PLOS ONE

RESEARCH ARTICLE

Assessing the heterogeneity in the transmission of infectious diseases from time series of epidemiological data

Günter Schneckenreither[1,2,5☯]*, **Lukas Herrmann**[3☯], **Rafael Reisenhofer**[4☯], **Niki Popper**[1,2,5], **Philipp Grohs**[3,4,6]

1 Institute of Information Systems Engineering, TU Wien, Vienna, Austria, **2** dwh GmbH, Vienna, Austria, **3** Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Linz, Austria, **4** Faculty of Mathematics, University of Vienna, Vienna, Austria, **5** Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria, **6** Research Network Data Science, University of Vienna, Vienna, Austria

☯ These authors contributed equally to this work.
* guenter.schneckenreither@tuwien.ac.at

Citation: Schneckenreither G, Herrmann L, Reisenhofer R, Popper N, Grohs P (2023) Assessing the heterogeneity in the transmission of infectious diseases from time series of epidemiological data. PLoS ONE 18(5): e0286012. https://doi.org/10.1371/journal.pone.0286012

Editor: Pablo Martin Rodriguez, Federal University of Pernambuco: Universidade Federal de Pernambuco, BRAZIL

Received: October 6, 2022

Accepted: May 5, 2023

Published: May 30, 2023

Data Availability Statement: The primary data used in this study is a time series of reported case numbers (daily positive COVID-19 tests) that are published by authorities in their respective countries and compiled in a single data repository by the Center for Systems Science and Engineering at Johns Hopkins University available at https://github.com/CSSEGISandData/COVID-19. Parameters of statistical distributions of disease intervals (published in literature or generated for this work) are collected in the supplementary

Abstract

Structural features and the heterogeneity of disease transmissions play an essential role in the dynamics of epidemic spread. But these aspects can not completely be assessed from aggregate data or macroscopic indicators such as the effective reproduction number. We propose in this paper an index of effective aggregate dispersion (EffDI) that indicates the significance of infection clusters and superspreading events in the progression of outbreaks by carefully measuring the level of relative stochasticity in time series of reported case numbers using a specially crafted statistical model for reproduction. This allows to detect potential transitions from predominantly clustered spreading to a diffusive regime with diminishing significance of singular clusters, which can be a decisive turning point in the progression of outbreaks and relevant in the planning of containment measures. We evaluate EffDI for SARS-CoV-2 case data in different countries and compare the results with a quantifier for the socio-demographic heterogeneity in disease transmissions in a case study to substantiate that EffDI qualifies as a measure for the heterogeneity in transmission dynamics.

Introduction

Tipping points that define significant transitions in the infection dynamics or ramifications of an epidemic have been a main focus of attention for researchers, decision makers, media outlets and the general public during the ongoing COVID-19 pandemic. In particular decision makers often cite thresholds for indicators such as the number of intensive care patients, the seven-day case rate, the hospitalization rate, or the effective reproduction number either as target values or triggers for specific containment measures. Most time-varying indicators that are commonly used to monitor an ongoing outbreak only measure the immediate effects of the outbreak on the population or, in the case of the effective reproduction number, aggregate properties of the current infection dynamics from a macroscopic point of view. However,

material (S1 Data) with references where applicable. The authors further used Austrian daily reported case numbers per gender, federal state, and five-year age-group, which is based on data that is administrated by the Austrian Agency for Health and Food Safety (AGES) and which was processed for our team in terms of a research cooperation (https://datenplattform-covid.goeg.at/prognosen) with the Austrian National Public Health Institute (GÖG). The data the authors received is subject to a non-disclosure agreement. Similar data is, however, published in the official Austrian COVID-19 dashboard (https://covid19-dashboard.ages.at). Under the same terms, the authors were also provided with a dataset on the time-span between recorded positive test results and corresponding symptom onset. This data contains the daily histogram of 'reporting delays' measured in days for reported cases in Austria. The results of statistical evaluations are provided in the supplementary material (S3 Data and S7 Text). Austrian demographic data was obtained from the national statistics bureau Statistics Austria (https://www.statistik.at). Data that was used in the study does not contain personal information and is fully anonymized, it specifically does not contain information on treatment, hospitalization, or fatalities.

Funding: R.R. gratefully acknowledges support from the Austrian Science Fund (www.fwf.ac.at) (FWF M 2528). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

outbreaks that are strongly governed by large infection clusters and superspreading events (SSEs) can only be fully understood by taking into account events occurring at a meso- or even microscopic scale of the epidemic. Such outbreaks are predominantly investigated with individual- or network-based models [1–9] and usually exhibit a large dispersion in the number of individual secondary infections, which has been observed to crucially affect the respective infection dynamics [1–4, 10–12].

Specifically during low-prevalence periods of an epidemic, such as in the onset of an outbreak, a strong variation in the number of individual secondary cases implies a higher degree of stochasticity in the observed daily case numbers and increases the likelihood of stagnation as compared to a situation in which the offspring distribution exhibits only a small variance [3, 10]. Once an outbreak with large dispersion in the individual secondary case numbers has taken off and grown beyond the emergence of isolated SSEs, it usually exhibits stable exponential growth with growth rates that are comparable to an outbreak with the same basic reproduction number but no variation in the number of secondary infections per infected individual [3, 10]. This suggests a phase transition occurring in the early stages of an outbreak with large dispersion in the number of individual secondary infections after which containment becomes increasingly difficult due to the ill nature of exponential growth.

Analogously, as seen during the current COVID-19 pandemic, we observe that measures such as rigorous contact tracing and personal quarantining are successful in low-prevalence episodes and particularly when isolated infection clusters or spreading trees can be identified and significantly contribute to the bulk of new infections. During periods of high prevalence or when infection clusters can no longer be demarcated, on the other hand, broader and more severe interventions such as partial lockdown and curfews are deemed to be effective for curbing the reproduction dynamics. Hence, correctly identifying tipping points in the course of an ongoing outbreak that indicate a transition between phases of clustered and diffusive spread independently of the general level of prevalent cases could be of great value for managing the implementation of countermeasures.

In this work, we propose a novel indicator that makes this phase transition transparent by quantifying the *effective aggregate dispersion* of epidemic outbreaks based on time series of daily reported case numbers and a statistical model for reproduction. Technically, our indicator can be seen as a time-varying measure for the stochasticity in time series of aggregate daily case numbers. It is important to note that simply computing standard measures of variation, such as the empirical variance, does not yield a meaningful metric in this setting. This is mainly because much of the variation is not due to the dispersion of secondary infections but can be explained by other factors like changes in the effective reproduction number or weekly periodic patterns that are caused by seasonalities in the behavior of the population and the underlying testing and reporting regime. Our statistical model and inference framework take into account artifacts and patterns that cannot be attributed to the dispersion in the number of secondary infections and provides an *effective aggregate dispersion index* (EffDI) that reflects the heterogeneity in individual transmission dynamics. We anticipate the EffDI to act as an 'early warning system' that allows decision makers to react to subtle but significant changes in the infection dynamics of an outbreak during periods of low-prevalence.

Investigating the infection dynamics on the individual or mesoscopic level requires a model for reproduction that considers the temporal aspects of disease transmission and case registration that manifest in aggregate reported case numbers. We initially revise the structure of existing models for inferring effective reproduction factors [7, 13–20] and basic dispersion parameters [1–4, 10–13] and generalize reproduction as the interplay between *infectious load*, which corresponds to the current contagious population that can be derived from reported case numbers, and *infectious activity*, which is the corresponding amount of new infections.

We then analyze the characteristics of the stochasticity in time series of reported case numbers and artifacts that result from the administration and procedures of case registration. Based on our findings, we develop a carefully designed infection model for the quantification of time-varying effective aggregate dispersion in daily incidence time series. This model relates the infectious load observed at a given day with the respective infectious activity and has certain desirable properties that enable the quantification of time-varying effective aggregate dispersion. Most importantly, it is capable of capturing daily seasonal patterns in the reproduction dynamics and considers an additional time-varying aggregate dispersion parameter. Based on this model, the EffDI will be determined by the time-varying minimal amount of aggregate dispersion required such that the observed infectious load and activity are plausible under the fitted model parameters.

With our approach, we exploit the connection between the observed stochasticity in reported case numbers and the degree to which SSEs and infection clusters are statistically relevant and traceable in the current infection dynamics. To support the developed indicator, we further investigate the occurrence of SSEs from a socio-demographic perspective based on reported case data that includes information about a small number of social attributes of the incidence population. We measure the statistical distance between the socio-demographic composition of infectious load and infectious activity, and assume that this supporting indicator reflects the emergence of infection clusters that are significant relative to the overall infectious activity. From the qualitative correspondence of both indicators, we conclude that EffDI can indeed be used to assess the significance of SSEs and infection clusters based on a time series of aggregate reported case numbers.

Results

Inference of reproduction dynamics

The common approach for investigating the reproduction dynamics of epidemic outbreaks aligns with a mathematical model that can be formalized as

$$\text{reproduction} \;:\; \text{infectious load} \mapsto \text{infectious activity}, \tag{1}$$

where *infectious load* refers to the currently contagious population or the induced potential for generating new infections among the susceptible population, and *infectious activity* refers to the amount of infections arising from transmissions by the currently infected population (compare also *prevalence* and *incidence* in the widest sense).

For inferring information about reproduction, load and activity are regarded as 'exogenous' variables that are derived from reported case numbers. This transformation must take into account the characteristics of the disease and of case reporting. An important simplification is to assume that load and activity are equally affected by under-reporting and by gradual changes of the detection rate. This becomes evident for a simple multiplicative reproduction factor in particular. Nevertheless, registration of cases is also subject to specific delays which result from the deferred exposure of symptoms, personal testing and monitoring schemes, and from the implemented administrative processes (compare 'nowcasting') [14, 21–24]. Furthermore, to distinguish between load and activity, the time between subsequent infections in a transmission pair must be considered [16, 17, 25–29]. This time period is specific to the disease and the contact behavior of the population and can only be measured in individually tracked infection pairs [30–37]. Fig 1A presents a schematic diagram of the time intervals occurring in transmission pairs. Fig 1B visualizes the general approach for inferring reproduction via load and activity from reported case numbers.

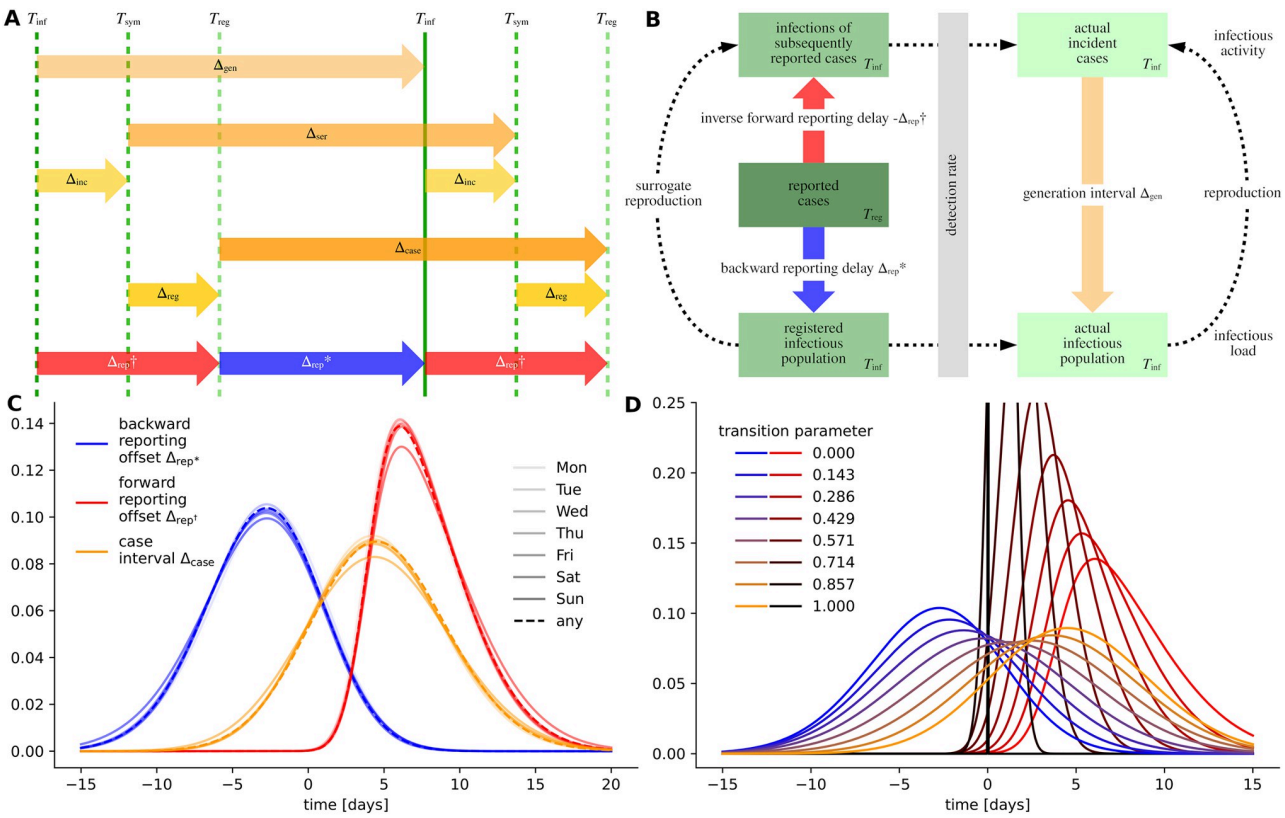

Fig 1. Disease intervals and reproduction. (A) Disease and transmission intervals. Schematic diagram showing the events *infection*, *symptom onset* and *registration* (denoted by T) in the timeline of a hypothetical infection pair. The time period between two consecutive disease transmissions is called the *generation interval*, here written as Δ_{gen}. The time period between the symptom onsets in an infection pair is called the *serial interval* Δ_{ser}. We further denote the time period between registration of two consecutive cases as the *case interval* Δ_{case}. Further intervals shown are the *incubation delay* Δ_{inc}, the *registration delay* Δ_{reg} and the *forward and backward reporting offsets* $\Delta_{\text{rep}^\dagger}$, Δ_{rep^*}. (B) Visual outline for the quantification of reproductive dynamics based on time series of reported cases. Infectious load and activity can be derived from reported cases using the statistical distributions of the reporting offsets. The obtained characterization of epidemic progression is only a surrogate for actual reproduction dynamics. (C) Probability densities of the backward and forward reporting offset distributions and the case interval distribution inferred from data. All interval distributions were calculated for different parameter settings; here, the weekday of case registration is encoded in the lightness of the color. (D) Transformation between different statistical models for the reporting offset intervals. We investigate our model and the resulting dispersion indicator (EffDI) under gradual transformation of the probability densities to analyze potential impacts of their characteristic features (S9 Text). Here, the continuous transformation—implemented by means of a transition parameter in the interval [0, 1] and displayed according to continuous color ramps—of the forward and backward reporting offset into a degenerate distribution and the case interval is portrayed.

https://doi.org/10.1371/journal.pone.0286012.g001

In the following, we investigate the statistical distributions of typical SARS-CoV-2 disease intervals and formalize the calculation of load and activity. We then analyze the stochasticity and artifacts in time series of reported SARS-CoV-2 case numbers which is the foundation for constructing a statistical framework for the quantification of heterogeneity in reproduction in the form of an aggregate dispersion parameter.

Obtaining infectious load and activity

We denote the time interval between infection and (potential) registration in a surveillance system as the *forward reporting offset* $\Delta_{\text{rep}^\dagger}$, and the time interval between the registration of a case and a (potential) secondary infection as the *backward reporting offset* Δ_{rep^*}. The *case interval* $\Delta_{\text{case}} = \Delta_{\text{rep}^*} + \Delta_{\text{rep}^\dagger}$ is the time between the registration of consecutive cases (compare

Fig 1A). For simplification and because case numbers are usually reported on a daily basis, we restrict our formulations to discrete time and assume that continuous distributions can be discretized accordingly [13, 20]. Independent of the actual statistical distributions, load and activity can be interpreted as simulated individual contagion events or as the corresponding aggregate time series,

$$\begin{aligned} T^{*}_{\text{inf}} &= T_{\text{reg}} + \Delta_{\text{rep}^{*}} \iff & I^{*}_{t} &= \sum_{\tau\in\mathbb{Z}} I_{t-\tau} u_{\tau} \\ T^{\dagger}_{\text{inf}} &= T_{\text{reg}} - \Delta_{\text{rep}^{\dagger}} \iff & I^{\dagger}_{t} &= \sum_{\tau\in\mathbb{Z}} I_{t-\tau} v_{-\tau} \end{aligned} \tag{2}$$

where I_t is the 'measurable' time series of reported cases, which, in turn, reflects all individual events of case registration (T_{reg}), and u_τ and v_τ are the probability masses of the backward and forward reporting offset distributions ($\Delta_{\text{rep}^{*}}$ and $\Delta_{\text{rep}^{\dagger}}$). Accordingly, the time series of infectious load I^{*}_{t} can be understood to reflect the (potential) events T^{*}_{inf} of individual persons transmitting the disease and the time series of infectious activity $I^{\dagger}_{t}$ counts the events of (detected) persons getting infected $T^{\dagger}_{\text{inf}}$ (compare Fig 1B). It is important to note, however, that infectious load and activity are time series of positive real numbers and that they serve as an abstract model for the occurrence of individual disease transmission events.

To obtain tangible statistical models for $\Delta_{\text{rep}^{*}}$ and $\Delta_{\text{rep}^{\dagger}}$ (i.e. the probability masses u_τ and v_τ), we formulate an algebraic equation system that relates the time intervals occurring in individual transmission pairs (Fig 1A). We then use available data about the statistical distributions of (individual) disease intervals to 'solve' this equation system using a Markov chain Monte Carlo (MCMC) approach (see Methods). To account for the most significant temporal changes and seasonalities in reporting, we separately perform our calculations for different stratifications of the data differentiating by the day of the week and for distinct time periods of the epidemic. We observe a gradual shortening of the obtained reporting offset intervals over the course of the pandemic (see S5 Text) and also marginally longer intervals for cases reported on weekends (see S5 Text and Fig 1C). The obtained statistical models are particularly valid for the Austrian setting with a relative large number of routine tests, but could indicate a general configuration that is also applicable to other countries. Similar approaches based on MCMC methods or Bayesian inference were applied for synthesising missing statistical information about disease intervals before [14, 20, 21, 27, 36, 38, 39]. Ultimately, the obtained probability masses can be used in Eq (2) to calculate the time series of load and activity.

In practice, frameworks that deal with the quantification of effective reproduction factors often use—in contrast to Eq (2)—a simplified model of disease and transmission intervals to calculate what we call load and activity. A reason for this is that particularly during the early stage of a pandemic detailed knowledge about the time intervals, and especially the reporting dynamics, is not available [13, 19, 20]. Often—partially justified by the characteristics of the observed disease—estimates of the serial interval provided by secondary literature and studies are used to approximate the generation or case interval [3, 7, 13, 19]. Methods for extracting the relevant inter-patient time distribution from a number of observed infection pairs were also directly included in the algorithms for calculating effective reproduction factors [20]. Furthermore, in the inference of effective reproduction factors, reported cases are widely used as a surrogate for actual contagion events [7, 13, 16, 17, 19, 20, 40, 41] (infectious activity). The corresponding 'simplified' configuration of the interval model and the associated forms of

infectious load and activity corresponds to the formula

$$\begin{aligned} \Delta_{\mathrm{rep}^*} &= \Delta_{\mathrm{case}} \iff T^*_{\mathrm{inf}} = T_{\mathrm{reg}} + \Delta_{\mathrm{case}} \iff I^*_t = \sum_{\tau\in\mathbb{Z}} I_{t-\tau} w_\tau \\ \Delta_{\mathrm{rep}^\dagger} &= 0 \iff T^\dagger_{\mathrm{inf}} = T_{\mathrm{reg}} \iff I^\dagger_t = I_t \end{aligned} \tag{3}$$

where w_τ are the probability masses of the distribution of the case interval Δ_{case}, which is usually replaced with either Δ_{ser} or Δ_{gen}. Under the condition that the case interval—or the used surrogate—is strictly positive, it can further be argued, that in contrast to 'complex' data-driven models (as discussed above and corresponding to Eq (2)), this form is better suited for the real-time assessment and forecasting of epidemics as no future values in the time series of the reported cases is required [13, 16, 17, 20].

Stochasticity and artifacts in reported case numbers

Reported (SARS-CoV-2) case data is typically characterized by weekly seasonal patterns [11, 15, 18, 21, 23], artifacts that result from reporting procedures, and a distinct—potentially temporally varying—level of fluctuation (variance). Nevertheless, statistical models for inferring effective reproduction factors usually directly operate on crude time series of reported case numbers according to the model in Eq (3). One way to account for the at times large variance in the data is to increase the dispersion of the statistical model for the individual number of secondary infections (offspring distribution). In other words, when the data is over-dispersed with respect to the employed statistical model, a model with greater variance in secondary infections must be used. Alongside improved correspondence with the data [3, 41, 42], such a modification can also provide—eventually in the form of a *basic dispersion parameter*—information about the inherent level of dispersion in the secondary infections during an outbreak [1, 4, 10, 12]. This, in turn, is a leverage point for the assessment of spreading characteristics such as the occurrence of SSEs [1–3, 10]. However, the stochasticity and regularity in time series of reported case numbers is not only affected by the transmission dynamics of the epidemic. It is also the employed testing and reporting regime that causes weekly seasonal patterns, making it necessary to separate artifacts of reporting from the stochasticity that is presumably caused by the transmission dynamics itself.

Furthermore, we observe that the general level of stochasticity in reported case numbers is high in low-incidence phases but diminishes when case numbers become large, rendering the data more regular with pronounced seasonal patterns in these periods. Fig 2 illustrates this separation for three different countries by comparing normalized segments of reported case numbers from low-prevalence periods with segments from high-prevalence periods.

In the following, we develop a statistical model for the reproduction dynamics of epidemic outbreaks that considers a temporally varying degree of dispersion and (a) explicitly decouples the amount of stochasticity from the general level of infectious load and (b) 'filters' seasonal patterns from the observed infectious activity. Ultimately, this leads to what we call a time-varying *aggregate dispersion parameter*, which, in turn, indicates periods of an epidemic that are likely driven by SSEs and episodes with diffusive spread. We implement both aspects in our extension of the state-of-the-art statistical inference framework in the next sections.

A model for quantifying effective aggregate dispersion

A common statistical approach for the inference of reproduction dynamics is based on a Poisson model for the individual number of secondary infections (offspring distribution) that is parameterized with an effective reproduction factor as the rate or expected value. Under the assumption that the individual reproduction factor (i.e. the expected number of secondary

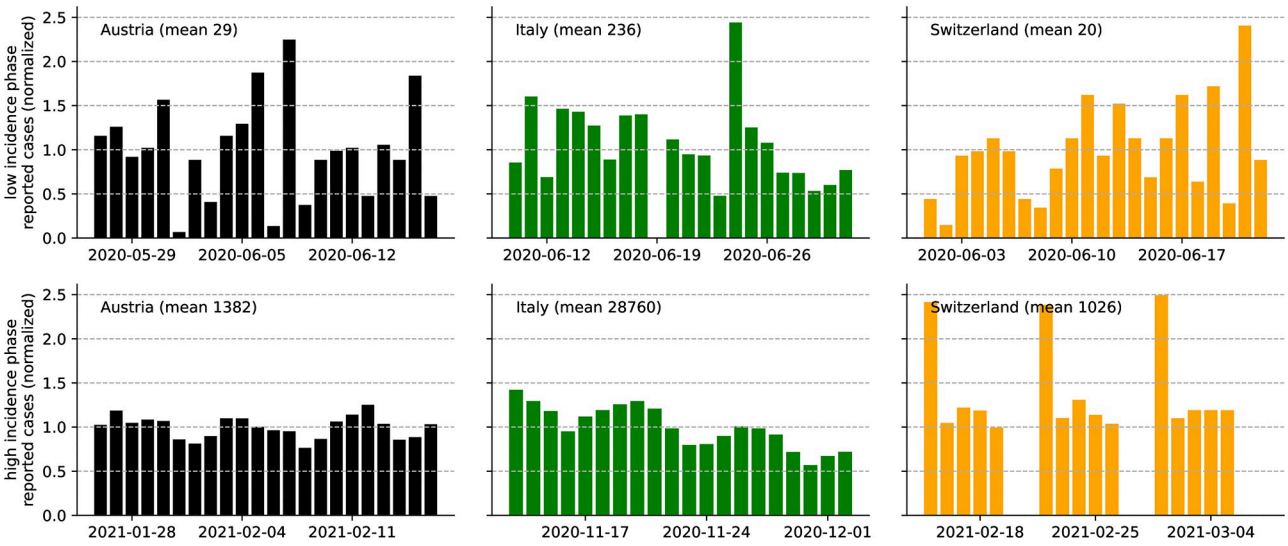

Fig 2. Varying regularity of reported case numbers. Normalized segments of daily reported case numbers [43] during low-prevalence (top row) and high prevalence (bottom row) periods in different countries (columns). Normalization was achieved by dividing the daily case numbers by their mean value during the respective three-week period. During the respective high prevalence periods, we observe a high degree of regularity and weekly seasonal patterns, whereas in low-prevalence phases, the segments seem to be more erratic.

https://doi.org/10.1371/journal.pone.0286012.g002

cases) is the same for all contagious individuals at a given time t and by using the additivity of the Poisson distribution, the total number of new cases $I_t^\dagger$ (infectious activity) can be modeled as

$$I_t^\dagger \sim \text{Poisson}(R_t I_t^*), \qquad P(I_t^\dagger = x) = \frac{(R_t I_t^*)^x e^{-R_t I_t^*}}{x!}, \tag{4}$$

where I_t^* is the number of contagious individuals (infectious load), and R_t denotes the common effective reproduction factor [13, 19, 20].

Not limited to the current SARS-CoV-2 pandemic, a large variety of techniques and software packages have been developed to infer time series of aggregate effective reproduction factors R_t based on Eq (4). A particular extension of this approach is to regard the individual reproduction factor as a random variable instead of a deterministic uniform value. Typical models for a stochastic individual reproduction factor are for instance the exponential distribution, or the gamma distribution. Due to the stochasticity encountered in aggregate reported case numbers, using an exponential model can lead to over-dispersion when fitting the resulting geometric model to the data [10, 18]. Considering identically and independently gamma distributed individual reproduction factors with shape parameter k and scale parameter R_t/k leads to a negative binomial model (gamma-Poisson mixture) for the total number of new cases, which can be written as

$$I_t^\dagger \sim \text{NB}\left(kI_t^*, \frac{R_t}{R_t+k}\right), \quad P(I_t^\dagger = x) = \binom{x+kI_t^*-1}{kI_t^*-1}\left(\frac{k}{R_t+k}\right)^{kI_t^*}\left(\frac{R_t}{R_t+k}\right)^x, \tag{5}$$

where the first parameter is understood as the number of allowed failures and the second parameter is the success probability. This approach allows the model to better comply with the observed stochasticity in reported case numbers [1, 3, 10, 18, 21, 41, 42] by considering an additional free parameter k, which can be viewed as a dispersion parameter that reflects the

variance in individual secondary cases. Large dispersion, which is associated with a scenario that shows great variability in the individual numbers of secondary cases, can be modeled in Eq (5) by choosing k small. Vice versa, small dispersion, which describes a situation where all infected individuals cause a similar number of secondary cases and where spread of the disease happens in a more 'homogeneous fashion', can be modeled by choosing k large. For $k \to \infty$, the limiting distribution falls back to the Poisson model Eq (4). The properties of the negative binomial distribution proved to provide good overall correspondence with the data in many case studies. Analogous to the basic reproduction number, *basic dispersion parameters* of SARS-CoV-2 outbreaks and other epidemics have been estimated from historical data and used for investigating the heterogeneity of infection dynamics [1, 2, 4, 10, 12].

According to our generalized concept of reproduction (Eq (1)), we presume that infectious load and infectious activity are continuous variables that are calculated as weighted sums (Eqs (2) and (3)) of original counting data. In S8 Text we demonstrate that the standard discrete-valued negative-binomial model in Eq (5) can be approximated with a corresponding continuous gamma model that is compatible with our approach to infectious load and activity. In the following, we discuss a further reproduction model based on the gamma distribution that was designed for investigating variable dispersion.

The observed variability of relative stochasticity in recorded time series data (Fig 2) and formal considerations regarding the model defined in Eq (5) indicate that the effect of dispersion on the infection dynamics is negligible in high-incidence periods. In order to capture the variable impact of dispersion on the dynamics of an outbreak with a quantitative measure, we consider a carefully designed infection model which introduces a time-varying *aggregate* dispersion parameter κ_t,

$$I_t^\dagger \sim \text{Gamma}\left(\kappa_t, \frac{I_t^* R_t}{\kappa_t}\right), \qquad f(x) = \Gamma(\kappa_t)^{-1} \left(\frac{\kappa_t}{I_t^* R_t}\right)^{\kappa_t} x^{\kappa_t - 1}\, e^{x \frac{\kappa_t}{I_t^* R_t}}. \tag{6}$$

A thorough motivation of Eq (6) and a description of the estimation procedure can be found in *Methods*. The key statistical aspects of the models Eqs (5) and (6) and additional aspects are recapitulated in S8 Text. Essentially, the models defined in Eqs (5) and (6) have the same expected value $I_t^* R_t$, but the variance of the former scales linearly with I_t^*, whereas the variance of the latter scales quadratically with I_t^*. Hence, given that the true infection dynamics of an outbreak yield an infectious activity $I_t^\dagger$ for which the variance scales subquadratically with the magnitude of the infectious load I_t^*—as observed in the data (c.f. Fig 2)—we can choose the parameter κ_t in the model Eq (6) increasingly large during high-incidence regimes without impeding the ability of the model to describe the variance present in the data. Ultimately, for a given day t, we apply Monte Carlo simulation to estimate the smallest plausible amount of aggregate dispersion (i.e., the largest possible value for κ_t) that is required for explaining the observed time series of daily aggregate case numbers.

Filtering of seasonal patterns

Inferring quantifiers such as effective reproduction factors R_t (using a statistical model based on Eq (4)) directly from reported case numbers (according to the load-activity model in Eq (3)) leads to strong oscillations in the obtained quantifier. A possible approach for obtaining less erratic numerical indicators is to employ the load-activity model in Eq (2), which utilizes the 'informed' offset distributions as convolution kernels [14, 16, 18] for calculating load and activity. However, with this approach, irregularities resulting from the transmission dynamics are also mostly removed from the exogenous variables (load and activity) and quantifying the stochasticity in transmission dynamics is no longer possible. Hence, we anticipate that the

'raw' form of load and activity according to Eq (3) is more suitable for harnessing the stochasticity in reported case data [43] than a 'regularizing' model Eq 2).

Many existing frameworks for estimating numerical indicators such as the effective reproduction factor use raw and unprocessed case numbers directly (Eq (3)). This allows to use the stochasticity in the data, for instance, to estimate the confidence intervals of the obtained indicator. To obtain smoother behavior of numerical indicators (e.g. R_t) often constant reproduction dynamics are assumed during a certain time window by simultaneously regarding all load-activity pairs that statistically occur during that window in an extended (Bayesian) inference approach [13]. Similarly, to remove the effects of reporting artifacts and seasonal patterns, we use a windowed linear model for R_t in Eq (6) that can explain trends and weekly periodic patterns in the data (see Methods) but retains the residual stochasticity such that the assessment of dispersion is still possible.

In S9 Text, we further address the question of how much of the stochasticity in reported case data can actually be filtered out in the preprocessing of load and activity before the assessment of stochasticity fails. To this end, we investigate the quantifier developed in this paper under a continuous transition between the 'pre-regularizing' model in Eq (2) and the 'raw' scenario defined in Eq (3). Technically this is achieved by simultaneously transforming the densities of the forward and backward reporting offset distributions into a degenerate distribution and the case interval respectively (cf. Fig 1D). Our results confirm what has been observed in existing research [13, 16, 17, 20]. Regularization, on the one hand, leads to implicit delays in the time series of load and activity, which is problematic in the real-time assessment of epidemic progression. Secondly, over-smooth input time series can hinder the assessment of stochasticity. On the other hand, when reporting artifacts or excess roughness in the underlying data is not filtered, the confidence intervals in statistical inference approaches are generally larger and numerical quantifiers can behave erratically. Furthermore, we conclude that if reporting artifacts are sufficiently filtered, inclusion of detailed data-driven statistical distributions of disease and transmission specific intervals can be evaded [16, 18, 41] and it is reasonable to map only the basic characteristics of these intervals.

Construction and evaluation of EffDI

Our approach for quantifying time varying dispersion is based on the temporal progression of the plausibility $p_t(\kappa) \in [0, 1]$ of the reproduction model Eq (6) with fixed dispersion parameters κ given the observed data. When for an assumed amount of dispersion (κ), the model becomes less plausible over time, increasing the dispersion (i.e. proceeding to a lower value of κ) allows the model to retain the initial amount of plausibility. Accordingly, we construct the time varying dispersion parameter κ_t as the largest κ for which the model provides the same amount of plausibility p throughout the course of an outbreak,

$$\kappa_t(p) = \sup\{\kappa : p_t(\kappa) \geq p\}.$$

Technical details on the corresponding Monte Carlo simulation approach and on the used test statistic are provided in Methods (cf. Eq (20)).

Fig 3B depicts the progression of the plausibility of the model for a range of dispersion parameters κ when fitting the corresponding model Eq (6) on aggregate reported SARS-CoV-2 case numbers in Austria [43]. Using a logarithmic scale for κ in Fig 3B, the transition from zero plausibility ($p = 0.0$) to high plausibility ($p = 1.0$) appears abrupt and uniform, suggesting that under such a transformation a plausibility level set could be a meaningful indicator. Ultimately, we define the *effective aggregate dispersion index* (EffDI) as the reciprocal square root

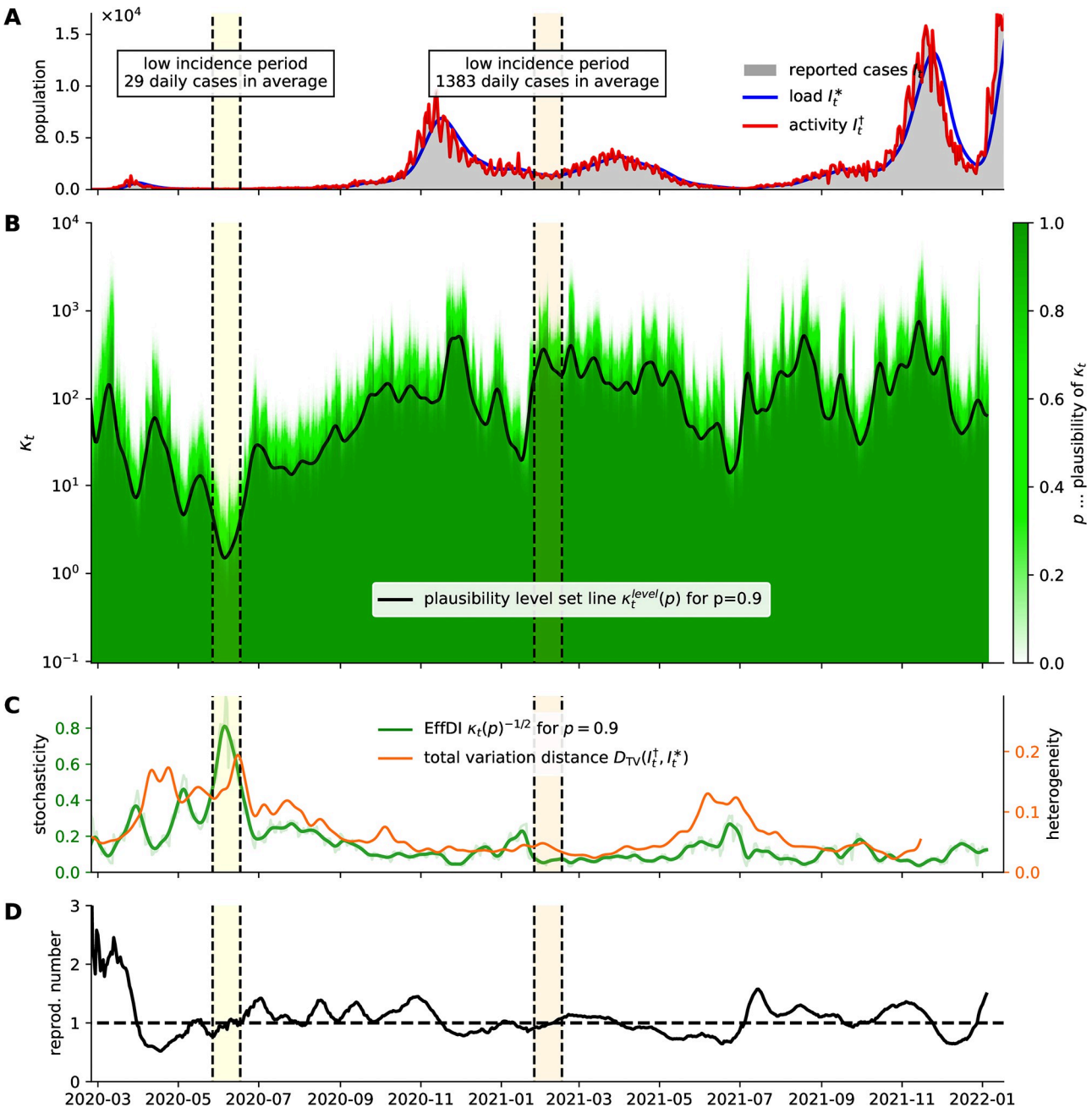

Fig 3. Quantification of effective aggregate dispersion. (A) Number of reported cases in Austria [43] and infectious load and activity according to Eq (3). The low- and a high-incidence phases from Fig 2 are indicated with vertical lines. (B) Plausibility diagram for the dispersion parameter κ_t. The transition between the plausible and implausible regime (e.g. $p = 0.9$) is abrupt. (C) EffDI results by the transformed level set line (green) $(\kappa_t(p = 0.9))^{-1/2}$. It is a quantifier for dispersion and corresponds to the coefficient of variation in the underlying statistical model Eq (6). The progression of dispersion aligns with the progression of a socio-demographic heterogeneity measure (orange). Socio-demographic heterogeneity is measured via the total variation distance between infectious load and infectious activity. (D) Resulting case reproduction number.

https://doi.org/10.1371/journal.pone.0286012.g003

of the time varying dispersion parameter

$$\mathrm{EffDI}_t = \kappa_t(p)^{-1/2}, \tag{7}$$

which has similar characteristics but also corresponds to the coefficient of variation of the model defined in Eq (6) (compare S8 Text).

Based on the construction of the time varying dispersion parameter, we anticipate similar qualitative behavior for arbitrary (initial) p. However, requiring high plausibility seems intuitive and we assume that the smoothed approximation of the level set $\kappa_t(p = 0.9)$ and the corresponding EffDI_t is a suitable indicator for separating plausible assumptions about dispersion from model configurations for which the data is over-dispersed.

In Fig 3A and 3B we observe that the dispersion parameter κ_t can in general be chosen large during high prevalence periods, and that low-prevalence periods of an outbreak are usually associated with a measurably higher degree of stochasticity relative to the current infectious load. For instance, we can observe a significant phase transition in the stochasticity of the time series of aggregate reported case numbers during the end of May and the first weeks of June in 2020.

In Fig 3C, we compare the EffDI with the progression of a measure for the socio-demographic heterogeneity in disease transmissions for the Austrian setting. The heterogeneity measure relies on data that includes additional information about cases and is developed in the next section. We motivate this measure as an independent indicator for the occurrence and significance of SSEs and infection clusters. As a consequence, qualitative correspondence between both measures substantiates that the progression of the time-varying aggregate dispersion parameter adequately reflects structural changes in the underlying infection dynamics of an outbreak.

Fig 4 shows the development of EffDI based on daily aggregate case numbers between February 25, 2020 and January 31, 2022 for six different countries. In the case of South Korea, the transition line closely follows the expected behavior of high stochasticity during low-prevalence periods and low stoachsticity during high-prevalence periods with a prominent phase transition during a particular period of low-prevalence around the beginning of May 2020. By far the longest period of high stochasticity can be found in the aggregate case numbers reported by Singapore, which yield a very high stochasticity measure for more than a year. In all of the five analyzed European countries (Austria, Switzerland, United Kingdom, Italy, and Germany), significant periods of high stochasticty can be observed between May and July 2020, indicating that during this time, the infection dynamics of the COVID-19 outbreak were mainly governed by comparatively few isolated events throughout many countries in Europe.

However, particularly in the cases of Germany and Italy, our analysis also yields periods of relatively high stochasticty during times of high prevalence. We assume that during these periods, systematic changes were made to the employed testing and reporting regime such that the measured stochasticity can not be exclusively attributed to the actual infection dynamics. In the example of Switzerland, all occurrences of high stochasticity in times of high prevalence can in fact be attributed to changes and irregularities in case reporting because aggregate case numbers were only reported for weekdays after 2020-09-19, and four additional dates in the spring of 2021 were missing from the time series of reported case numbers.

We further contextualize in S10 Text the behavior of EffDI with the emergence of new variants of the Corona virus in Austria. Generally, the gradual introduction of a new variant does not imply a change in the progression of our heterogeneity measure. However, we observe that aggressive variants with increased transmissibility (e.g. Omicron) can become dominant in a

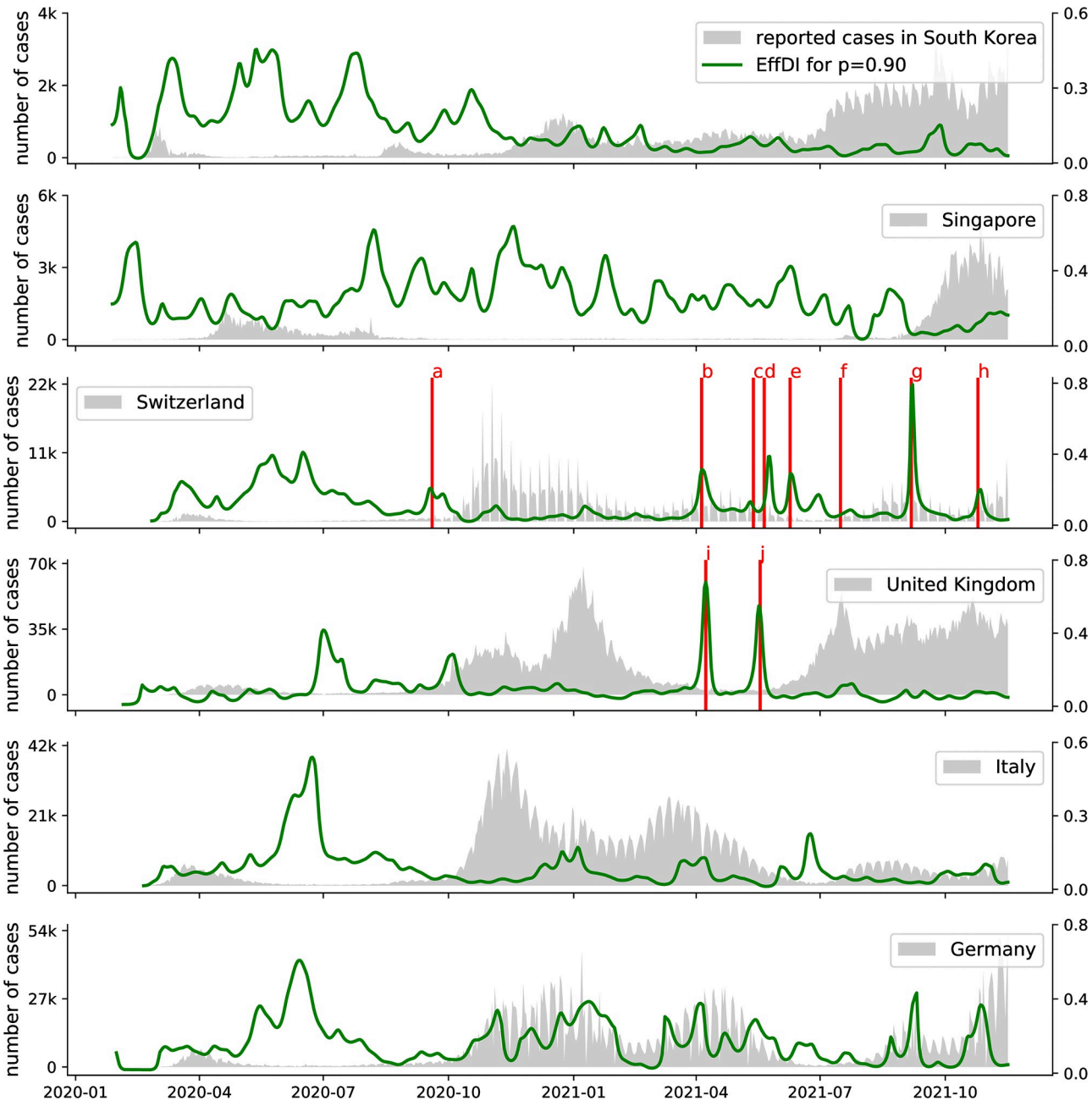

Fig 4. EffDI evaluated for different countries. Reported case numbers [43] are shown in gray. In the case of Switzerland and the UK, certain dates are highlighted to illustrate the effect of irregularities or changes in the reporting regime. a: After September 19, 2020, Switzerland only reported aggregate case numbers on weekdays (cf. Fig 2). b, c, d, e, f, h: Missing aggregate case numbers in Switzerland on several days in 2021. g: Missing entries on a Friday and the following Monday in Switzerland during September 2021. i, j: Reporting of negative aggregate case numbers on two days in April and May 2021 in the UK.

https://doi.org/10.1371/journal.pone.0286012.g004

relatively short time and abruptly change the reproduction dynamics coinciding with elevated values for EffDI.

Social heterogeneity in transmission dynamics

To support our findings about the temporal variability of aggregate dispersion, we investigate the progression of socio-demographic heterogeneity of the infected population and in reproduction using statistical distance measures. In particular, we calculate the *total variation distance* (TV)

$$D_{\mathrm{TV}}(p,q) \quad = \frac{1}{2}\sum_{x\in X}|p(x)-q(x)|, \tag{8}$$

between normalized histograms p and q of two populations with respect to a discrete and categorical feature space X. Here X either refers to age-compartments, gender, geographic region or the product space.

Let $I_t^\dagger$ be the time series of vector-valued infectious activity consisting of normalized histograms of the newly infected population. Let Q be the socio-demographic configuration of the total population. We denote the difference $D_{\mathrm{TV}}(I_t^\dagger, Q)$ as the *heterogeneity of infectious activity*. If this quantifier is small, we expect that infectious activity is distributed homogeneously across the population. If the distance measure is large, infectious activity is concentrated in distinct social compartments. We use the previously found statistical interval models to extract vector-valued infectious activity according to Eq (2) from an Austrian socio-demographic case dataset and quantify the distance to the total population. In Fig 5B–5D the resulting heterogeneity of infectious activity with respect to gender, age and administrative affiliation is visualized. In Fig 5E the corresponding evaluated distance measures are plotted over time. With the visual as well as the quantitative approach it is possible to distinguish phases with heterogeneous infectious activity and phases with homogeneous activity.

We further compare the socio-demographic configuration of infectious activity $I_t^\dagger$ with the configuration of infectious load I_t^* and denote the progression of the statistical distance $D_{\mathrm{TV}}(I_t^\dagger, I_t^*)$ as the socio-demographic *heterogeneity in reproduction*. When the distance measure is small, then the socio-demographic configuration of the infected population remains the same, when the distance measure is large, we expect that the socio-demographic configuration of the infected population is about to change. Hence, large heterogeneity can indicate that predominant infectious activity shifts to previously unaffected social strata. We further assume that abrupt changes in the socio-demographic composition of the infected population point to the emergence of infection clusters that are composed of individuals with similar social attributes (e.g. geographic location). Hence, we conclude that (analogously to EffDI) a measure for the social heterogeneity in reproduction can indicate the emergence and dissipation of significant infection clusters. In Fig 5F this measure is evaluated for Austrian case data.

In Fig 5 a selection of significant SSEs and infection clusters is indicated in the timeline (compare [1, 3]). Details about case numbers and the corresponding media coverage are found in the supplement (S7 Text). We observe an initial cluster (C1) in Ischgl, Tyrol (region code 7) affecting mostly younger and middle-aged persons. During the first epidemic wave, news media report about infection clusters in retirement homes in different provinces (C2). During the summer of 2020 we further observe mostly smaller infection clusters with occasional larger SSEs (C3) in logistic centers in and around Vienna (region code 9) as well as other SSEs (C4) in the state of Lower Austria (region code 3). A large regional cluster (C5) was detected in Upper Austria (region code 4). In the onset of the second epidemic wave SSEs are reported to have occurred during private meetings (C6) as well as in public events (C7). After a high-

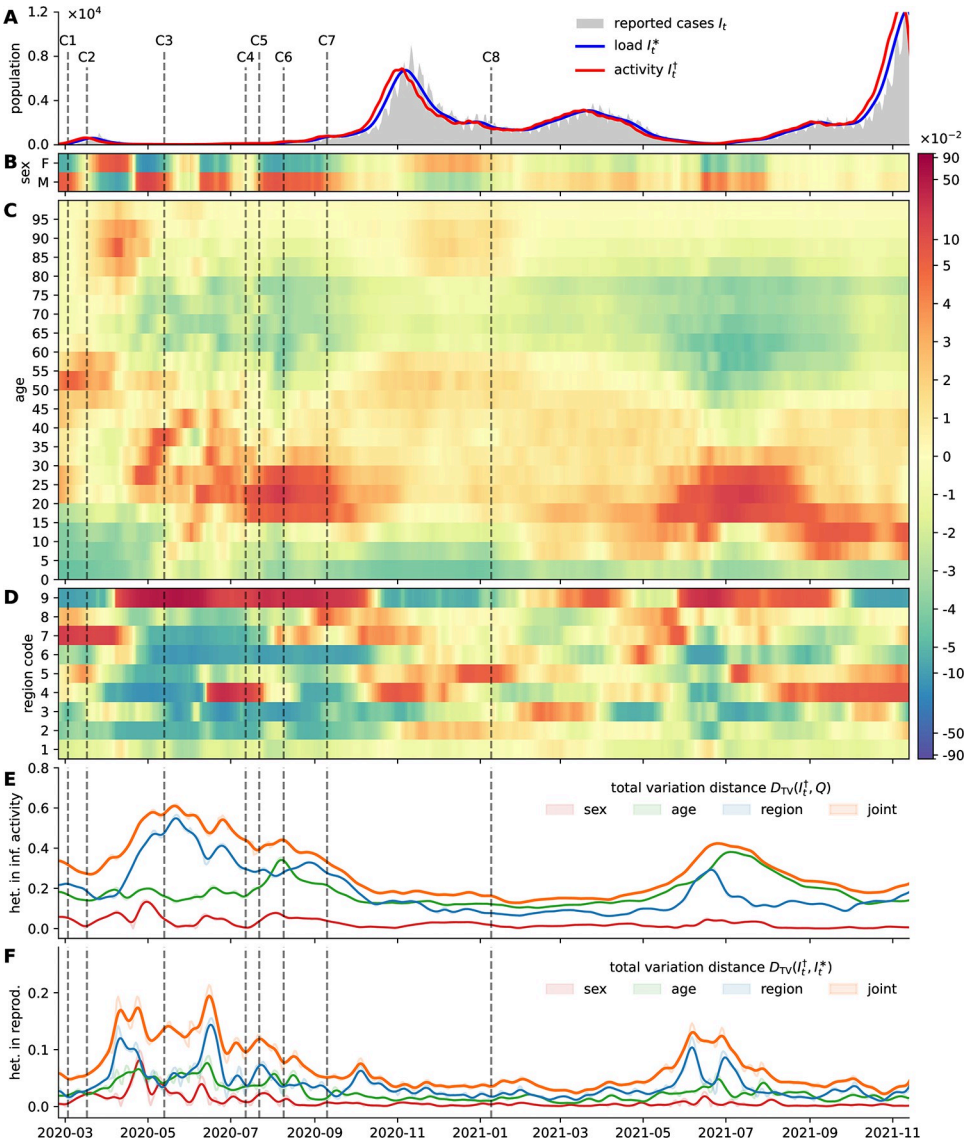

Fig 5. Heterogeneity of infectious activity and in reproduction. Media coverage about certain SSEs in Austria is indicated by vertical dashed lines and the identifiers C1-C8 (see main text). (A) Crude number of reported cases in Austria; estimate for the time series of infections of subsequently reported cases (infectious activity); and estimated time series of the currently infectious population (infectious load). (B-D) For the social dimensions sex, age, and geographic region, the difference of the distribution of the estimate newly infected population (infectious activity) to the distribution of the total population $I_t^\dagger(x) - Q(x)$ is visualized. High values (red) indicate over-representation of infectious activity and low values (blue) indicate under-representation. (E) Quantification of the heterogeneity of infectious activity using the total variation distance measure. To increase lucidity, a smoothed version of the resulting time series is shown. (F) Quantification of the heterogeneity in reproduction; if the distance measure is small, spread is confined to specific social strata; if the distance is large infections shift to previously unaffected social compartments.

https://doi.org/10.1371/journal.pone.0286012.g005

incidence phase, infection clusters in retirement homes across the country were observed (C8). Media reports about specific (noticeable) SSEs occurred in particular during the low-incidence phases of the epidemic. Analogously, the measure for heterogeneity in infectious activity is large only during the same periods indicating that individual SSEs can be distinguished only during such phases of an epidemic. In high-incidence phases, singular SSEs can

be recognized in measured heterogeneity only if they are very significant in size (relative to the current incidence numbers) involving hundreds of infections.

Discussion

The methods proposed in this paper aim to provide insight into the mesoscopic dynamics of epidemic outbreaks based on time series of reported case numbers. A statistical framework was designed to harness the stochasticity in such data for inferring a time-varying *effective aggregate dispersion index* (EffDI) that reflects the progression of heterogeneity in the configuration of individual secondary infections. Large dispersion is in general associated with clustering and the occurrence of superspreading events (SSEs), whereas low dispersion implies homogeneous reproduction and spread [1–4, 10, 42]. Technically, the proposed indicator quantifies the time-varying minimal plausible amount of dispersion that is required for explaining the stochasticity in observed incidence numbers via a statistical model for reproduction. As a consequence, this novel approach serves to distinguish phases of an outbreak, in which infection clusters and SSEs are definite and play a significant role, from phases, in which infection clusters either appear indistinct or are nonexistent. Since the individual distribution of secondary cases is only indirectly reflected in our model, we use the notion of a dimensionless *aggregate* dispersion parameter to emphasize the absence of a direct 'physical' interpretation. Correspondingly, the proposed indicator does not quantify the (plausible) variance in individual secondary infections but rather measures the plausible heterogeneity of the spreading dynamics on a mesoscopic scale. For instance, we assume that EffDI also accounts for the occurrence of multi-generation clusters, instead of assessing the likelihood for individual superspreaders.

Evaluating the EffDI for a number of countries we observe that the transition from periods with moderate case numbers to periods with high incidence ('epidemic waves' or 'peaks') is often accompanied by the early decline of the EffDI. We speculate that this behavior indicates a pending shift in the quality of infectious spread towards the diffusive regime with exponentially growing case numbers. Hence, the EffDI could be a valuable tool for monitoring qualitative changes in the mesoscopic spreading behavior. We anticipate that our novel indicator will be relevant in the planning of containment measures (e.g., deciding between individual-level contact tracing policies and large-scale containment measures) [2–4, 6, 11].

Our approach is founded on a branch of existing statistical models for reproduction that have been developed for inferring the progression of effective reproduction factors [7, 13–20] and a constant *basic dispersion parameter* [1–4, 10–13]. In alignment with the general concept of reproduction, the employed model portrays a relation between the currently contagious population (infectious load) and the new emerging infected population (infectious activity). For the extraction of both quantities from time series of reported case numbers, specific time intervals such as the incubation and generation period as well as delays in case registration must be considered. Furthermore, the implemented testing and reporting procedures often lead to weekly seasonal patterns and artifacts in the data. Experiments showed that modeling the seasonalities of case reporting in the statistical distributions of the disease intervals in a data-driven approach can successfully eliminate artifacts of reporting, but also obliterates the stochasticity, which presumably contains the effects of heterogeneous spreading dynamics. As a consequence we include a technique for the dynamic filtering of weekly periodic patterns within our statistical framework which allows to retain exactly the residual stochasticity of the data while at the same time correctly reproducing the relevant times intervals. Analogous to existing research [13, 16, 18, 20], we confirm that in this setting there exists a trade-off between

synthesizing load and activity based on 'accurate' data-driven models for disease and transmission intervals, and the careful abstraction of transmission dynamics.

We evaluate the progression of EffDI for reported SARS-CoV-2 case numbers in different countries and confirm that inferred effective aggregate dispersion is usually higher in low-incidence regimes. This can be explained by a generally higher relative stochasticity for small sample sizes [42] but also aligns with individual infection clusters only having a significant effect on the overall infection dynamic in such periods. Hence, EffDI can assess the presence and demarcation of infection clusters and SSEs in low-incidence periods, but in high-incidence periods a large number of simultaneous clusters can appear as homogeneous spreading and obliterate the significance of singular spreading events. Nevertheless, our indicator is not trivial in the sense that the time periods with elevated effective aggregate dispersion do not simply agree with periods of large relative variance in the data, constantly low case numbers or a large effective reproduction factor.

To further validate the proposed indicator, we use an Austrian SARS-CoV-2 data-set that includes aggregated social attributes of reported cases and compare the progression of socio-demographic heterogeneity in disease transmissions with the progression of the detected effective aggregate dispersion. Whereas our original method is data-economical, the validation approach relies on higher-resolution case data for quantifying the statistical distance between the socio-demographic configuration of the present and newly infected population (between infectious load and infectious activity). Typically, close-proximity interaction communities, which are the main driver for infection clusters and SSEs, are characterized by distinct social attributes like age or geographic affiliation [1–5, 11]. Hence, we assume that the demarcation and significance of individual infection clusters and SSEs can be deduced from relatively abrupt changes in the observed socio-demographic composition of the incidence population. The indicators show good qualitative correspondence and we can substantiate that EffDI is suitable for differentiating regimes with clustered spreading characteristics form regimes with diffusive spread.

The EffDI complements existing indicators (e.g., effective reproduction factor) which are used for assessing and anticipating the dynamics of epidemic outbreaks. Furthermore, the separation and indication of periods with different spreading characteristics could also improve individual-based [4, 8, 9, 44] and aggregate models for the simulation and forecasting of epidemics and could provide a reference for combining or switching between different modeling approaches. We also assume that providing indications about the time-varying progression of heterogeneity in epidemic outbreaks could contribute to the general perception of epidemic spreading as complex dynamics on multiple scales. Generally, the usability of quantifiers and numerical indicators is inevitably affected by the availability and accuracy of data. Especially in the epidemiology of infectious diseases, the data often provides an indirect view on the actual situation and makes it necessary to combine data of different origin and to map the dynamic aspects of spreading by the means of mathematical models. An immediate consequence is to include also the variability or stochasticity of reported case data itself in such models and in the calculation of numerical indicators. For instance, in the inference of effective reproduction factors, the relative stochasticity of the data is used for assessing the uncertainty of statistical estimates [13, 20, 41]. Analogously, we designed our approach to use the stochasticity in the data for inferring information about the infection dynamics on a mesoscopic scale, that can not directly be concluded from aggregate case numbers. However, for investigating the plain stochasticity or regularity in time series of reported case numbers there exist more suitable mathematical approaches like approximate entropy or spectral methods that do not integrate a statistical model for reproduction.

We provide the algorithms developed in this paper as a Python programming package (EffDI Python package) that can operate on any time series of crude reported case numbers. We anticipate that the proposed concept of an indicator for effective aggregate dispersion can further be improved in terms of the reflection of artifacts and seasonal patterns in the data and by investigating the robustness against changes in the assumed characteristics of the disease. Certain changes in the characteristics of the disease should be factored into the calculation of infectious load and activity. On the other hand, changes in the heterogeneity of spreading dynamics, that are caused by the emergence of new viral variants or by the implementation of intervention measures, should be reflected in the temporal progression of an indicator for transmission heterogeneity. This suggests that numerical indicators such as EffDI should not be investigated without regarding additional epidemiological data. In our current approach, abrupt changes in the seasonality of the data, which are caused by systematic changes in case reporting operations, can lead to the artificial inflation of inferred effective aggregate dispersion. Furthermore, our framework does currently not regard truncation or other aspects that should be considered in the analysis of time series data. Besides technical improvements, we assume that investigating effective aggregate dispersion with synthetic data that was obtained from individual-based or network models [4–9, 44] could provide additional insight and improvements.

Methods

Disease and transmission intervals

Motivation. In the study of the epidemiology of transmissible diseases usually only a subset of the relevant time intervals is accessible via collected data. We relate an extended set of disease and transmission specific intervals in a stochastic model to infer information that cannot be retrieved directly from collected data in most cases.

Corresponding to the notation introduced in *Results*, we denote the offset between infection and (potential) case registration as the *forward reporting offset* $\Delta_{\text{rep}^\dagger}$ and the time between detection and a (potential and subsequently detected) secondary infection as the *backward reporting offset* Δ_{rep^*}. For an infection pair AB let T^A_{reg} be the point in time when a patient A was registered, let T^B_{inf} be the point in time when the secondary case (patient B) acquired the infection from patient A and let T^B_{reg} be the time of registration of patient B. Then

$$\begin{aligned} T^B_{\text{inf}} &= T^A_{\text{reg}} + \Delta^{AB}_{\text{rep}^*}, \\ T^B_{\text{inf}} &= T^B_{\text{reg}} - \Delta^B_{\text{rep}^\dagger}, \end{aligned}$$

where $\Delta^B_{\text{rep}^\dagger}$ and $\Delta^{AB}_{\text{rep}^*}$ are the specific reporting offsets for the infection pair AB. In addition to the specific infection and reporting times T_{inf} and T_{reg}, also regard the time of symptom onset T_{sym} and denote the time intervals

$$\begin{aligned} \Delta^{AB}_{\text{gen}} &= T^B_{\text{inf}} - T^A_{\text{inf}}, \\ \Delta^{AB}_{\text{ser}} &= T^B_{\text{sym}} - T^A_{\text{sym}}, \\ \Delta^{AB}_{\text{case}} &= T^B_{\text{reg}} - T^A_{\text{reg}}, \end{aligned}$$

as the *generation*, *serial* and *case* interval. We further write

$$\begin{aligned} \Delta^A_{\text{inc}} &= T^A_{\text{sym}} - T^A_{\text{inf}}, \\ \Delta^A_{\text{reg}} &= T^A_{\text{reg}} - T^A_{\text{sym}}. \end{aligned}$$

for the *incubation period* and the *registration delay* (for patient *B* respectively). A diagram relating all disease and transmission intervals is provided in Fig 1A.

Data. For quantifying infectious load and activity (compare Eq (2)), the time series of reported cases and the statistical distributions of the reporting offsets are required. In literature and studies, however, most often the generation [27, 36, 45] and serial intervals [29, 30, 32, 35, 37, 46] as well as the incubation period [31, 34, 35, 47] are investigated and the underlying data usually stems from a relatively small number of closely tracked infection pairs. Parameters of statistical distributions found in literature and visual comparison are provided in S1 Data and S1 Text. In the presented stochastic inference approach we use 'averaged' versions of the found distributions.

Delays in registration and reporting, on the other hand, are less frequently studied because they are very specific to administrative processes including changing reporting procedures and laboratory schedules. We use a data-set containing daily histograms about the time between symptom onset and recording that was extracted for us from the Austrian national surveillance system. We removed outliers reporting large negative and positive delays, subdivided the data by registration date at 2020-06-01, 2020-10-01 and 2021-03-01, and fitted gamma distributions separately for each of the four resulting periods and for each day of the week. The inferred parameters are provided in S2 Data. In our analysis, we observe smaller delays in the later stages of the epidemic; presumably due to improved testing and registration procedures. But it is also possible that the emergence of new variance with different incubation times could influence the duration of reporting delays. Weekly patterns can be recognized with slightly longer reporting delays for cases that first showed symptoms during the weekends. The obtained interval distributions are specific to the Austrian setting but we assume that qualitatively similar results should be obtained for other countries. Visualizations of the data is provided in S2 Text.

Population specific time intervals are less frequently studied than intervals in transmissions that are specific to the disease. As a consequence, in models for the inference of reproductive numbers, often the serial or generation interval is used as an approximation of the case interval [3, 7, 13, 19] (compare Eq (3)). In fact, we find that $\mathbb{E}[\Delta_{\text{gen}}] \approx \mathbb{E}[\Delta_{\text{ser}}] \approx \mathbb{E}[\Delta_{\text{case}}]$. Nevertheless, the generation interval is always a positive time period, whereas the serial and case intervals can take negative duration. We distinguish statistical distributions for the serial interval found in literature that assume strictly positive values ('pSI') and models that allow negative serial intervals ('nSI'). We use the latter model throughout our presentation but highlight some effects and differences when choosing the 'pSI' model in S6 Text. Due to the accumulated variability in individual disease progression and reporting, we may further assume that variance in the case interval distribution is larger than in the generation and serial interval distributions. Furthermore, the incubation period Δ_{inc} is always a positive time period, and we anticipate for the registration delay Δ_{reg} a distribution with its mass concentrated close to and right of the origin. Negative registration delays can occur if symptom onset happens after the infection was discovered. Finally, for the backward reporting offset Δ_{rep^*} we anticipate a statistical distribution that yields mostly negative values because once a patient is discovered, the likelihood of posterior secondary infections should be minimal. The forward reporting offset $\Delta_{\text{rep}^\dagger}$ presumably is a strictly positive interval with a shape strongly influenced by the incubation period distribution. All intervals can be subject to gradual change over time. This can be attributed to the emergence of more aggressive variants of a virus, which may lead to decreased incubation periods, to the improvement of surveillance and contact tracing, which can lead to smaller reporting delays (observed in the data), or to a generally altered social contact behavior and immunity.

Inference approach. The following equation model corresponds to the schematic in Fig 1A and brings into relation above time intervals occurring in infection pairs *AB*,

$$\begin{aligned}
\Delta_{\text{ser}}^{AB} &= \Delta_{\text{gen}}^{AB} - \Delta_{\text{inc}}^{A} + \Delta_{\text{inc}}^{B} \\
\Delta_{\text{case}}^{AB} &= \Delta_{\text{gen}}^{AB} - \Delta_{\text{inc}}^{A} - \Delta_{\text{reg}}^{A} + \Delta_{\text{inc}}^{B} + \Delta_{\text{reg}}^{B} \\
\Delta_{\text{case}}^{AB} &= \Delta_{\text{ser}}^{AB} - \Delta_{\text{reg}}^{A} + \Delta_{\text{reg}}^{B} \\
\Delta_{\text{rep}^*}^{A} &= \Delta_{\text{gen}}^{AB} - \Delta_{\text{inc}}^{A} - \Delta_{\text{reg}}^{A} \\
\Delta_{\text{rep}^*}^{A} &= \Delta_{\text{ser}}^{AB} - \Delta_{\text{inc}}^{B} - \Delta_{\text{reg}}^{A} \\
\Delta_{\text{rep}^\dagger}^{A} &= \Delta_{\text{inc}}^{A} + \Delta_{\text{reg}}^{A} \\
\Delta_{\text{rep}^\dagger}^{B} &= \Delta_{\text{inc}}^{B} + \Delta_{\text{reg}}^{B}.
\end{aligned} \tag{9}$$

We designate the generation, serial and incubation period (Δ_{gen}, Δ_{ser}, Δ_{inc}) as exogenous variables that can be modeled with statistical distributions found in literature (S1 Data and S1 Text). A further exogenous variable is the registration delay Δ_{reg} (S2 Data and S2 Text). Accordingly, the remaining 'endogenous' variables left for inference are the case interval Δ_{case} and the forward and backward reporting offsets $\Delta_{\text{rep}^\dagger}$, Δ_{rep^*}. To map the equation system in a stochastic simulation framework, we differentiate between distinct instances of random variables on the left-hand side in Eq (9) and model the difference between instances of the same variable as normal errors with a standard deviation of 3 days. We then implement the resulting expressions (S3 Text) in a stochastic simulation framework (Python PyMC3 library [48]) and employ gradient-based MCMC algorithms to obtain stochastic samples of all random variables that comply with the formulated constraints (error terms). Corresponding Python programming code is provided in S1 Code.

We compare the obtained parameters of the fitted distributions of all exogenous variables with their prior counterparts that were found in literature and data (S4 Text). By this we can assure that the posterior distributions only deviate marginally from the initially provided data and models while simultaneously adhering to the equation model. We further fit statistical models (skew normal distributions) to the synthetic samples of the remaining endogenous variables (case interval and reporting offset distributions) and collate their qualitative characteristics with above considerations. To honour the temporal transformation of disease intervals and weekly patterns in reporting delays, this procedure was separately performed for distinct time periods of the SARS-CoV-2 epidemic in Austria and for each day of the week. Hence, for distinct time periods and for different days of the week specific statistical models for the reporting offset distributions were obtained (compare Fig 1C).

In S3 Data we provide the obtained distribution parameters for all scenarios (weekday and phase of the epidemic in Austria). In S5 Text we visualize the obtained reporting offset distributions for all scenarios. In S6 Text we compare the results of our approach for real and positive serial intervals.

Framework for the inference of effective aggregate dispersion

Motivation. In the common approach for inferring reproduction dynamics from time series of reported cases (see Results) it is assumed that infectious activity ($I^\dagger$) can be modeled with a negative binomial distribution (see also Eq (5)),

$$Y_t \sim \text{NB}\left(kI_t^*, \frac{R_t}{R_t + k}\right), \tag{10}$$

where I_t^* is the current infectious load, R_t is the effective reproduction number, and $k > 0$ is a constant dispersion parameter. We obtain for the mean and variance,

$$\mu_{Y_t} = I_t^* R_t, \qquad \sigma_{Y_t}^2 = I_t^* R_t + \frac{I_t^* R_t^2}{k} \tag{11}$$

and Chebyshev's inequality yields

$$\mathrm{P}\left(\frac{|Y_t - \mu_{Y_t}|}{\mu_{Y_t}} \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\left(\frac{1}{I_t^* R_t} + \frac{1}{k I_t^*}\right), \qquad \varepsilon > 0. \tag{12}$$

Hence, the relative deviation of infectious activity converges to zero in probability for increasing load and fixed k, that is

$$\operatorname*{plim}_{I_t^* \to \infty} \frac{|Y_t - \mu_{Y_t}|}{\mu_{Y_t}} = 0. \tag{13}$$

This means that with the model Y_t, independent of the dispersion parameter k, the relative 'stochasticity' of infectious activity diminishes with increasing infectious load. In other words, the above model assumes highly regular time series of infectious activity during regimes with increased load irrespective of the presence of dispersion in the transmission dynamics. In turn, this indicates that Y_t is not suitable for assessing the variable amount of traceable or *effective* dispersion in the transmission dynamics of an outbreak.

We now consider a slightly modified model for infectious activity (see also Eq (6))

$$Z_t \sim \mathrm{Gamma}\left(\kappa_t, \frac{I_t^* R_t}{\kappa_t}\right), \tag{14}$$

where R_t is again the effective reproduction number and $\kappa_t > 0$ is a time-dependent aggregate dispersion parameter. It holds that

$$\mu_{Z_t} = R_t I_t^*, \qquad \sigma_{Z_t}^2 = \frac{(R_t I_t^*)^2}{\kappa_t}. \tag{15}$$

Whereas $\mu_{Z_t} = \mu_{Y_t}$, the main difference between the random variables Z_t and Y_t is how their respective standard deviations scale with the infectious load I_t^*. In particular, we have $\sigma_{Y_t} = O(\sqrt{I_t^*})$ and $\sigma_{Z_t} = O(I_t^*)$ for $I_t^* \to \infty$. Hence, in contrast to Eq (12), applying the Chebyshev's inequality for Z_t we obtain

$$\mathrm{P}\left(\frac{|Z_t - \mu_{Z_t}|}{\mu_{Z_t}} \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2 \kappa_t}, \qquad \varepsilon > 0, \tag{16}$$

which does not imply a reduction of the relative deviation from the expectation with increased load I_t^*. Note that in general this behavior can only be obtained when σ_{Z_t} grows at least linearly with I_t^* (further technical comparison of both models is provided in S8 Text). Hence, in contrast to Y_t, when small dispersion and homogeneous spreading is assumed (large κ_t) in model Z_t, then the 'stochasticity' of simulated infectious activity is small irrespective of the current infectious load.

We now use Z_t as a benchmark model for investigating time-varying aggregate dispersion in the following sense: If we can further extend the model Z_t in such a way that it inherently explains most of the (seasonal) variance in $I_t^\dagger$ caused by external factors such as the testing and reporting regime, it is possible to choose the time-dependent aggregate dispersion parameter

κ_t increasingly large only during phases of low relative 'stochasticity' without making the observed infectious activity implausible.

Residual stochasticity. When attempting to quantify the amount of variance in a time series of aggregate reported case numbers that can be attributed to the dispersion in the number of individual secondary infections, we are faced with two major problems.

First, models for the reproduction dynamics such as Y_t in Eq (10) are defined by time-varying parameters that can easily change between two consecutive days. The observed time series of aggregate case numbers, however, only yields a single entry for each day. Hence, in general, we only have a single sample for a fitted model Y_t, from which it is clearly impossible to obtain any notion of empirical variation. One way to address this issue is to assume that the parameters of a model for the reproduction dynamics of an outbreak remain unchanged for a given number of days [13]. Under this assumption, we can sample infectious load and infectious activity pairs for each day in the selected time window and measure how well a model that was fitted on these pairs explains the data observed across the respective time frame.

Secondly, large parts of the variance in a time series of daily aggregate case numbers is caused by independent factors such as the individual testing behavior, the current setup of the testing regime, or the workflow of laboratories and governmental institutions. These contributions to the overall variance in the observed time series need to be separated from the variation caused by the dispersion in the number of secondary infections.

In the following, we define a model for the number of secondary infections that mitigates both of these issues and thus enables us to quantify how much dispersion at least needs to be assumed during a given period in the pandemic such that the observed variance of daily aggregate case numbers is in fact plausible.

Inference approach. To assess the plausibility of the observed data around a day t given an aggregate dispersion parameter κ_t in the model Eq (14), we first fit the effective reproduction numbers R_t on a fixed window around t during which we assume that the parameters describing the current infection dynamics remain constant. A typical approach in the literature is to use a constant model for R_t [13]. However, a constant model cannot account for variance in the data caused by daily seasonalities in the reporting or strong linear trends, both of which are usually present in daily incidence time series. Here, we will consider a model which combines a linear trend with a daily seasonal constant, namely

$$R_t = c_0 t + s_{t \bmod 7}, \tag{17}$$

where $c_0, s_0, \ldots, s_6 \in \mathbb{R}$. Let here for convenience $\mathbf{e}_i \in \mathbb{R}^7$ for $i \in \{0, \ldots, 6\}$ denote the $(i+1)$-th unit row vector and let $I_t^\dagger = I_t$ be the infectious activity on day t according to Eq (3). By assuming that the parameters c_0, s_0, . . ., s_6 remain constant τ_0 days before and τ_1 days after t, we define the linear problem

$$\begin{pmatrix} 0 & I^*_{t-\tau_0}\mathbf{e}_{(t-\tau_0) \bmod 7} \\ I^*_{t-\tau_0+1} & I^*_{t-\tau_0+1}\mathbf{e}_{(t-\tau_0+1) \bmod 7} \\ \vdots & \vdots \\ \tau_0 I^*_t & I^*_t \mathbf{e}_{t \bmod 7} \\ \vdots & \vdots \\ (\tau_0+\tau_1-1)I^*_{t+\tau_1-1} & I^*_{t+\tau_1-1}\mathbf{e}_{(t+\tau_1+-1) \bmod 7} \\ (\tau_0+\tau_1)I^*_{t+\tau_1} & I^*_{t+\tau_1}\mathbf{e}_{(t+\tau_1) \bmod 7} \end{pmatrix} \cdot \begin{pmatrix} c_0 \\ s_0 \\ \vdots \\ s_6 \end{pmatrix} = \begin{pmatrix} I^\dagger_{t-\tau_0} \\ I^\dagger_{t-\tau_0+1} \\ \vdots \\ I^\dagger_t \\ \vdots \\ I^\dagger_{t+\tau_1-1} \\ I^\dagger_{t+\tau_1} \end{pmatrix}, \tag{18}$$

which can be solved with an ordinary least squares approach. Substituting the obtained parameters in Eq (17) then yields a fit for the effective reproduction numbers $R_{t-\tau_1}, \ldots, R_{t+\tau_1}$ on the selected window of days around t. For a fixed time-dependent aggregate dispersion parameter κ_t, which we assume to also remain constant in the period from $t - \tau_0$ to $t + \tau_1$, this fully defines the model Z_t for the selected time window. To calculate the plausibility of this model given the observed data within the selected time window, we define the following test statistic based on the squared distance from the daily expectation:

$$\mathrm{T}(x_{-\tau_0}, \ldots, x_0, \ldots, x_{\tau_1}) = \sum_{n=-\tau_0}^{\tau_1} (x_n - \mu_{Z_{t-n}})^2, \tag{19}$$

where $\mu_{Z_{t-n}} = R_{t-n} I^*_{t-n}$. The plausibility of the model Z_t given the observed case data and a fixed parameter κ_t can now be assessed via the p-value that represents the probability that the test statistic Eq (19) for samples from the random variables $Z_{t-\tau_0}, \ldots, Z_{t+\tau_1}$ is greater or equal than the test statistic for the estimated infectious activities $I^\dagger_t$, that is,

$$p_t(\kappa_t) = \mathrm{P}(\mathrm{T}(I^\dagger_{t-\tau_0}, \ldots, I^\dagger_{t+\tau_1}) \leq \mathrm{T}(Z_{t-\tau_0}, \ldots, Z_{t+\tau_1})). \tag{20}$$

Note that the test statistic can be applied to integer- and real-valued time series alike. For a plausibility threshold $p \in [0, 1]$, we write

$$\kappa_t(p) = \sup\{\kappa \colon p_t(\kappa) \geq p\}, \tag{21}$$

to denote the largest aggregate dispersion parameter for which the model Z_t is plausible given the observed case data around day t. The effective aggregate dispersion index (EffDI) is then defined as the square root of the reciprocal of $\kappa_t(p)$, that is,

$$\mathrm{EffDI}_t = \kappa_t(p)^{-1/2}. \tag{22}$$

Supporting information

S1 Code. Python programming code for the stochastic inference of disease interval distributions. The programming code implements the stochastic inference approach for the reporting offset distributions presented in *Methods*. The program uses the data provided in and S1 and S2 Data.
(PY)

S1 Data. Parameters of statistical distributions of disease intervals. We provide a machine-readable file in the JSON format containing the parameters and statistics of the distributions of disease intervals found in literature. The references are listed in S1 Text.
(JSON)

S2 Data. Parameters of reporting delay distributions. We provide a machine-readable file in the JSON format containing the parameters of reporting delay distributions that were fitted to Austrian data.
(JSON)

S3 Data. Parameters of the inferred distributions of disease intervals. We provide a machine-readable file in the JSON format containing the parameters and statistics of inferred forward and backward reporting offset distributions and the inferred distributions of the case interval.
(JSON)

S1 Text. Visualization of statistical distributions of disease intervals. Visualization of the data provided in S1 Data.
(PDF)

S2 Text. Visualization of reporting delay distributions. Visualization of the statistical distributions provided in S2 Data.
(PDF)

S3 Text. Equations and constraints for stochastic simulation. Formal representation of the equations defining the stochastic model for the MCMC inference approach.
(PDF)

S4 Text. Posterior distributions of the MCMC inference approach. Analysis of the posterior distributions obtained by stochastic inference.
(PDF)

S5 Text. Visual display of the inferred disease interval distributions. Visualization of the statistical distributions provided in S3 Data.
(PDF)

S6 Text. Comparison of inferred interval models under different conditions for the positivity of the serial intervals. Additional visualization of the statistical distributions provided in S3 Data.
(PDF)

S7 Text. List of selected superspreading events in Austria. Details on the SSEs that are indicated in Fig 5.
(PDF)

S8 Text. Additional details on the statistical models for reproduction. Formal analysis of the reproduction models in Eqs 4–6.
(PDF)

S9 Text. Parameter variation studies. EffDI under the variation of the parameters of the statistical distributions of disease intervals. We investigate the robustness of EffDI against the shape of the reporting offset distributions and the behavior of EffDI under the transition between Eqs 2 and 3.
(PDF)

S10 Text. Emergence of SARS-CoV-2 variants and EffDI. This supporting text compares the emergence of new variants with the behavior of EffDI based on Austrian case data.
(PDF)

Acknowledgments

We thank the Austrian National Public Health Institute (GÖG, www.goeg.at) for their support and the provisioning of research data.

Code availability

The EffDI package is available on the Python Package Index (PyPI) and GitHub (https://github.com/mdsunivie/EffDI). The Python programing code for inferring statistical distributions of disease intervals via MCMC methods is available in the supplementary material (S1 Code).

Author Contributions

Conceptualization: Günter Schneckenreither, Lukas Herrmann, Rafael Reisenhofer.

Formal analysis: Günter Schneckenreither, Lukas Herrmann, Rafael Reisenhofer, Niki Popper, Philipp Grohs.

Funding acquisition: Niki Popper, Philipp Grohs.

Investigation: Günter Schneckenreither, Lukas Herrmann, Rafael Reisenhofer.

Methodology: Günter Schneckenreither, Lukas Herrmann, Rafael Reisenhofer.

Project administration: Niki Popper, Philipp Grohs.

Resources: Niki Popper, Philipp Grohs.

Software: Günter Schneckenreither, Lukas Herrmann, Rafael Reisenhofer.

Supervision: Niki Popper, Philipp Grohs.

Visualization: Günter Schneckenreither, Lukas Herrmann, Rafael Reisenhofer.

Writing – original draft: Günter Schneckenreither, Lukas Herrmann, Rafael Reisenhofer.

Writing – review & editing: Günter Schneckenreither, Lukas Herrmann, Rafael Reisenhofer, Niki Popper, Philipp Grohs.

References

1. Adam DC, Wu P, Wong JY, Lau EHY, Tsang TK, Cauchemez S, et al. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. Nature Medicine. 2020; 26(11):1714–1719. https://doi.org/10.1038/s41591-020-1092-0 PMID: 32943787
2. Sun K, Wang W, Gao L, Wang Y, Luo K, Ren L, et al. Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. Science. 2021; 371(6526). https://doi.org/10.1126/science.abe2424 PMID: 33234698
3. Althouse BM, Wenger EA, Miller JC, Scarpino SV, Allard A, Hébert-Dufresne L, et al. Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for interventions and control. PLOS Biology. 2020; 18(11):e3000897. https://doi.org/10.1371/journal.pbio.3000897 PMID: 33180773
4. Sneppen K, Nielsen BF, Taylor RJ, Simonsen L. Overdispersion in COVID-19 increases the effectiveness of limiting nonrepetitive contacts for transmission control. Proceedings of the National Academy of Sciences. 2021; 118(14):e2016623118. https://doi.org/10.1073/pnas.2016623118 PMID: 33741734
5. Schneckenreither G, Popper N. Dynamic multiplex social network models on multiple time scales for simulating contact formation and patterns in epidemic spread. 2017 Winter Simulation Conference (WSC). 2017; p. 4324–4335.
6. Thurner S, Klimek P, Hanel R. A network-based explanation of why most COVID-19 infection curves are linear. Proceedings of the National Academy of Sciences. 2020; 117(37):22684–22689. https://doi.org/10.1073/pnas.2010398117 PMID: 32839315
7. Wallinga J. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. American Journal of Epidemiology. 2004; 160(6):509–516. https://doi.org/10.1093/aje/kwh255 PMID: 15353409
8. Bicher M, Rippinger C, Urach C, Brunmeir D, Siebert U, Popper N. Evaluation of Contact-Tracing Policies against the Spread of SARS-CoV-2 in Austria: An Agent-Based Simulation. Medical Decision Making. 2021; 41(8):1017–1032. https://doi.org/10.1177/0272989X211013306 PMID: 34027734
9. Rippinger C, Bicher M, Urach C, Brunmeir D, Weibrecht N, Zauner G, et al. Evaluation of undetected cases during the COVID-19 epidemic in Austria. BMC Infectious Diseases. 2021; 21(1). https://doi.org/10.1186/s12879-020-05737-6 PMID: 33441091
10. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005; 438(7066):355–359. https://doi.org/10.1038/nature04153 PMID: 16292310

11. Donnat C, Holmes S. Modeling the heterogeneity in COVID-19's reproductive number and its impact on predictive scenarios. Journal of Applied Statistics. 2021; p. 1–29. https://doi.org/10.1080/02664763.2021.1941806

12. Endo A, Abbott S, Kucharski AJ, Funk S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. Wellcome Open Research. 2020-07; 5:67. https://doi.org/10.12688/wellcomeopenres.15842.3 PMID: 32685698

13. Cori A, Ferguson NM, Fraser C, Cauchemez S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. American Journal of Epidemiology. 2013; 178 (9):1505–1512. https://doi.org/10.1093/aje/kwt133 PMID: 24043437

14. Abbott S, Hellewell J, Thompson RN, Sherratt K, Gibbs HP, Bosse NI, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open Research. 2020; 5:112. https://doi.org/10.12688/wellcomeopenres.16006.2

15. O'Driscoll M, Harry C, Donnelly CA, Cori A, Dorigatti I. A Comparative Analysis of Statistical Methods to Estimate the Reproduction Number in Emerging Epidemics, With Implications for the Current Coronavirus Disease 2019 (COVID-19) Pandemic. Clinical Infectious Diseases. 2020.

16. Fraser C. Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. PLoS ONE. 2007; 2(8):e758. https://doi.org/10.1371/journal.pone.0000758 PMID: 17712406

17. Gostic KM, McGough L, Baskerville EB, Abbott S, Joshi K, Tedijanto C, et al. Practical considerations for measuring the effective reproductive number, Rt. PLOS Computational Biology. 2020; 16(12): e1008409. https://doi.org/10.1371/journal.pcbi.1008409 PMID: 33301457

18. Koyama S, Horie T, Shinomoto S. Estimating the time-varying reproduction number of COVID-19 with a state-space method. PLOS Computational Biology. 2021; 17(1):e1008679. https://doi.org/10.1371/journal.pcbi.1008679 PMID: 33513137

19. Obadia T, Haneef R, Boëlle PY. The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. BMC Medical Informatics and Decision Making. 2012; 12(1). https://doi.org/10.1186/1472-6947-12-147 PMID: 23249562

20. Thompson RN, Stockwin JE, van Gaalen RD, Polonsky JA, Kamvar ZN, Demarsh PA, et al. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. Epidemics. 2019; 29:100356. https://doi.org/10.1016/j.epidem.2019.100356 PMID: 31624039

21. Günther F, Bender A, Katz K, Küchenhoff H, Höhle M. Nowcasting the COVID-19 pandemic in Bavaria. Biometrical Journal. 2020; 63(3):490–502. https://doi.org/10.1002/bimj.202000112 PMID: 33258177

22. an der Heiden M, Hamouda O. Schätzung der aktuellen Entwicklung der SARS-CoV-2- Epidemie in Deutschland—Nowcasting. Epidemiologisches Bulletin. 2020; 2020(17):10–15. https://doi.org/10.25646/6692.4

23. Nunes B, Natário I, Carvalho ML. Nowcasting influenza epidemics using non-homogeneous hidden Markov models. Statistics in Medicine. 2012; 32(15):2643–2660. https://doi.org/10.1002/sim.5670 PMID: 23124850

24. Richter L, Schmid D, Stadlober E. Imputation des Erkrankungs-Datums bei unvollständigen Daten im Rahmen der COVID-19 Epidemie, Östrreich. Austrian Agency for Health and Food Safety (AGES); 2020. Available from: https://wissenaktuell.ages.at/imputation-des-erkrankungs-datums-bei-unvollstaendigen-daten-im-rahmen-der-covid-19-epidemie-oesterreich/.

25. Challen R, Brooks-Pollock E, Tsaneva-Atanasova K, Danon L. Meta-analysis of the SARS-CoV-2 serial interval and the impact of parameter uncertainty on the COVID-19 reproduction number. medRxiv. 2020. https://doi.org/10.1101/2020.11.17.20231548

26. Fine PEM. The Interval between Successive Cases of an Infectious Disease. American Journal of Epidemiology. 2003; 158(11):1039–1047. https://doi.org/10.1093/aje/kwg251 PMID: 14630599

27. Ganyani T, Kremer C, Chen D, Torneri A, Faes C, Wallinga J, et al. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. Eurosurveillance. 2020; 25(17). https://doi.org/10.2807/1560-7917.ES.2020.25.17.2000257 PMID: 32372755

28. Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis CI, Russell TW, et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. The Lancet Global Health. 2020; 8(4):e488–e496. https://doi.org/10.1016/S2214-109X(20)30074-7 PMID: 32119825

29. He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nature Medicine. 2020; 26(5):672–675. https://doi.org/10.1038/s41591-020-0869-5 PMID: 32296168

30. Ali ST, Wang L, Lau EHY, Xu XK, Du Z, Wu Y, et al. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. Science. 2020; 369(6507):1106–1109. https://doi.org/10.1126/science.abc9004 PMID: 32694200

31. Backer JA, Klinkenberg D, Wallinga J. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. Eurosurveillance. 2020; 25(5). https://doi.org/10.2807/1560-7917.ES.2020.25.5.2000062 PMID: 32046819

32. Du Z, Xu X, Wu Y, Wang L, Cowling BJ, Meyers LA. Serial Interval of COVID-19 among Publicly Reported Confirmed Cases. Emerging Infectious Diseases. 2020; 26(6):1341–1343. https://doi.org/10.3201/eid2606.200357 PMID: 32191173

33. Griffin J, Casey M, Collins Á, Hunt K, McEvoy D, Byrne A, et al. Rapid review of available evidence on the serial interval and generation time of COVID-19. BMJ Open. 2020; 10(11):e040263. https://doi.org/10.1136/bmjopen-2020-040263 PMID: 33234640

34. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Annals of Internal Medicine. 2020; 172(9):577–582. https://doi.org/10.7326/M20-0504 PMID: 32150748

35. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. New England Journal of Medicine. 2020; 382(13):1199–1207. https://doi.org/10.1056/NEJMoa2001316 PMID: 31995857

36. Ng SHX, Kaur P, Kremer C, Tan WS, Tan AL, Hens N, et al. Estimating Transmission Parameters for COVID-19 Clusters by Using Symptom Onset Data, Singapore, January–April 2020. Emerging Infectious Diseases. 2021; 27(2):582–585. https://doi.org/10.3201/eid2702.203018 PMID: 33496243

37. Nishiura H, Linton NM, Akhmetzhanov AR. Serial interval of novel coronavirus (COVID-19) infections. International Journal of Infectious Diseases. 2020; 93:284–286. https://doi.org/10.1016/j.ijid.2020.02.060 PMID: 32145466

38. Cauchemez S, Carrat F, Viboud C, Valleron AJ, Boëlle PY. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Statistics in Medicine. 2004; 23 (22):3469–3487. https://doi.org/10.1002/sim.1912 PMID: 15505892

39. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature. 2020; 584(7820):257–261. https://doi.org/10.1038/s41586-020-2405-7 PMID: 32512579

40. Ferguson NM, Cummings DAT, Cauchemez S, Fraser C, Riley S, Meeyai A, et al. Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature. 2005; 437(7056):209–214. https://doi.org/10.1038/nature04017 PMID: 16079797

41. Johnson KD, Beiglböck M, Eder M, Grass A, Hermisson J, Pammer G, et al. Disease momentum: Estimating the reproduction number in the presence of superspreading. Infectious Disease Modelling. 2021; 6:706–728. https://doi.org/10.1016/j.idm.2021.03.006 PMID: 33824936

42. Lloyd-Smith JO. Maximum Likelihood Estimation of the Negative Binomial Dispersion Parameter for Highly Overdispersed Data, with Applications to Infectious Diseases. PLoS ONE. 2007; 2(2):e180. https://doi.org/10.1371/journal.pone.0000180 PMID: 17299582

43. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases. 2020; 20(5):533–534. https://doi.org/10.1016/S1473-3099(20)30120-1 PMID: 32087114

44. Popper N, Zechmeister M, Brunmeir D, Rippinger C, Weibrecht N, Urach C, et al. Synthetic Reproduction and Augmentation of COVID-19 Case Reporting Data by Agent-Based Simulation. Data Science Journal. 2021; 20. https://doi.org/10.5334/dsj-2021-016

45. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L, et al. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science. 2020; 368(6491): eabb6936. https://doi.org/10.1126/science.abb6936 PMID: 32234805

46. Richter L, Schmid D, Chakeri A, Maritschnik S, Pfeiffer S, Stadlober E. Schätzung des seriellen Intervalles von COVID19, Österreich. Austrian Agency for Health and Food Safety (AGES); 2020. Available from: https://wissenaktuell.ages.at/schaetzung-des-seriellen-intervalles-von-covid19-oesterreich/.

47. Zhang J, Litvinova M, Wang W, Wang Y, Deng X, Chen X, et al. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. The Lancet Infectious Diseases. 2020; 20(7):793–802. https://doi.org/10.1016/S1473-3099(20)30230-9 PMID: 32247326

48. Salvatier J, Wiecki TV, Fonnesbeck C. Probabilistic programming in Python using PyMC3. PeerJ Computer Science. 2016; 2:e55. https://doi.org/10.7717/peerj-cs.55