the FI change of variables formula $\mathbb{F}_I(\mathcal{R}_t) = \mathbb{F}_I(R_t)\left(\frac{\partial R_t}{\partial \mathcal{R}_t}\right)^2$ (ref. [30]). This transformation has several optimal statistical properties[29,37] and so we will commonly work with $\mathcal{R}_t := 2\sqrt{R_t}$.

As we are interested in evaluating the informativeness or reliability of the entire I_1^τ time series for inferring transmission dynamics we require the total FI it provides for all estimable reproduction numbers, R_1^τ. As we noted above, the inverse of the square root of the FI for a single R_t corresponds to an uncertainty (or confidence) interval. Generalizing this to multiple dimensions yields an uncertainty ellipsoid with volume inversely proportional to the square root of the determinant of the FI matrix[33,37]. This matrix has diagonals given by $\mathbb{F}_I(R_t)$ and off-diagonals defined as $\mathbb{E}[-\frac{\partial^2 \ell(R_1^\tau)}{\partial R_t R_s}]$ for $1 \le t,s \le \tau$.

Maximizing this non-negative determinant, which we denote the total information $\mathbb{T}(I_1^\tau)$ from the data I_1^τ, corresponds to what is known as a D-optimal design[36]. This design minimizes the overall asymptotic uncertainty around estimates of the vector R_1^τ. As the renewal model in equation (9) treats every R_t as independent, off-diagonal terms are 0 and $\mathbb{T}(I_1^\tau)$ is a product of the diagonal FI terms. Transforming $R_t \to \mathcal{R}_t$ we then obtain

$$\mathbb{T}(I_1^\tau) = \prod_{t=1}^{\tau} \sqrt{\mathbb{F}_I(\mathcal{R}_t)} = \prod_{t=1}^{\tau} \sqrt{\Lambda_t}. \tag{12}$$

If we work directly in R_t we obtain $\prod_{t=1}^{\tau} \Lambda_t^{\frac{1}{2}} R_t^{-\frac{1}{2}}$ instead. In two dimensions (that is, $\tau = 2$) our ellipsoid becomes an ellipse and equation (12) intuitively means that its area is proportional to a product of lengths $\mathbb{F}_I(\mathcal{R}_1)^{-\frac{1}{2}}\mathbb{F}_I(\mathcal{R}_2)^{-\frac{1}{2}}$, which factors in the uncertainty from each estimate.

We will use this recipe of formulating a log-likelihood for R_1^τ given some data source and then computing the total information, $\mathbb{T}(.)$, it provides about these parameters to quantify the reliability of case, death and other I_1^τ proxies for inferring transmissibility. Comparing data source quality will involve ratios of these total information terms. Metrics such as equation (12) are valuable because they measure the usable information within a time series and also delimit the possible distributions that a model can describe given these data (see refs. [33,62] for more on these ideas, which emerge from information geometry). Transforms such as $\mathcal{R}_t = 2\sqrt{R_t}$ stabilize these metrics (that is, maximize robustness) to unknown true values[29,37].

Epidemic noise sources and surveillance models. We investigate two important and common sources of noise, under-reporting and reporting delay, which limit our ability to precisely monitor I_1^τ, the true time series of new infections. We quantify how much information is lost due to these noise processes by examining how these imperfections degrade $\mathbb{T}(I_1^\tau)$, the total information obtainable from I_1^τ under perfect (noiseless) surveillance for estimating parameter vector R_1^τ (equation (12)). Figure 1 illustrates how these two main noise sources individually alter the shape and size of incidence curves.

(1) Under-reporting or underascertainment. Practical surveillance systems generally detect some fraction of the true number of infections occurring at any given t. If this proportion is $\rho_t \le 1$ then the number of cases, C_t, observed is generally modeled as $C_t \sim \mathrm{Bin}(I_t, \rho_t)$ (refs. [23,56]), where Bin indicates the binomial distribution. The under-reported fraction is $1-\rho_t$ and so $C_t \sim \mathrm{Pois}(\rho_t \Lambda_t R_t)$. Reporting protocols are defined by choices of ρ_t. CONR is the simplest and most popular, assuming every $\rho_t = \rho$ (ref. [21]). VARR describes general time-varying protocols where every ρ_t can differ[28].

(2) Reporting delays or latencies. There can be notable lags between an infection and when it is reported[13]. If $\boldsymbol{\delta}$ defines the distribution of these lags with δ_x as the probability of a delay of $x \ge 0$ time units, then the new cases reported at t, C_t, sums infections actually occurring at t but not delayed and those from previous days that were delayed[10]. This is commonly modeled as $C_t \sim \mathrm{Pois}\left(\sum_{x=0}^{t-1} \delta_x \Lambda_{t-x} R_{t-x}\right)$ (refs. [20,28]) and means that true incidence I_t splits over future times as $\sim \mathrm{Mult}(I_t, \boldsymbol{\delta})$, where Mult denotes multinomial[12]. The C_t time series is OBNR if we later learn about the past I_t splits (right censoring), else we say data are never reported (NEVR).

We make some standard assumptions[8,11,21,28] in incorporating the above noise sources within renewal model frameworks. We only consider stationary delay distributions, that is, $\boldsymbol{\delta}$ and any related distributions do not vary with time, and we neglect co-dependences between reporting and transmissibility. Additionally, we assume that these distributions and all reporting or ascertainment fractions, that is, ρ_t and related parameters, are inferred from other data (for example, contact tracing studies or line lists)[12]. In the absence of these assumptions R_1^τ would be non-identifiable and the inference problem ill-defined. In Results we examine how noise sources (1) and (2) in combination limit the information available about epidemic transmissibility.

FI derivations for practical data. We derive the FI of parameters R_1^τ given the case curve C_1^τ under the model of equation (1). This procedure mirrors that used above to obtain equation (12). We initially assume that reporting delays are OBNR, that is, we eventually learn the source time of cases at a later date. This corresponds to a right censoring that can be compensated for using nowcasting techniques[13]. Later we prove that this not only defines a practical noise model but also serves as an upper bound on the information available from NEVR delays, where the true timestamps of cases are never resolved. Mathematically, the OBNR assumption lets us decompose the sum in equation (1). We can therefore identify the component of C_t that is informative about R_x. This follows from the statistical relationship $C_t \,|\, R_x \sim \mathrm{Pois}(\delta_{t-x}\rho_x \Lambda_x R_x)$.

As we are interested in the total information that C_1^τ contains about every R_t we collect and sum contributions from every C_t. We can better understand this process by constructing the matrix Q in the following equation, which expands the convolution of the reporting fractions with the delay probabilities over the entire observed time series.

$$Q = \begin{bmatrix} \delta_0\rho_\tau & \delta_1\rho_{\tau-1} & \delta_2\rho_{\tau-2} & \cdots & \delta_{\tau-1}\rho_1 \\ 0 & \delta_0\rho_{\tau-1} & \delta_1\rho_{\tau-2} & \cdots & \delta_{\tau-2}\rho_1 \\ \vdots & 0 & \delta_0\rho_{\tau-2} & \ddots & \vdots \\ \vdots & \vdots & 0 & \ddots & \delta_1\rho_1 \\ 0 & 0 & \cdots & \cdots & \delta_0\rho_1 \end{bmatrix}. \tag{13}$$

We work with the vector $\boldsymbol{\mu} = [\mu_\tau, \mu_{\tau-1}, \ldots, \mu_1]^{\mathrm{T}}$ with $\mu_t = \Lambda_t R_t$ and T denoting the transpose operation. Then $Q\boldsymbol{\mu} = [\mathbb{E}[C_\tau], \mathbb{E}[C_{\tau-1}], \ldots, \mathbb{E}[C_1]]^{\mathrm{T}}$ with $\mathbb{E}[C_t]$ as the mean of the reported case incidence at t.

The components of C_1^τ that contain information about every R_t parameter follow from $Q\mathrm{T}\boldsymbol{\mu} = [\delta_0\rho_\tau\mu_\tau, (\delta_0+\delta_1)\rho_{\tau-1}\mu_{\tau-1}, \ldots, (\delta_0+\ldots+\delta_{\tau-1})\rho_1\mu_1]$. The elements of this vector are Poisson means formed by collecting and summing the components of C_1^τ that inform about $[R_\tau, R_{\tau-1}, \ldots, R_1]$, respectively. Hence we obtain the key relationship in the following equation with $F_{\tau-t} := \sum_{x=0}^{\tau-t} \delta_x$ as the cumulative probability delay distribution.

$$C_1^\tau \,|\, R_t \sim \mathrm{Pois}\left(\rho_t F_{\tau-t} \Lambda_t R_t\right). \tag{14}$$

The ability to decompose the row or column sums from Q into the Poisson relationships of equation (14) is a consequence of the independence properties of renewal models and the infinite divisibility of Poisson formulations.

Using equation (14) and analogs to Poisson log-likelihood definitions from equation (10) we derive the FI that C_1^τ contains about R_t as follows:

$$\mathbb{F}_C(R_t) = \rho_t F_{\tau-t} \Lambda_t R_t^{-1}. \tag{15}$$

As in equation (11) we recompute the FI in equation (15) under the transform $\mathcal{R}_t = 2\sqrt{R_t}$ to obtain $\mathbb{F}_C(\mathcal{R}_t) = \rho_t F_{\tau-t} \Lambda_t$. It is clear that under-reporting and delays can substantially reduce our information about instantaneous reproduction numbers. As we might expect, if $\rho_t = 0$ (no reports at time unit t) or $F_{\tau-t} = 0$ (all delays are larger than $\tau - t$) then we have no information on R_t at all from C_1^τ. If reporting is perfect then $\rho_t = 1$, $F_{\tau-t} = 1$ and $\mathbb{F}_C(R_t)$ is equal to the FI from I_1^τ in equation (11).

The MLE, $\hat{R}_t$, also follows from equation (14) (see subsections above) as $(\sum_{x=t}^{\tau} C_x \,|\, R_t)(\rho_t F_{\tau-t} \Lambda_t)^{-1}$, with $C_x \,|\, R_t$ as the component of C_x containing information about R_t. By comparison with the MLE under perfect surveillance we see that $(\sum_{x=t}^{\tau} C_x \,|\, R_t)(\rho_t F_{\tau-t})^{-1}$ is equivalent to applying a nowcasting correction as in refs. [12,13]. An important point to make here is that, while such corrections can remove bias, allowing inference despite these noise sources, they cannot improve on the information (in this case equation (15)) inherently available from the data. This is known as the data processing inequality[63,64].

If we cannot resolve the components of every C_t from equation (1) as $\sum_{x=1}^{t} \mathrm{Pois}(\delta_{t-x}\rho_x \Lambda_x R_x)$, then the reporting delay is classed as NEVR (that is, we never uncover case source dates). Hence we know $Q\boldsymbol{\mu}$ but not $Q\mathrm{T}\boldsymbol{\mu}$. Accordingly, we must use equation (1) to construct an aggregated log-likelihood $\ell(R_1^\tau) = \log \mathbb{P}(C_1^\tau \,|\, R_1^\tau) = \sum_{t=1}^{\tau} \log \mathbb{P}(C_1^\tau \,|\, R_t)$. This gives the following equation with the aggregate term $h(R_1^t) := \sum_{x=1}^{t} \delta_{t-x}\rho_x \Lambda_x R_x$. We ignore constants that do not depend on any R_t in this likelihood.

$$\ell(R_1^\tau) = \sum_{t=1}^{\tau} C_t \log h(R_1^t) - h(R_1^t). \quad (16)$$

For every given R_t we decompose $h(R_1^s)$ for $s \geq t$ into the form $\delta_{s-t}\rho_t\Lambda_t R_t + a_t$, where a_t collects all terms that are not informative about this specific R_t. Here $s \geq t$ simply indicates that information about R_t is distributed across later times due to the reporting delays.

We can then obtain the FI contained in C_1^τ about R_t by computing $\mathbb{E}[-\frac{\partial^2 \ell(R_1^\tau)}{\partial R_t^2}]$, yielding the following equation (see Supplementary Notes for derivation details), with $b_t := a_t(\delta_{s-t}\rho_t\Lambda_t R_t)^{-1}$.

$$\mathbb{F}_C(R_t) = \sum_{x=t}^{\tau} \delta_{x-t}\rho_t\Lambda_t(R_t + b_x)^{-1}. \quad (17)$$

If we could decouple the interactions among the reproduction numbers then the b_x terms would disappear and we would recover the expressions derived under OBNR delay types. Since b_x is a function of other reproduction numbers, the overall FI matrix for R_1^τ is not diagonal (there are non-zero terms from evaluating $\mathbb{E}[-\frac{\partial^2 \ell(R_1^\tau)}{\partial R_t R_x}]$).

However, we find that this matrix can be reduced to a triangular form with determinant equal to the product of terms (across t) in equation (17). We show this for the example scenario of $\tau = 3$ in Supplementary Notes. As a result, the FI term for R_t in equation (17) does behave like and correspond to that in equation (15). Interestingly, as $b_x \geq 0$, equation (17) yields the revealing inequality $\mathbb{F}_C(R_t) \leq \rho_t F_{\tau-t}\Lambda_t R_t^{-1}$. This proves that OBNR delays upper bound the information available from NEVR delays. Last, we note that robust transforms cannot be applied to remove the dependence of equation (17) on the unknown R_t parameters.

The best we can do is evaluate equation (17) at the MLEs $\hat{R}_t$. These MLEs emerge as the joint maxima of the set of coupled differential equations $\frac{\partial \ell(R_1^\tau)}{\partial R_t} = \sum_{x=t}^{\tau} \frac{I_x}{b_x + R_t} - \delta_{x-t}\rho_t\Lambda_t$, that is, numerical solutions of equation (18) for all t.

$$\sum_{x=t}^{\tau} C_x(\hat{R}_t + b_x)^{-1} = \rho_t F_{\tau-t}\Lambda_t. \quad (18)$$

Here sums start at t as they include only time points that contain information about R_t. Expectation-maximization algorithms, such as the deconvolution approaches outlined in ref. [10], are viable means of computing these MLEs or equivalents. Note that the nowcasting methods used to correct for OBNR delays do not help here[12] and that for both OBNR and NEVR delays the cumulative probability terms must be aggregated to match chosen time units (for example, if empirical delay distributions are given in days but t is in weeks then F_x sums over $7x$ days).

Data availability

Source data for Figs. 1–5 are available with this manuscript.

Code availability

All data and source code (MATLAB v2021a) for reproducing the analyses and figures in this manuscript, as well as for applying the methodology we have developed here, are freely available at https://github.com/kpzoo/information-in-epidemic-curves with a citable release at ref. [65]. We include a template function (in MATLAB and R) that can be easily modified to compute our metrics with user-defined noise estimates.

Received: 1 March 2022; Accepted: 8 August 2022;
Published online: 26 September 2022

References

1. Anderson, R. et al. *Reproduction Number (R) and Growth Rate (r) of the COVID-19 Epidemic in the UK: Methods of Estimation, Data Sources, Causes of Heterogeneity, and Use as a Guide in Policy Formulation* Technical Report (Royal Society, 2020).
2. Flaxman, S., Mishra, S. & Gandy, A. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020).
3. Li, Y., Campbell, H. & Kulkarni, D. The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (*R*) of SARS-CoV-2: a modelling study across 131 countries. *Lancet Infect. Dis.* **21**, 193–202 (2020).
4. Cauchemez, S., Hoze, N. & Cousien, A. How modelling can enhance the analysis of imperfect epidemic data. *Trends. Parasitol.* **35**, 369–379 (2019).
5. Funk, S., Camacho, A. & Kucharski, A. Assessing the performance of real-time epidemic forecasts: a case study of Ebola in the western area region of Sierra Leone, 2014–15. *PLoS Comput. Biol.* **15**, e1006785 (2019).
6. GOV.UK The *R* value and growth rate. https://www.gov.uk/guidance/the-r-value-and-growth-rate (2021).
7. Riley, S., Ainslie, K. & Eales, O. Resurgence of SARS-CoV-2: detection by community viral surveillance. *Science* **372**, 990–995 (2021).
8. Gostic, K., McGough, L. & Baskerville, E. Practical considerations for measuring the effective reproductive number, R_t. *PLoS Comput. Biol.* **16**, e1008409 (2020).
9. White L. & Pagano, M. Reporting errors in infectious disease outbreaks, with an application to pandemic influenza A/H1N1. *Epidemiol. Perspect. Innov.* **7** (2010).
10. Goldstein, E., Dushoff, J. & Ma, J. Reconstructing influenza incidence by deconvolution of daily mortality time series. *Proc. Natl Acad. Sci. USA* **106**, 21825–21829 (2009).
11. Wallinga, J. & Teunis, P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.* **160**, 509–516 (2004).
12. Yang, P. & Chowell, G. *Quantitative Methods for Investigating Infectious Disease Outbreaks* (Texts in Applied Mathematics Vol. 70, Springer, 2019).
13. Lawless, J. Adjustments for reporting delays and the prediction of occurred but not reported events. *Can. J. Stat.* **22**, 15–31 (1994).
14. Salmon, M., Schumacher, D. & Stark, K. Bayesian outbreak detection in the presence of reporting delays. *Biom. J.* **57**, 1051–1067 (2015).
15. Gunther, F., Bender, A. & Katz, K. Nowcasting the COVID-19 pandemic in Bavaria. *Biom. J.* **63**, 490–502 (2021).
16. Dalziel, B., Lau, M. & Tiffany, M. Unreported cases in the 2014–2016 Ebola epidemic: spatiotemporal variation, and implications for estimating transmission. *PLOS Negl. Trop. Dis.* **12**, e0006161 (2018).
17. Funk, S., Bansal, S. & Bauch, C. Nine challenges in incorporating the dynamics of behaviour in infectious diseases models. *Epidemics* **10**, 21–25 (2015).
18. Pullano, G., Di Domenico, L. & Sabbatini, C. Underdetection of cases of COVID-19 in France threatens epidemic control. *Nature* **590**, 134–139 (2021).
19. Pitzer, V., Chitwood, M. & Havumaki, J. The impact of changes in diagnostic testing practices on estimates of COVID-19 transmission in the United States. *Am. J. Epidemiol.* **190**, 1908–1917 (2021).
20. Nouvellet, P., Bhatia, S. & Cori, A. Reduction in mobility and COVID-19 transmission. *Nat. Commun.* **12**, 1090 (2021).
21. Cori, A., Ferguson, N. & Fraser, C. A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.* **178**, 1505–1512 (2013).
22. Parag, K., Donnelly, C. & Jha, R. An exact method for quantifying the reliability of end-of-epidemic declarations in real time. *PLoS Comput. Biol.* **16**, e1008478 (2020).
23. Fraser, C., Donnelly, C. & Cauchemez, S. Pandemic potential of a strain of influenza A (H1N1): early findings. *Science* **324**, 1557–1561 (2009).
24. Hohle, M. & der Heiden, M. Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. *Biometrics* **70**, 993–1002 (2014).
25. Ali, S., Kadi, A. & Ferguson, N. Transmission dynamics of the 2009 influenza (H1N1) pandemic in India: the impact of holiday-related school closure. *Epidemics* **5**, 157–163 (2013).
26. Li, R., Pei, S. & Chen, B. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489–493 (2020).
27. Fraser, C. Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS ONE* **8**, e758 (2007).
28. Azmon, A., Faes, C. & Hens, N. On the estimation of the reproduction number based on misreported epidemic data. *Stat. Med.* **33**, 1176–1192 (2014).
29. Bartlett, M. The use of transformations. *Biometrics* **3**, 39–52 (1947).
30. Lehmann, E. & Casella, G. *Theory of Point Estimation* 2nd edn (Springer, 1998).
31. Polonsky, J., Baidjoe, A. & Kamvar, Z. Outbreak analytics: a developing data science for informing the response to emerging pathogens. *Phil. Trans. R. Soc. B* **374**, 20180276 (2019).
32. Brauner, J., Mindermann, S. & Sharma, M. Inferring the effectiveness of government interventions against COVID-19. *Science* **371**, eabd9338 (2021).
33. Grunwald, P. *The Minimum Description Length Principle* (MIT Press, 2007).
34. Ali, S., Wang, L. & Lau, E. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. *Science* **369**, 1106–1109 (2020).
35. Wallinga, J. & Lipsitch, M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B* **274**, 599–604 (2007).
36. Atkinson, A. & Donev, A. *Optimal Experimental Designs* (Oxford Univ. Press, 1992).
37. Parag, K. & Pybus, O. Robust design for coalescent model inference. *Syst. Biol.* **68**, 730–743 (2019).
38. Marshall, A., Olkin, I. & Arnold, B. *Inequalities: Theory of Majorization and its Applications* 2nd edn (Springer, 2011).

39. Casella, F. Can the COVID-19 epidemic be controlled on the basis of daily test reports? *IEEE Control Syst. Lett.* **5**, 1079–1084 (2021).
40. Vanni, F., Lambert, D. & Palatella, L. On the use of aggregated human mobility data to estimate the reproduction number. *Sci. Rep.* **11**, 23286 (2021).
41. Huisman, J., Scire, J. & Caduff, L. Wastewater-based estimation of the effective reproductive number of SARS-CoV-2. *Environ. Health Perspect.* **130**, 057011 (2022).
42. Rissanen, J. Fisher information and stochastic complexity. *IEEE Trans. Inf. Theory* **42**, 40–47 (1996).
43. Parag, K. & Donnelly, C. Adaptive estimation for epidemic renewal and phylogenetic skyline models. *Syst. Biol.* **69**, 1163–1179 (2020).
44. Meyerowitz-Katz, G. & Merone, L. A systematic review and meta-analysis of published research data on COVID-19 infection fatality rates. *Int. J. Infect. Dis.* **101**, 138–148 (2020).
45. Centers for Disease Control and Prevention *Estimated Covid-19 Burden* (2022).
46. Huisman, J. et al. Estimation and worldwide monitoring of the effective reproductive number of SARS-CoV-2. *eLife* **11**, e71345 (2022). https://doi.org/10.7554/eLife.71345
47. Irons, N. & Raftery, A. Estimating SARS-CoV-2 infections from deaths, confirmed cases, tests, and random surveys. *Proc. Natl Acad. Sci. USA* **118**, e2103272118 (2021).
48. WHO Ebola Response Team Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections. *N. Engl. J. Med.* **371**, 1481–1495 (2014).
49. Global.health—a Data Science Initiative (2022). https://global.health/
50. De Angelis, D., Presanis, A. & Birrell, P. Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics* **10**, 83–87 (2015).
51. Hartung, J., Knapp, G. & Sinha, B. *Statistical Meta-Analysis with Applications* (Wiley Series in Probability and Statistics, Wiley, 2008).
52. Liu, Q., Ajelli, M. & Aleta, A. Measurability of the epidemic reproduction number in data-driven contact networks. *Proc. Natl Acad. Sci. USA* **115**, 12680–12685 (2018).
53. COVID-19 Forecasting Team Variation in the COVID-19 infection–fatality ratio by age, time, and geography during the pre-vaccine era: a systematic analysis. *Lancet* **399**, 1469–1488 (2022).
54. Parag, K. & Donnelly, C. Fundamental limits on inferring epidemic resurgence in real time using effective reproduction numbers. *PLoS Comput. Biol.* **18**, e1010004 (2022).
55. Fraser, C., Cummings, D. & Klinkenberg, D. Influenza transmission in households during the 1918 pandemic. *Am. J. Epidemiol.* **174**, 505–514 (2011).
56. Churcher, T., Cohen, J. & Ntshalintshali, N. Measuring the path toward malaria elimination. *Science* **344**, 1230–1232 (2014).
57. Bourhy, H., Nakoune, E. & Hall, M. Revealing the micro-scale signature of endemic zoonotic disease transmission in an African urban setting. *PLoS Pathog.* **12**, e1005525 (2016).
58. Parag, K. Sub-spreading events limit the reliable elimination of heterogeneous epidemics. *J. R. Soc. Interface* **18**, 20210444 (2021).
59. Bracher, J. & Held, L. A marginal moment matching approach for fitting endemic–epidemic models to underreported disease surveillance counts. *Biometrics* **77**, 1202–1214 (2020).
60. Brunel, N. & Nadal, J. Mutual information, Fisher information, and population coding. *Neural Comput.* **10**, 1731–1757 (1998).
61. Parag, K., Pybus, O. & Wu, C. Are skyline plot-based demographic estimates overly dependent on smoothing prior assumptions? *Syst. Biol.* **71**, 121–138 (2022).
62. Myung, I., Balasubramanian, V. & Pitt, M. Counting probability distributions: differential geometry and model selection. *Proc. Natl Acad. Sci. USA* **97**, 11170–11175 (2000).
63. Zamir, R. A proof of the Fisher information inequality via a data processing argument. *IEEE Trans. Inf. Theory* **44**, 1246–1250 (1998).
64. Cover, T. & Thomas, J. *Elements of Information Theory* 2nd edn (Wiley, 2006).
65. Parag, K. kpzoo/information-in-epidemic-curves: information-in-epidemic-curves (v1.0.0). *Zenodo* https://doi.org/10.5281/zenodo.6962446 (2022).

Acknowledgements

Thanks to M. Hickman for providing useful and interesting comments on the manuscript. K.V.P. acknowledges support from the NIHR Health Protection Research Unit in Behavioural Science and Evaluation at the University of Bristol. C.A.D. thanks the UK National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Emerging and Zoonotic Infections for funding (grant HPRU200907). K.V.P. and C.A.D. acknowledge funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth and Development Office (FCDO), under the MRC/FCDO Concordat agreement and also part of the EDCTP2 program supported by the European Union. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

Conceptualization, investigation, methodology, formal analysis, funding acquisition and writing (original draft preparation): K.V.P. Validation: K.V.P. and A.E.Z. Writing (review and editing): all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-022-00313-1.

Correspondence and requests for materials should be addressed to Kris V. Parag.

Peer review information *Nature Computational Science* thanks Lauren McGough, Laura White and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Fernando Chirigati, in collaboration with the *Nature Computational Science* team. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.