Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The uncertainty of infectious disease outbreaks is underestimated

Matthew Penn (matthew.penn@st-annes.ox.ac.uk)
University of Oxford
Daniel Laydon (d.laydon@imperial.ac.uk)
Imperial College London https://orcid.org/0000-0003-4270-3321
Joseph penn (joseph.penn@seh.ox.ac.uk)
University of Oxford
Charles Whittaker (charles.whittaker16@imperial.ac.uk)
Imperial College London https://orcid.org/0000-0002-5003-2575
Christian Morgenstern (c.morgenstern@imperial.ac.uk)
Imperial College London https://orcid.org/0000-0003-3735-1130
Oliver Ratmann (oliver.ratmann05@imperial.ac.uk)
Imperial College London
Swapnil Mishra (swapnil.mishra@sund.ku.dk)
University of Copenhagen
Mikko Pakkanen (m.pakkanen@imperial.ac.uk)
Imperial College London
Christl Donnelly (christl.donnelly@stats.ox.ac.uk)
Oxford University https://orcid.org/0000-0002-0195-2463
Samir Bhatt (samir.bhatt@sund.ku.dk)
University of Copenhagen https://orcid.org/0000-0002-0891-4611

Article

Keywords:

DOI: https://doi.org/

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

The uncertainty of infectious disease outbreaks is underestimated

Matthew J. Penn[1,*], Daniel J. Laydon[2], Joseph Penn[1], Charles Whittaker[2], Christian Morgenstern[2], Oliver Ratmann[2], Swapnil Mishra[3], Mikko S. Pakkanen[4,2], Christl A. Donnelly[1,2], and Samir Bhatt[2,3,*]

[1]University of Oxford

[2]Imperial College London

[3]University of Copenhagen

[4]University of Waterloo

[*]corresponding authors

Abstract

Uncertainty can be classified as either aleatoric (intrinsic randomness) or epistemic (imperfect knowledge of parameters). The majority of frameworks assessing infectious disease risk consider only epistemic uncertainty. We only ever observe a single epidemic, and therefore cannot empirically determine aleatoric uncertainty. Here, for the first time, we characterise both epistemic and aleatoric uncertainty using a time-varying general branching processes. Our framework explicitly decomposes aleatoric variance into mechanistic components, quantifying the contribution to uncertainty produced by each factor in the epidemic process, and how these contributions vary over time. The aleatoric variance of an outbreak is itself a renewal equation where past variance affects future variance. Perhaps surprisingly, superspreading is not necessary for substantial uncertainty, and profound variation in outbreak size can occur even without overdispersion in the offspring distribution (i.e. the distribution of the number

of secondary infections an infected person produces). Crucially, aleatoric forecasting uncertainty grows dynamically and rapidly, and so forecasting using only epistemic uncertainty is a significant underestimate. Therefore, failure to account for aleatoric uncertainty will ensure that policymakers are misled about the substantially higher true extent of potential risk. We demonstrate our method, and the extent to which potential risk is underestimated, using two historical examples: firstly the 2003 Hong Kong severe acute respiratory syndrome (SARS) outbreak, and secondly the early 2020 UK COVID-19 epidemic. Our framework provides analytical tools to estimate epidemic uncertainty with limited data, to provide reasonable worst-case scenarios and assess both epistemic and aleatoric uncertainty in forecasting, and to retrospectively assess an epidemic and thereby provide a baseline risk estimate for future outbreaks. Our work strongly supports the precautionary principle in pandemic response.

Infectious diseases remain a major cause of human mortality. Understanding their dynamics is essential for forecasting cases, hospitalisations, and deaths, and to estimate the impact of interventions. The sequence of infection events defines a particular epidemic trajectory – the outbreak – from which we infer aggregate, population-level quantities. The mathematical link between individual events and aggregate population behaviour is key to inference and forecasting. The two most common frameworks for modelling aggregate data are susceptible-infected-recovered (SIR) models [18] or renewal equation models [15, 28]. Under certain specific assumptions, these frameworks are deterministic and equivalent to each other [7]. While stochastic versions exist [2, 28], most analytic frameworks make strong simplifying assumptions (e.g. Markov or Gaussian) regarding the probabilities of individual events that lead to emergent aggregate behaviour.

Randomness at individual level affects randomness at population level. We can classify uncertainty as either aleatoric (due to randomness) or epistemic (imprecise knowledge of parameters) [19]. Consider fitting a renewal equation to observed incidence data [28], where infectiousness is known but the rate of infection events $\rho(\cdot)$ must be fitted. The secondary infections produced by an infected individual will occur randomly over their infectious period g, depending on their infectiousness ν. The population mean rate of infection events is given by $\rho(t)$, and we assume that this mean does not differ between individuals (although each individual has a different random draw of their number of secondary infections). In Bayesian settings, inference yields multiple posterior estimates for ρ, and therefore multiple incidence values. This is epistemic uncertainty: any given

value of ρ corresponds to a single realisation of incidence. However, each posterior estimate of ρ is in fact only the mean of an underlying offspring distribution (i.e. the distribution of the number of secondary infections an infected person produces). If an epidemic governed by identical parameters were to happen again, but with different random draws of infection events, each realisation would be different, thus giving aleatoric uncertainty.

Infectious disease models usually consider epistemic uncertainty only, and those that consider aleatoric uncertainty (e.g. individual-based models) are challenging to use in inference. We can characterise aleatoric uncertainty by the final epidemic size from multiple historical outbreaks [8, 34] but these are confounded by temporal, cultural, epidemiological, and biological context, and therefore parameters vary between each outbreak. Here we analyse aleatoric uncertainty by studying an epidemiologically-motivated stochastic process, serving as a proxy for repeated realisations of an epidemic. We find that using epistemic uncertainty alone is a vast underestimate, and accounting for aleatoric uncertainty shows potential risk to be much higher. We demonstrate our method using two historical examples: firstly the 2003 severe acute respiratory syndrome (SARS) outbreak in Hong Kong, and secondly the early 2020 UK COVID-19 epidemic.

An analytical framework for aleatoric uncertainty

A time-varying general branching processes proceeds as follows: first, an individual is infected, and their infectious period is distributed with probability density function g (with corresponding cumulative distribution function G). Second, while infectious, individuals randomly infect others (via a counting process with independent increments), driven by their infectiousness $\nu(t)$ and a rate of infection events $\rho(t)$ i.e. an individual infected at time l, will generate secondary infections at a rate $\rho(t)$ and proportional to their infectiousness $\nu(t-l)$) per unit time. $\rho(t)$ is closely related to the time-varying reproduction number $R(t)$ (see Methods and [28] for further details). We allow multiple infection events to occur simultaneously, and assume individuals behave independently once infected, thus allowing mathematical tractability [16]. Briefly, we model an individual's secondary infections using a stochastic counting process, which gives rise to secondary infections (i.e. offspring) that are either Poisson or Negative Binomial distributed in their number, and Poisson distributed in their timing (see Methods and Supplementary Information for details). We study

the aggregate of these events (prevalence or incidence) through closed-form probability generating functions and probability mass functions. Our approach models epidemic evolution through intuitive individual-level characteristics while retaining analytical tractability. Importantly, the mean of our process follows a renewal equation [1, 28, 29]. Our formulation unifies mechanistic and individual-based modelling within a single analytical framework based on branching processes. Figure 1 shows a schematic of this process. Formal derivation is in the supplementary information.

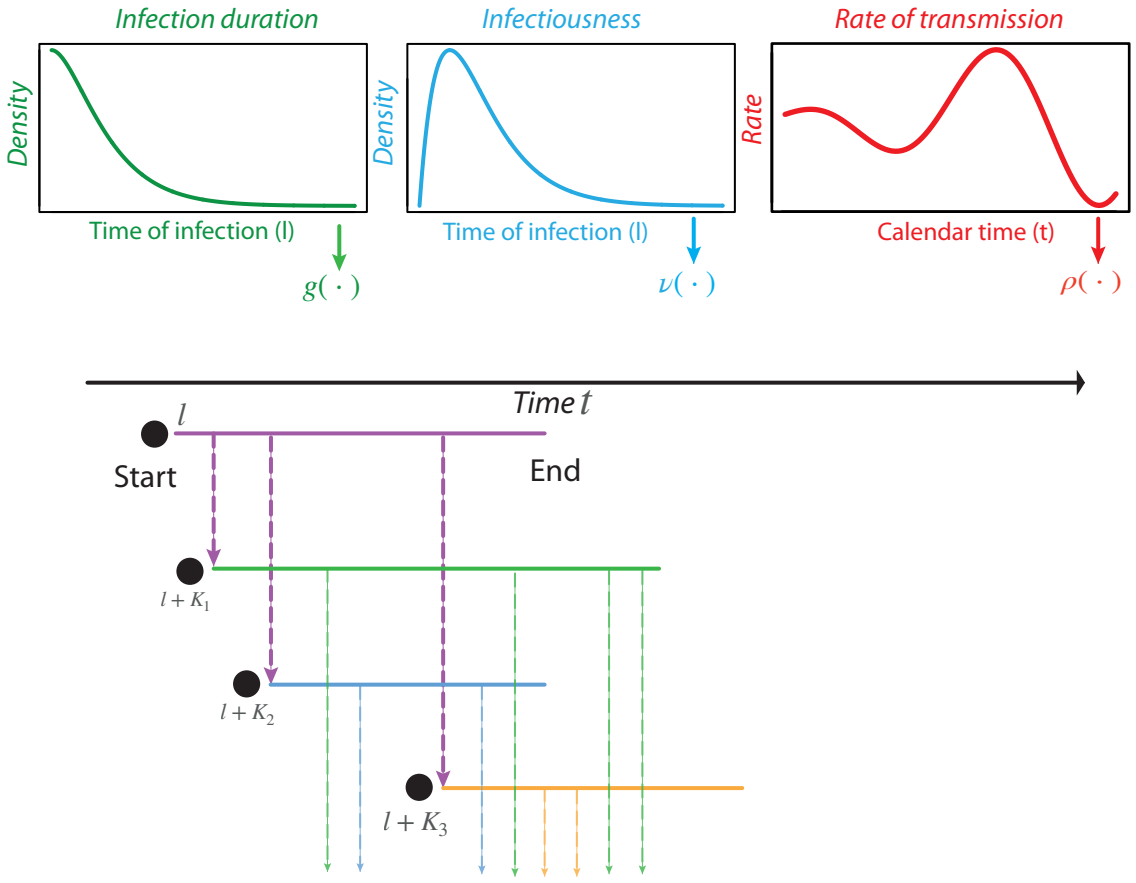


Figure 1: Schematic of a time-varying general branching process. Top row shows schematics for the infectious period (left), an individual's time-varying infectiousness (middle) (both functions of time post infection t^*), and the population-level mean rate of infection events (right). The infectious period is given by probability density function g. For each individual their (time-varying) infectiousness and rate of infection events are given by ν and ρ respectively. In the example above, an individual is infected at time l, and infects three people (random variables K, black dashed lines) at times $l + K_1$, $l + K_2$ and $l + K_3$. The times of these infections are given by a random variable with probability density function $\sim \frac{\rho(t)\nu(t-l)}{\int_l^t \rho(u)\nu(u-l)du}$. Each new infection then has its own infectious period and secondary infections (thinner coloured lines).

Randomness occurs at individual level, and there is a distribution of possible realisations of the epi-

demic given identical parameters. Simulating our general branching process would be cumbersome using the standard approach of Poisson thinning [27], and inference from simulation is more challenging still. Using probability generating functions, we analytically derive important quantities from the distribution of the number of infections, including the (central) moments and marginal probabilities given ρ, g and ν (with or without epistemic uncertainty). We additionally use the probability generating function to prove general, closed-form, analytical results such as the decomposition of variance into mechanistic components, and the conditions under which overdispersion exists. Finally, we derive a general probability mass function (likelihood function) for incidence.

If infection event $k = 0, \ldots, n$ occurred at time τ_k and produced y_k infections, let x_{kj} denote the end time of the infectious period of the j^{th} infection at event k. Note that $\tau_0 = l$ is the time of the first infection event and $y_0 = 1$. Then the likelihood $L_{\text{InfPeriod}}$ of each infected person's infectious period is a product over all infections given by

$$L_{\text{InfPeriod}} = \prod_{k=0}^{n} \prod_{j=1}^{y_k} g(x_{kj} - \tau_k, \tau_k). \tag{1}$$

The likelihood of there being y_k infections at time τ_k is given by

$$L_{\text{InfTime}} = \prod_{k=1}^{n} \left(\sum_{i=0}^{k-1} \sum_{j=1}^{y_i} \mathbb{1}_{\{x_{ij} < \tau_k\}} p_{y_k}(\tau_k, \tau_i) \right), \tag{2}$$

where $p_{y_k}(\tau_k, \tau_i)$ is the (infinitesimal) rate at which an individual infected at τ_i causes y_k infections at time τ_k, provided it is still infectious. Finally, the probability that no other infections occurred between the infection events at times $(\tau_k)_{k=0}^{n}$ is given by

$$L_{\text{Only}} = \exp\left(- \sum_{i=0}^{n} \sum_{j=1}^{y_i} \int_{\tau_i}^{\min(t, x_{ij})} r(u, \tau_i) du \right), \tag{3}$$

where r is the "infection event" rate and t is the current time. Our full likelihood L_{Full} is then

$$L_{\text{Full}} = L_{\text{InfPeriod}} \times L_{\text{InfTime}} \times L_{\text{Only}}. \tag{4}$$

Full derivations of these quantities are provided in the supplementary information. If discrete time is assumed and we ignore death times equation 4 simplifies to a likelihood commonly used for inference [9]. Markov Chain Monte Carlo can be used on equation 4 to sample aleatoric incidence realisations, but it is often simpler to solve the probability generating function with complex

integration. The probability generating function, equations for the variance, and derivations of the probability mass function are found in the supplementary information and a summary of the main analytical results is found in the Methods.

The dynamics of Uncertainty

We derive the mean and variance of our branching process. The general variance Equation 7 (see Methods) captures uncertainty in prevalence over time, where individual-level parameters govern each infection event. This equation comprises three terms: the timing of secondary infections from the infectious period (Equation 7a); the offspring distribution (Equation 7b); and propagation of uncertainty through the descendants of the initial individual (Equation 7c). Importantly, this last term depends on past variance, showing that the infection process itself contributes to aleatoric variance, and does not arise only from uncertainty in individual-level events. In short, unlike common Gaussian stochastic processes, the general variance in disease prevalence is described through a renewal equation. Therefore, future uncertainty depends on past uncertainty, and so the uncertainty around subsequent epidemic waves has *memory*. Additionally, uncertainty is driven by a complex interplay of time-varying factors, and not simply proportional to the mean. For example, a large first wave of infection can increase the variance of the second wave. As such, the general variance equation 7 disentangles and quantifies the causes of uncertainty, which remain obscured in brute-force simulation experiments [2].

Consider a toy simulated epidemic with $\rho(t) = 1.4 + \sin(0.15t)$, where the offspring distribution is Poisson in both timing and number of secondary infections, and where infectiousness ν is given by the probability density function $\nu \sim \text{Gamma}(3, 1)$, and, similarly, the infectious period $g \sim$ Gamma(5,1). The resulting variance is counterintuitive. We prove analytically that overdispersion emerges despite a non-overdispersed Poisson offspring distribution. The second wave has a lower mean but a higher variance than the first wave (Figure 2), because uncertainty is propagated. If the variance were Poisson, i.e. equal to the mean, the second wave would instead have a smaller variance due to fewer infections. Initially, uncertainty from individuals is largest, but as the epidemic progresses, compounding uncertainty propagated from the past dominates [Figure 2, bottom right]. Note that in this example with zero epistemic uncertainty (we know the parameters

perfectly), aleatoric uncertainty is large.

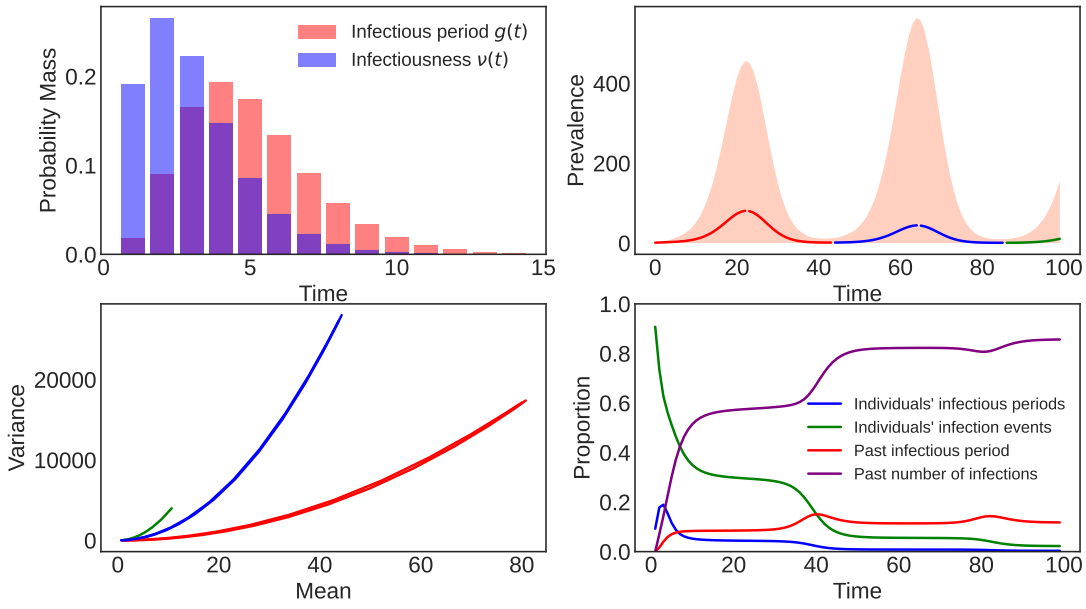


Figure 2: Aleatoric uncertainty without overdispersed offspring distribution. Plots show simulated epidemic where $\rho(t) = 1.4 + \sin(0.15t)$, with a Poisson offspring distribution. We use infectiousness $\nu \sim \text{Gamma}(3, 1)$, and infectious period $g \sim \text{Gamma}(5,1)$. (Top left) Overlap between g and the infectiousness ν, where g controls when the infection ends e.g. by isolation. (Top right) Predicted mean and 95% aleatoric uncertainty intervals for prevalence. Note there is no epistemic uncertainty as the parameters are known exactly (Bottom left) Phase plane plot showing the mean plotting against the variance. (Bottom right) Proportional contribution to the variance from the individual terms in Equation 7. Compounding uncertainty from past events is the dominant contributor to overall uncertainty.

In Equation 7, the first two terms account for uncertainty in the infectious periods of all infected individuals. The third term denotes the uncertainty from the offspring distribution. By construction, the *timing* of infections is an inhomogenous Poisson process, where at each infection time the *number* of infections is random. The third term (Equation 7b) contains the second moment of the offspring distribution, which is the variability around its mean (i.e. $\rho(t)$). The second mo-

ment quantifies the extent of possible "superspreading". In contrast to other studies [23, 35], we find that individual-level overdispersion in the offspring distribution is less important than explosive epidemics. Under a null Poisson model, with no overdispersion (see Poisson case in Figure 2), substantial aleatoric uncertainty arises from a Poisson offspring distribution combined with variance propagation. We rigorously prove via the Cauchy-Schwarz inequality that, under a mild condition on the possible spread of the epidemic, the variance of number of infections at a given time is always greater than the mean, and hence is overdispersed (see supplementary information). Overdispersion in the offspring infection distribution is therefore not necessary for high aleatoric uncertainty, although it still increases variance at both individual-level and population-level.

We derive the conditional variance, with known past events but unknown future events. Conditional variance grows proportionally to the square of the mean, with additional terms containing the previous variance. Therefore aleatoric uncertainty grows and forecasting exercises based only on epistemic uncertainty greatly underestimates the risk of very large epidemics, and this underestimation becomes more severe as the forecast horizon expands or as the epidemic grows.

Aleatoric uncertainty over the SARS 2003 epidemic

To demonstrate the importance of aleatoric uncertainty, we analyse daily incidence of symptom onset in Hong Kong during the 2003 severe acute respiratory syndrome (SARS) outbreak [10, 17, 21]. The epidemic struck Hong Kong in March-May 2003, with a case fatality ratio of 15%. We fit a Bayesian renewal equation assuming a random walk prior distribution for the rate of infection events ρ [28], using Equation 4 for inference. We ignore g and assume that the distribution of generation times mirrors the distribution of infectiousness, i.e. that the infectiousness ν equals the generation time [21]. Note these parameter choices are illustrative and do not affect our main conclusions. The fitted $\rho(t)$ in Figure 3 (top left) shows two major peaks, consistent with the major transmission events in the epidemic [17]. Figure 3 (top right) shows the mean epistemic fit, with epistemic (posterior) uncertainty tightly distributed around the data. Figure 3 (bottom left) shows the aleatoric uncertainty under optimistic and pessimistic scenarios (i.e. the upper and lower bounds of $\rho(t)$ in Figure 3 (top right)). The pessimistic scenario includes the possibility of extinction, but also an epidemic that could have been more than six times larger than that

observed. The optimistic scenario suggests we would observe an epidemic of at worst comparable size to that observed. Finally, Figure 3 (bottom right) shows epistemic and aleatoric forecasts at day 60 of the epidemic, fixing $\rho(t)$ using the 95% epistemic uncertainty interval to be constant at either $\rho(t \geq 60) = 0.38$ or $\rho(t \geq 60) = 0.83$ and simulating forwards. While the epistemic forecast does contain the true unobserved outcome of the epidemic, it underestimates true forecast uncertainty, which is 1.3 times larger. The range of the constant ρ for forecast is below 1, and yet we still see substantial aleatoric uncertainty. If ρ were above 1 for a sustained period, aleatoric uncertainty would play a smaller role [5], but this is rare with real epidemics, where susceptible depletion, behavioural changes or interventions keep ρ around 1. Our results therefore highlight that epistemic uncertainty drastically underestimates potential epidemic risk.

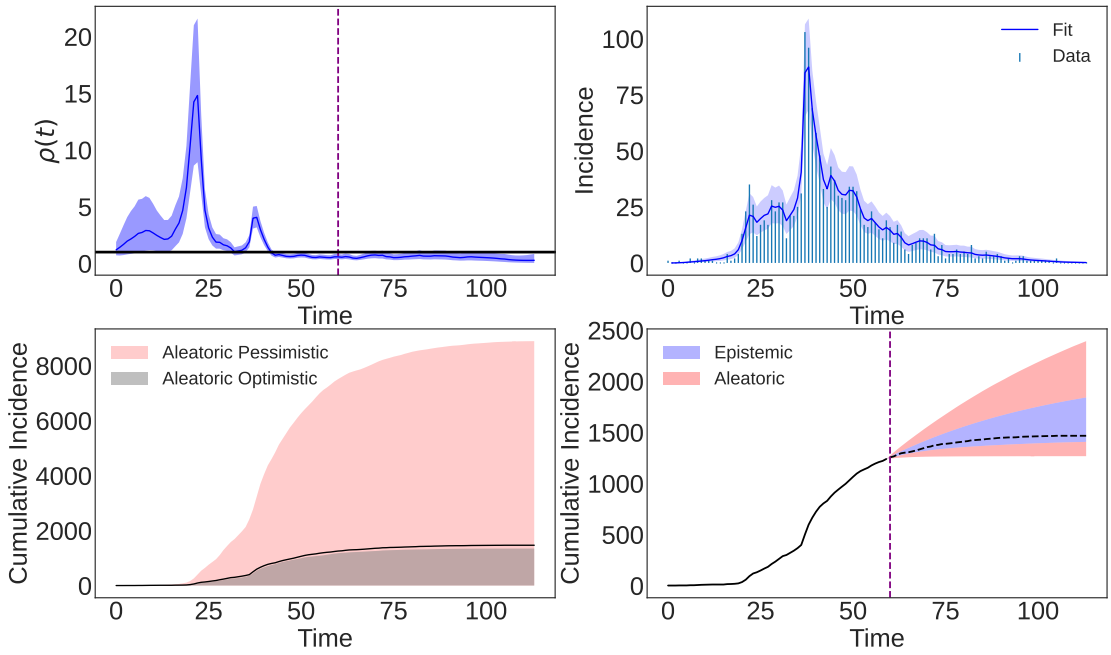


Figure 3: Aleatoric and epistemic uncertainty of the 2003 severe acute respiratory syndrome (SARS) epidemic in Hong Kong [10, 21]. (Top left) $\rho(t)$ with 95% epistemic uncertainty. (Top right) Fitted incidence mean, 95% epistemic uncertainty with observational noise from using Equation 4. Data is daily incidence of symptom onset. (Bottom left) Aleatoric uncertainty from the start of the epidemic under an optimistic and pessimistic $\rho(t)$. (Bottom right) Aleatoric uncertainty in forecasting keeping ρ constant at the forecast data (dotted line) Forecasting is from day 60.

Aleatoric risk assessment in the early 2020 COVID-19 pandemic in the UK

To demonstrate the practical application of our model, we retrospectively examine the early stage of the COVID-19 pandemic in the UK, using only information available at the time. While the date of the first locally transmitted case in the UK remains unknown (likely mid-January 2020 [30]), COVID-19 community transmission was confirmed in the UK by late January 2020, and

we therefore start our simulated epidemic on January 31st 2020. We consider uncertainty in the predicted number of deaths on March 16th 2020 [12], during which time decisions regarding non-pharmaceutical interventions were made. Testing was extremely limited during this period, and COVID-19 death data were unreliable. For this illustration, we assume that we did not know the true number of COVID-19 deaths, as was the case for many countries in early 2020. Policymakers then needed estimates of the potential death toll, given limited knowledge of COVID-19 epidemiology and unreliable national surveillance.

We simulated an epidemic from a time-varying general branching process with a Negative Binomial offspring distribution, using parameters that were largely known at by March 16th 2020 (Table 1). The infection fatality ratio, infection-to-onset distribution and onset-to-death distribution were convoluted with incidence [28] to estimate numbers of deaths. Estimated COVID-19 deaths and uncertainty estimates between January 31st and March 16th 2020 are shown in Figure 4 (Top). While the epistemic uncertainty contains the true number of deaths, it is still an underestimate, and including aleatoric uncertainty, we find that the epidemic could have had more than four times as many deaths. Consider a hypothetical intervention on March 17th 2020 (Figure 4 (bottom)) that completely stops transmission. Deaths would still occur from those already infected but no new infections would arise. In this hypothetical case, the aleatoric uncertainty would still be 2.5 times the actual deaths that occurred (when in fact transmission was never zero or close to it). This hypothetical scenario highlights the scale of aleatoric uncertainty, and demonstrates that our method can be useful in assessing risk in the absence of data by giving a reasonable worst case. Further, we observe that using only epistemic uncertainty provides a reasonably good fit in a relatively short time-horizon (Figure 4, Top), but soon afterwards greatly underestimates uncertainty (Figure 4, Bottom). The fits using aleatoric uncertainty provide a more reasonable assessment of uncertainty. While we concentrate on the upper bound, the lower bound on the worst-case scenario still exceeds zero, and therefore the epidemic going extinct by March 16th in the worst-case with no external seeding would have been very unlikely. Aleatoric uncertainty highlights a more informative reasonable worst-case estimate than epistemic uncertainty alone, and could be a useful metric for a policymaker in real time, with low-quality data, without requiring simulations from costly, individual-based models.

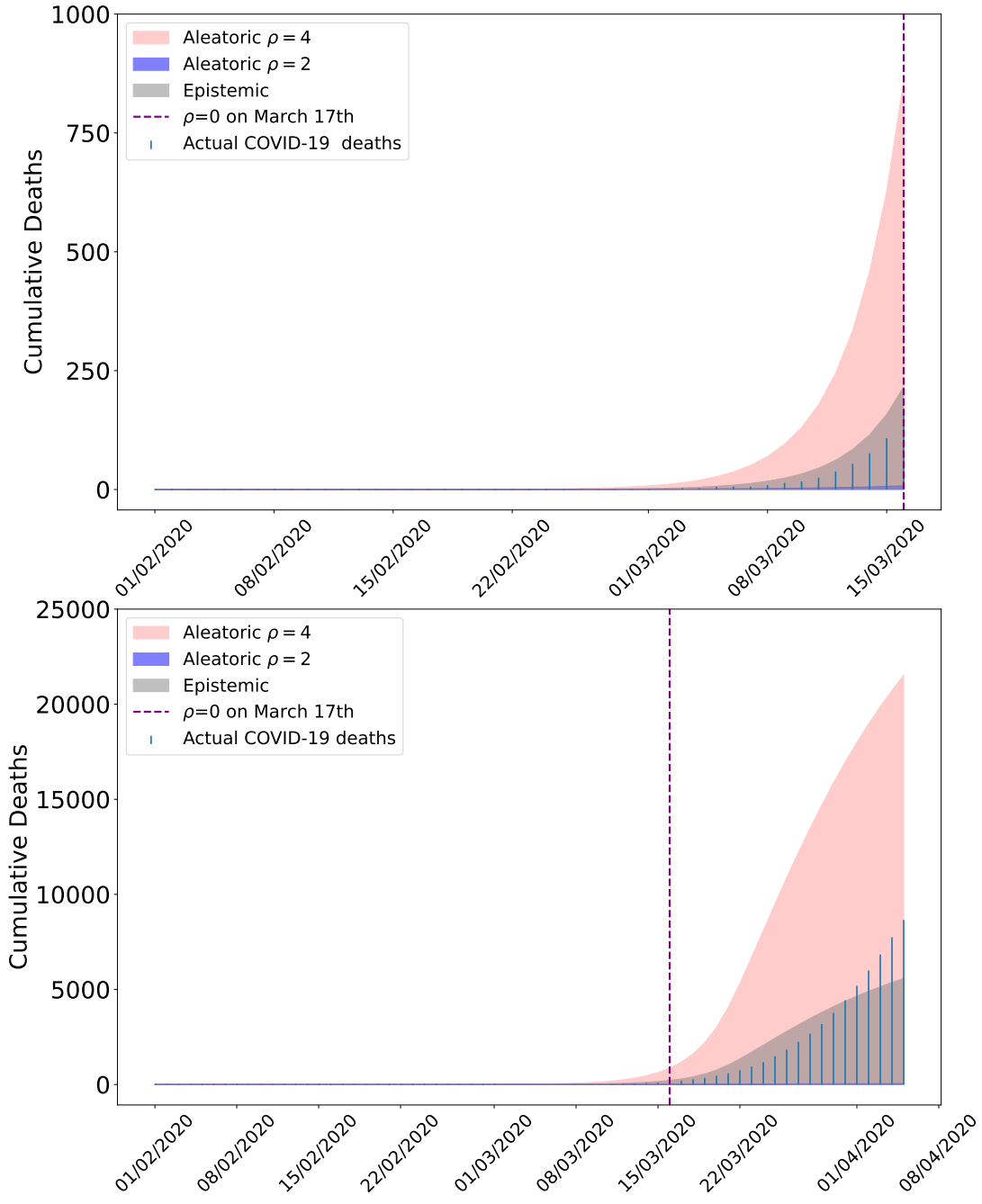


Figure 4: Early 2020 COVID-19 pandemic in the UK. Top plot shows a simulated epidemic using parameters available on March 16th 2020 (Table 1), for a plausible range of $\rho = R_0$ between 2 and 4. Blue bars indicate actual COVID-19 deaths, which we assume no knowledge of. The purple line is March 17th 2020, we set transmission to zero i.e. $\rho = 0$, to simulate an intervention that stops transmission completely. The grey envelope is the epistemic uncertainty and the red envelope the aleatoric uncertainty. Bottom plot is the same as the top plot, except time is extended past March 17th with transmission being zero.

Epidemiological Parameter	**Value or Distribution**	**Citation**
Infection Fatality Ratio	0.9%	[6, 32]
Basic Reproduction Number	$2-4$	[12, 22]
Serial Interval Distribution	$\sim$ Gamma$(7.82, \frac{1}{0.62})$	[14, 31, 32]
Onset-to-Death Distribution	$\sim$ Gamma$(1.45, 10.43)$	[25, 32]
Infection-to-Onset Distribution	$\sim$ Gamma$(35.16, 6.9)$	[14, 32]
Overdispersion Coefficient	0.53	[20]

Table 1: Epidemiological parameters available on March 16th 2020 used in branching process simulation

Towards a more realistic quantification of uncertainty

Stochastic models more realistically model natural phenomena than deterministic equations [26], and particularly so with infection processes [3]. Accordingly, individual-based models have found much success [13, 33] in capturing the complex dynamics that emerge from infectious disease outbreaks, and have been highly influential in policy [12]. However, analytical frameworks still tend to be deterministic [10, 11, 14], and only consider statistical, epistemic parameter uncertainty. Frameworks that expand deterministic, mechanistic equations to include stochasticity use a Gaussian noise process [2], or restrict the process to be Markovian. Markovian branching processes require the infection period or generation time to be exponentially distributed - a fundamentally unrealistic choice for most infectious diseases. Further, a Gaussian noise process is unlikely to be realistic [8].

Our results show that individual-level uncertainty is overshadowed by uncertainty in the infection process itself. Profound overdispersion in infectious disease epidemics is not simply a result of overdispersion in the offspring distribution, but is fundamental and inherent to the branching process. We rigorously prove that even with a Poisson offspring distribution (not characterized by overdispersion), overdispersion in resulting prevalence or incidence is still virtually always guaranteed. We show that forecast uncertainty increases rapidly, and therefore common forecasting methods almost certainly underestimate true uncertainty. Our framework provides a new method-

ological tool to evaluate uncertainty in the presence of little to no data, assess uncertainty in forecasting, and retrospectively assess an epidemic.

We have considered only a small number of mechanisms that generate uncertainty. Cultural, behavioural and socioeconomic factors will introduce even greater randomness. Therefore our framework also underestimates of true uncertainty in infectious disease epidemics, underlining that potential risk is much higher than previously thought.

Funding

S.B. C.A.D and D.J.L. acknowledge support from the MRC Centre for Global Infectious Disease Analysis (MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement, and also part of the EDCTP2 programme supported by the European Union. S.B. acknowledges support from the Novo Nordisk Foundation via The Novo Nordisk Young Investigator Award (NNF20OC0059309), which also supports S.M.Ṡ.B. acknowledges support from the Danish National Research Foundation via a chair position. S.B. and C.M. acknowledges support from The Eric and Wendy Schmidt Fund For Strategic Innovation via the Schmidt Polymath Award (G-22-63345). S.B. acknowledges support from the National Institute for Health Research (NIHR) via the Health Protection Research Unit in Modelling and Health Economics. D.J.L. acknowledges funding from Vaccine Efficacy Evaluation for Priority Emerging Diseases (VEEPED) grant, (ref. NIHR:PR-OD-1017-20002) from the National Institute for Health Research. M.J.P. acknowledges funding from a EPSRC DTP Studentship. C.W. acknowledges support from the Wellcome Trust.

Data and code availability

All model code to reproduce Figures 2, 3 and 4 is available at `https://github.com/MLGlobalHealth/uncertainity_infectious_diseases.git`. Data from Figure 3 is available via the R-Package EpiEstim[10], and data from Figure 4 is available at `https://imperialcollegelondon.github.io/covid19local` and via official UK Government reporting (`https://www.ons.gov.uk/`).

Methods

A time-varying general branching process proceeds as follows: first, a single individual is infected at some time l, and their infectious period L is distributed with probability density function g (and cumulative distribution function G). Second, during their infectious period, they randomly infect other individuals, affected by their infectiousness $\nu(t-l)$, and their mean number of secondary infections, which is assumed to be equal to the population-level rate of infection events $\rho(t)$. $\rho(t)$ is closely related to the time-varying reproduction number $R(t)$ (see [28] for details). The infectious period g accounts for variation in individual behaviour. If people take preventative action to reduce onward infections, their reduced infection period can stop transmission despite remaining infectious. Where infectious individuals do not change their behaviour, g can be ignored and individual-level transmission is controlled by infectiousness ν only. Each newly infected individual then proceeds independently by the same mechanism as above.

Formally, if an individual is infected at time s, their number of secondary infections is given by a stochastic counting process $\{N(t,s)\}_{t\geq s}$, which is independent other individuals and has independent increments. We assumehere that the epidemic occurs in continuous time, and hence that $N(t,s)$ is continuous in probability, although we consider discrete-time epidemics in the supplementary information. To aid calculation, we suppose $N(t,s)$ can be defined from a Lévy Process $\mathcal{N}(t)$ - that is, a process with both independent and identically distributed increments - via $N(t,s) = \mathcal{N}\left(\int_s^t r(k,s)dk\right)$ for some non-negative rate function r. It is assumed that each counting process $\{N(t,s)\}_{t\geq s}$ is defined from an independent copy of $M(t)$. This formulation has two advantages: first, the dependence of $N(t,s)$ on s is restricted to the rate function r; and second, if $J_{\mathcal{N}}(t)$ counts the number of infection events in $\mathcal{N}(t)$ (where here "infection events" refer to an increase, of any size, in $N(t,s)$), then $J_{\mathcal{N}}(t)$ is a Poisson process with some rate κ [4]. We can then define $J(t,s)$ to be the counting process of infection events in $N(t,s)$, and $Y(v)$ to be size of the infection event (i.e. the number of secondary infections that occur) t time v. We assume that Y is independent of s, although such a dependence would curtail superspreading to depend on infectiousness, and could be incorporated into the framework. Therefore $J(t,s)$ is an inhomogeneous Poisson Process (and so $N(t,s)$ has been characterised as an inhomogeneous compound Poisson Process). We consider the cases where $N(t,s)$ is itself an inhomogeneous Poisson process,

and where $N(t, s)$ is a Negative Binomial process. This allows us to examine effects of overdispersion in the number of secondary infections, although our framework allows for more complicated distributions.

Here, $r(t, l) = \rho(t)\nu(t - l)$ where $\rho(t)$ models the population-level rate of infection events, and $\nu(t - l)$ models the infectiousness of an individual infected at time l. If $\nu(t - l)$ is sufficiently well characterised by the generation time (i.e. where the timing of secondary infections mirrors tracks their infectiousness) , and the infectious period can be ignored, then the integral $\int_l^t r(s, l)ds$ has the same scale as the commonly used reproduction number $R(t)$ [28]. The branching process yields a series of "birth" and "death" times for each individual (i.e. the time of infection and the end of the infectious period respectively), from which prevalence (the number of infections at any given time) or cumulative incidence (the total number of infections up to any time) can be defined.

Probability generating function

We derive the probability generating function for a time-varying age-dependent branching process, allowing derivation of the mean and higher-order moments. We consider two special cases for the number of new infections $Y(v)$ at each infection event: a Poisson distribution and a logarithmic (log series) distribution. In both cases, we assume that the distribution of $Y(v)$ is equal for all values of v. In the Poisson case, the number of new infections at each infection time is, by definition, one. Therefore the number of infections an individual creates is Poisson distributed, and closely clustered around the mean rate of infection events. The logarithmic case, which causes $N(t, l)$ to be a Negative Binomial process, more realistically allows multiple infections to occur at each infection time, and so the number of infections an individual causes is overdispersed. We derive (with details in the supplementary information) the following pgf (probability generating function):

$F(t,l;s) = E(s^{Z(t,l)})$, for either the prevalence or the cumulative incidence

$$\underbrace{F(t,l;s)}_{\text{pgf}} = \underbrace{\big(1-G(t-l,l)\big)}_{P(L>t-l)}\, q_1\underbrace{\left(\int_0^{t-l} \overbrace{f(\underbrace{F(t,l+k;s)}_{\text{pgf of process started at } l+k})}^{\text{pgf of } Y}\ \overbrace{\rho(l+k)\nu(k)dk}^{\text{infection rate at time } l+k}\right)}_{\text{pgf of } J(t,l)}$$

$$+\int_0^{t-l} \underbrace{q_2\left(\int_0^{u} \overbrace{f(\underbrace{F(t,l+k;s)}_{\text{pgf of process started at } l+k})}^{\text{pgf of } Y}\ \overbrace{\rho(l+k)\nu(k)}^{\text{infection rate at time } l+k}\ dk\right)}_{\text{pgf of } J(l+u,l)} \underbrace{g(u,l)du}_{P(L=l+u)}, \quad (5)$$

where $q_1(z;s) = se^z$, and where $q_2(z) = e^z$ in the case where $Z(t,l)$ refers to prevalence, whereas $q_2(z;s) = se^z$ in the case where $Z(t,l)$ refers to cumulative incidence. Note also that $f(z) = z-1$ in the Poisson case and $f(z) = -\phi\left[\log\left(1-(1-\frac{\phi}{1+\phi}z)\right) - \log\left(\frac{\phi}{1+\phi}\right)\right]$ in the log-series case and that the constant κ has absorbed into ρ.

The key intuition in understanding Equation 5 is that for an integer random variable X and iid (independent and identically distributed) random variables Y_i, $E(s^{\sum_{i=1}^{X} Y_i}) = G_X(G_Y(s))$, where G_X and G_Y are the generating functions of X and Y_i respectively. Thus, we expect the pgfs of the various parts of our model to combine via composition, as occurs in the equation above.

Mean incidence can recovered from both prevalence (via back calculation [28]) and cumulative incidence. In Equation 5 for the Negative Binomial case, ϕ is the degree of overdispersion. Equation 5 is solvable using via quadrature and the fast Fourier transform via a result from complex analysis [24] and scales easily to populations with millions of infected individuals, and the probability mass function can be computed to machine precision.

Variance decomposition

For simplicity, we only summarise the decomposition for prevalence, but an analogous and highly similar derivation for cumulative incidence can be found in the supplementary information. We can derive an analytical equation for the mean and variance of the entire branching process. The

mean prevalence $M(t,l)$ is given by

$$M(t,l) = (1 - G(t-l,l)) + \int_0^{t-l} M(t,l+u)\rho(l+u)\nu(u)\mathbb{E}(Y)(1-G(u,l))du. \tag{6}$$

Note, ρ can be scaled to absorb the $E(Y)$ and κ constants. Equation 6 is consistent with that previously derived in [28]. The second moment, $W(t,l) := \mathbb{E}(Z(t,l)(Z(t,l)-1))$ allows us to determine the variance, $V(t,l)$ as $V(t,l) = W(t,l) + M(t,l) - M(t,l)^2$. The variance can be decomposed into three mechanistic components (see supplementary information).

$$\begin{aligned}
V(t,l) = & \underbrace{\int_0^{t-l} \left[\int_0^u M(t,l+k)\rho(l+k)\nu(k)dk\right]^2 g(u,l)du - M(t,l)^2}_{\text{(7a): uncertainty from the infectious period}} \\
& + \underbrace{(1-G(t-l,l))\left[1 + 2\int_0^{t-l} M(t,l+u)\rho(l+u)\nu(u)du + \left(\int_0^{t-l} M(t,l+u)\rho(l+u)\nu(u)du\right)^2\right]}_{\text{(7a continued): uncertainty from the infectious period}} \\
& + \underbrace{\int_0^{t-l} M(t,l+u)^2\mathbb{E}(Y^2)\rho(l+u)\nu(u)(1-G(u,l))du}_{\text{(7b): uncertainty from the offspring distribution}} \\
& + \underbrace{\int_0^{t-l} V(t,l+u)\rho(l+u)\nu(u)(1-G(u,l))du}_{\text{(7c): uncertainty propagated from the past}}.
\end{aligned} \tag{7}$$

The general variance equation 7 captures the evolution of uncertainty in population-level disease prevalence over time, where fixed individual-level disease transmission parameters govern each infection event. Unlike the simple Galton-Watson process, we find that previously unknown factors also determine aleatoric variation in disease prevalence. Specifically, the general variance equation 7 comprises three terms, one for the infectious period (Equation 7a), one for the number and timing of secondary infections (Equation 7b), and a term that propagates uncertainty through descendants of the initial individual (Equation 7c). Importantly, the last term (Equation 7c) depends on past variance, showing that the infection process itself contributes to aleatoric variance, and this is distinct from the uncertainty in individual infection events. In short, and unlike Gaussian stochastic processes, the general variance in disease prevalence is described through a renewal equation. Intuitively then, uncertainty in an epidemic's future trajectory is contingent on past infections, and that the uncertainty around consecutive epidemic waves are connected. As such, the general variance equation 7 allows us to disentangle important aspects of infection dynamics that remain obscured in brute-force simulations [2].

Overdispersion

We define an epidemic to be "expanded" if at time t there is a non-zero probability that the prevalence, not counting the initial individual or its secondary infections, is non-zero.

Note that this is a very mild condition on an epidemic - in a realistic setting, the only way for an epidemic to not be expanded is if it is definitely extinct by time t, or if t is small enough that tertiary infections have not yet occurred.

Large aleatoric variance intrinsic to our branching process implies that the prevalence of new infections (that is, prevalence excluding the deterministic initial case) is always strictly overdispersed at time t, providing the epidemic is expanded at time t. A full proof is given in the supplementary information, but we provide here a simpler justification in the special case that $G(t-l,l)=1$.

In this case, prevalence of new infections is equal to standard prevalence, and the equations for $M(t,l)$ and $V(t,l)$ simplify significantly. Switching the order of integration in the equation for $M(t,l)$ gives

$$M(t,l)=\int_0^{t-l} M(t,l+u)\rho(l+u)\nu(u)\mathbb{E}(Y)(1-G(u,l))du=\int_0^{t-l}\left[\int_0^u M(t,l+k)\rho(l+k)\mathbb{E}(Y)\nu(k)dk\right]g(u,l)du \tag{8}$$

and hence, the Cauchy-Schwarz Inequality shows that

$$M(t,l)^2\leq\left(\int_0^{t-l}\left[\int_0^u M(t,l+k)\mathbb{E}(Y)\rho(l+k)\nu(k)dk\right]^2 g(u,l)du\right) \tag{9}$$

as $\int_0^{t-l} g(u,l)du=1$. Thus, the first term, (7a), in the variance equation is non-negative.

The remaining terms can be dealt with as follows. (7a) is equal to zero, and the sum of (7c) is (using $Y(l+u,l)^2\geq Y(l+u,l)$) bounded below by $\int_0^{t-l}\mathbb{E}(Z(t,l+u)^2)\mathbb{E}(Y)\rho(l+u)\nu(u)(1-G(u,l))du$. Finally, noting that $Z(t,l+u)^2\geq Z(t,l+u)$, this is bounded below by $\int_0^{t-l} M(t,l+u)\mathbb{E}(Y)\rho(l+u)\nu(u)(1-G(u,l))du=M(t,l)$. Hence, $V(t,l)\geq M(t,l)$ holds.

To show strict overdispersion, note that for $V(t,l)=M(t,l)$ to hold, it is necessary that

$$\int_0^{t-l}\mathbb{E}(Z(t,l+u)^2)\mathbb{E}(Y)\rho(l+u)\nu(u)(1-G(u,l))du=\int_0^{t-l} M(t,l+u)\mathbb{E}(Y)\rho(l+u)\nu(u)(1-G(u,l))du \tag{10}$$

and hence, for each u (as $\mathbb{E}(Y) > 0$)

$$\mathbb{E}\Big[Z(t, l+u)(Z(t, l+u) - 1)\Big] = 0 \quad \text{or} \quad \rho(l+u)\nu(u)(1 - G(u, l)) = 0 \tag{11}$$

As proved in the Supplementary Material, if new infections can be caused, then more than one new infection can be caused. Thus, if an individual infected at $l+u$ has $\mathbb{E}\Big[Z(t, l+u)(Z(t, l+u)-1)\Big] = 0$, this individual cannot cause new infections whose infection trees have non-zero prevalence at time $l + u$. Hence, the condition 11 is equivalent to the epidemic being non-expanded at time t, as at each time $l+u$, either no infections are possible from the initial individual, or any individuals that are infected at time $l + u$ contribute zero prevalence at time t from the new infections they cause.

Hence, $Z(t, l)$ is strictly overdispersed for expanded epidemics. This means that Gaussian approximations are unlikely to be useful.

Variance midway through an epidemic

It is important to calculate uncertainty starting midway through an epidemic, conditional on previous events. This derivation is significantly more algebraically involved than the other work in this paper. For simplicity, we assume that $N(t, l)$ is an inhomogeneous Poisson Process, and that $L = \infty$ for each individual.

Suppose that prevalence (here equivalent to cumulative incidence) $Z(t, l) = n + 1$. We create a strictly increasing sequence $l = B_0 < B_1 < \cdots < B_n$ of $n+1$ infection times, which has probability density function

$$\underbrace{f_{\boldsymbol{B}}(\boldsymbol{b})}_{\text{Joint pdf}} = \underbrace{\frac{1}{P(Z(t, l) = n + 1)}}_{\text{normalising constant}} \underbrace{\prod_{i=1}^{n} \left(\rho(b_i) \sum_{j=0}^{i-1} \nu(b_i - b_j) \right)}_{\text{infection rates at each } b_i} \underbrace{\exp\left[- \sum_{i=0}^{n} \int_0^{t-b_i} \rho(s + l)\nu(s) ds \right]}_{\text{probability of no other infections}}, \tag{12}$$

where pdf is short for probability mass function. Then, the variance at time $t+s$ is given by

$$\underbrace{\mathrm{var}(Z(t+s,l))}_{\text{variance}} = \underbrace{\int_{b=0}^{t}\sum_{i=0}^{n} V^*(t+s,b)f_{B_i}(b)db}_{\text{variance from subsequent cases}} ...$$

$$... + \underbrace{\int_{b=0}^{t}\int_{c=0}^{t}\sum_{i=0}^{n}\sum_{j=0}^{n} M^*(t+s,b)M^*(t+s,c)(f_{B_i,B_j}(b,c) - f_{B_i}(b)f_{B_j}(c))dbdc}_{\text{variance from unknown infection times}}, \tag{13}$$

where $M^*(t+s,b)$ and $V^*(t+s,b)$ are the mean and variance of the size of the infection tree (i.e. prevalence or cumulative incidence) at time $t+s$, caused by an individual infected at time b, ignoring all individuals they infected before time t. These quantities are calculated from M and V (supplementary information). Note also that f_{B_i} and f_{B_i,B_j} are the one-and-two-dimensional marginal distributions from $f_{\boldsymbol{B}}$.

References

[1] S. Abbott et al. *EpiNow2: Estimate Real-Time Case Counts and Time-Varying Epidemiological Parameters*. 2020. DOI: 10.5281/zenodo.3957489.

[2] L. J. S. Allen. "A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis". en. In: *Infect Dis Model* 2.2 (May 2017), pp. 128–142.

[3] P. W. Anderson. "More is different". en. In: *Science* 177.4047 (Aug. 1972), pp. 393–396.

[4] D. Applebaum. *Lévy Processes and Stochastic Calculus*. en. Cambridge University Press, Apr. 2009.

[5] A. Barbour and G. Reinert. "Approximating the epidemic curve". en. In: *ejp* 18.none (Jan. 2013), pp. 1–30.

[6] N. F. Brazeau et al. "Estimating the COVID-19 infection fatality ratio accounting for seroreversion using statistical modelling". en. In: *Commun. Med.* 2 (May 2022), p. 54.

[7] D. Champredon et al. "Two approaches to forecast Ebola synthetic epidemics". en. In: *Epidemics* 22 (Mar. 2018), pp. 36–42.

[8] P. Cirillo and N. N. Taleb. "Tail risk of contagious diseases". en. In: *Nat. Phys.* 16.6 (May 2020), pp. 606–613.

[9] A. Cori et al. “A new framework and software to estimate time-varying reproduction numbers during epidemics”. In: *American Journal of Epidemiology* (2013). DOI: 10.1093/aje/kwt133.

[10] A. Cori et al. “A new framework and software to estimate time-varying reproduction numbers during epidemics”. en. In: *Am. J. Epidemiol.* 178.9 (Nov. 2013), pp. 1505–1512.

[11] N. R. Faria et al. “Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil”. en. In: *Science* (Apr. 2021).

[12] N. Ferguson et al. *Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand*. Tech. rep. Mar. 2020.

[13] N. M. Ferguson et al. “Strategies for containing an emerging influenza pandemic in Southeast Asia”. en. In: *Nature* 437.7056 (Sept. 2005), pp. 209–214.

[14] S. Flaxman et al. “Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe”. en. In: *Nature* (June 2020).

[15] C. Fraser. “Estimating individual and household reproduction numbers in an emerging epidemic”. In: *PLoS ONE* (2007). DOI: 10.1371/journal.pone.0000758.

[16] T. E. Harris et al. *The theory of branching processes*. Vol. 6. Springer Berlin, 1963.

[17] L. S. Hung. “The SARS epidemic in Hong Kong: what lessons have we learned?” en. In: *J. R. Soc. Med.* 96.8 (Aug. 2003), pp. 374–378.

[18] W. O. Kermack and A. G. McKendrick. “A contribution to the mathematical theory of epidemics”. In: *Proc. R. Soc. Lond. A Math. Phys. Sci.* 115.772 (Aug. 1927), pp. 700–721.

[19] A. D. Kiureghian and O. Ditlevsen. “Aleatory or epistemic? Does it matter?” In: *Struct. Saf.* 31.2 (Mar. 2009), pp. 105–112.

[20] A. J. Kucharski et al. “Early dynamics of transmission and control of {COVID}-19: a mathematical modelling study”. In: *Lancet Infect Dis* 3099.20 (2020), p. 2020.01.31.20019901. DOI: 10.1101/2020.01.31.20019901. URL: http://medrxiv.org/content/early/2020/02/18/2020.01.31.20019901.abstract.

[21] M. Lipsitch et al. “Transmission dynamics and control of severe acute respiratory syndrome”. en. In: *Science* 300.5627 (June 2003), pp. 1966–1970.

[22] Y. Liu et al. "The reproductive number of COVID-19 is higher compared to SARS coronavirus". en. In: *J. Travel Med.* 27.2 (Mar. 2020).

[23] J. O. Lloyd-Smith et al. "Superspreading and the effect of individual variation on disease emergence". en. In: *Nature* 438.7066 (Nov. 2005), pp. 355–359.

[24] J. N. Lyness. "Numerical algorithms based on the theory of complex variable". In: *Proceedings of the 1967 22nd national conference*. ACM '67. Association for Computing Machinery, Jan. 1967, pp. 125–133.

[25] S. Mishra et al. "On the derivation of the renewal equation from an age-dependent branching process: an epidemic modelling perspective". In: (June 2020). arXiv: 2006.16487 [q-bio.PE].

[26] Mumford. "The dawning of the age of stochasticity". In: *Mathematics: frontiers and perspectives* (2000).

[27] Y. Ogata. "On Lewis' simulation method for point processes". In: *IEEE Trans. Inf. Theory* 27.1 (Jan. 1981), pp. 23–31.

[28] M. S. Pakkanen et al. "Unifying incidence and prevalence under a time-varying general branching process". In: (July 2021). arXiv: 2107.05579 [q-bio.PE].

[29] K. V. Parag and C. A. Donnelly. "Using information theory to optimise epidemic models for real-time prediction and estimation". en. In: *PLoS Comput. Biol.* 16.7 (July 2020), e1007990.

[30] O. Pybus, A. Rambaut, COG-UK-Consortium, et al. "Preliminary analysis of SARS-CoV-2 importation & establishment of UK transmission lineages". In: *Virological. org* (2020).

[31] M. Sharma et al. "Understanding the effectiveness of government interventions against the resurgence of COVID-19 in Europe". en. In: *Nat. Commun.* 12.1 (Oct. 2021), p. 5820.

[32] R. Verity et al. "Estimates of the severity of {COVID}-19 disease". In: *Lancet Infect Dis* in press (2020). DOI: https://doi.org/10.1101/2020.03.09.20033357.

[33] L. Willem et al. "Lessons from a decade of individual-based models for infectious disease transmission: a systematic review (2006-2015)". en. In: *BMC Infect. Dis.* 17.1 (Sept. 2017), p. 612.

[34] F. Wong and J. J. Collins. *Evidence that coronavirus superspreading is fat-tailed*. 2020.

[35] M. E. Woolhouse et al. “Heterogeneities in the transmission of infectious agents: implications for the design of control programs”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 94.1 (Jan. 1997), pp. 338–342.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Pennetal2022Theuncertaintyofinfectiousdiseaseoutbreaksisunderestimatedsupplementarymaterial.pdf
- Pennetal2022Theuncertaintyofinfectiousdiseaseoutbreaksisunderestimatedsupplementarymaterial.pdf