Unjustified Poisson assumptions lead to overconfident estimates of the effective reproductive number

Barbora Němcová[1,2], Isaac H. Goldstein[3], Jessalyn Sebastian[4], Volodymyr M. Minin[4], Johannes Bracher[1,5]

[1]Institute of Statistics, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Helmholtz Information and Data Science School for Health, Heidelberg/Karlsruhe, Germany
[3]Department of Statistics, Stanford University, Stanford, CA, USA
[4]Department of Statistics, University of California, Irvine, CA, USA
[5]Computational Statistics Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

Abstract

Time-varying effective reproductive numbers of infectious diseases are commonly estimated using renewal equation models. In the widely applied `R` package `EpiEstim`, this approach is combined with a Poisson distributional assumption. This has been criticized on various occasions, mostly on grounds of general model realism or a desire to estimate overdispersion parameters. Here we argue that an important issue arising from the Poisson assumption is that inference about the effective reproductive number becomes overconfident in presence of overdispersion. By how much standard errors are underestimated follows in a straightforward manner from theory on generalized linear models. We therefore recommend to replace the Poisson assumption by quasi-Poisson or negative binomial extensions, and contrast their respective properties. We illustrate our arguments in three examples on Ebola, pandemic influenza and COVID-19.

1 Introduction

The effective reproductive number, defined as "the expected number of new infections caused by an infectious individual in a population where some individuals may no longer be susceptible" (Gostic et al., 2020), is a key quantity characterizing disease outbreak dynamics. It is commonly estimated using renewal equation models, often combined with Poisson distributional assumptions. This approach was popularized by the `R` package `EpiEstim` (Cori et al., 2013), which due to its broad applicability, ease of use, and thorough documentation has become a standard tool (Nash et al., 2022). It has also been integrated in numerous refined methods (e.g., Alvarez et al. 2021, Brizzi et al. 2022, Green et al. 2022, Nash et al. 2023). In this note we are concerned with the uncertainty surrounding estimated reproduction numbers based on the renewal equation and a Poisson assumption for the observed incidence. We demonstrate that the associated standard errors are underestimated in the presence of overdispersion, a common feature of infectious disease surveillance data. Especially when incidence is high, uncertainty intervals based on Poisson assumptions get vanishingly narrow, overstating the certainty of the inference. While this has been pointed out empirically (Brockhaus et al., 2023; Goldstein et al., 2023), we provide a concise summary of relevant statistical properties. We moreover contrast several methods to account for overdispersion, which turn out to impact not only estimated standard errors, but also point estimates. We illustrate our arguments in case studies on COVID-19, Ebola and pandemic influenza.

2 The Poisson renewal equation model

The *instantaneous* reproduction number is often denoted by R_e or R_t, but as in `EpiEstim` it is assumed constant over an estimation window of T days, we simply write R. Daily or weekly incidence counts $X_t, t = 1, \dots, T$ are linked to R via the renewal equation and a Poisson assumption,

$$X_t \mid \text{past} \sim \text{Pois}(\mu_t), \quad \mu_t = R \times \Lambda_t, \quad \Lambda_t = \sum_{d=1}^{D} w_d \times X_{t-d}. \tag{1}$$

Here, μ_t is the expectation of the Poisson distribution. The discrete-time serial interval distribution, expressed by probabilities $w_1, \ldots, w_D$, is considered known. The Poisson assumption is chosen as it makes Bayesian inference straightforward without the need for Monte Carlo methods (Cori et al., 2013).

Equation (1) can be read as a Poisson regression with an identity link, no intercept and Λ_t as the sole covariate (Goldstein et al., 2023; Brockhaus et al., 2023). The properties of the resulting estimates can hence be studied in the framework of generalized linear regression (Fahrmeir et al., 2009). This is straightforward for maximum likelihood (ML) estimates; the results also translate to the Bayesian approach by Cori et al. (2013), which behaves similarly unless incidence is very low (Appendix C.1.2). The ML estimator takes the intuitive form

$$\hat{R}_{\mathrm{Po}} = \frac{\sum_{t=1}^{T} X_t}{\sum_{t=1}^{T} \Lambda_t}, \tag{2}$$

where "Po" indicates the underlying distributional assumption. The estimated standard error

$$\widehat{\mathrm{se}}(\hat{R}_{\mathrm{Po}}) = \frac{\hat{R}_{\mathrm{Po}}}{\sqrt{\sum_{t=1}^{T} X_t}} \tag{3}$$

depends on $\hat{R}_{\mathrm{Po}}$ and the total incidence, but not the "noisiness" of the time series. This is a consequence of the equidispersion property of the Poisson distribution, $\mathrm{Var}(X_t \mid \text{past}) = \sigma_t^2 = \mu_t$. Derivations of equations (2) and (3) are provided in Supplement C.1.1, where we also discuss challenges due to the dynamic nature of model (1).

3 Overdispersion and how to account for it

3.1 Impact on the Poisson estimator

In practice, infectious disease incidence data usually display conditional overdispersion rather than equidispersion, i.e., $\sigma_t^2 > \mu_t$. The Poisson assumption will then cause our inference for R to be overconfident; see Supplementary Section A for an intuitive illustration. Well-known theory (Fahrmeir et al., 2009, Sec. 5.5) implies that expression (3) underestimates the standard error of $\hat{R}_{\mathrm{Po}}$ by a factor of $\sqrt{\phi}$, with

$$\phi = \frac{\sum_{t=1}^{T} \sigma_t^2}{\sum_{t=1}^{T} \mu_t}. \tag{4}$$

The *variance inflation factor* ϕ depends on the relationship between μ_t and σ_t^2. It can be large in practice, meaning that standard errors estimated under the Poisson assumption may be considerably too small.

In the following we will outline three options to account for overdispersion, all of which have been applied in the literature. While this is typically motivated by the desire to estimate overdispersion parameters (Ho et al., 2023) and general model realism (Gressani et al., 2022), our focus is on uncertainty estimation.

3.2 Quasi-Poisson regression

Quasi-Poisson regression, used by Goldstein et al. (2023), serves to correct the uncertainty quantification for the Poisson model by estimating the variance inflation factor ϕ from equation (4). Assuming a linear mean-variance relationship, it can be estimated from the residuals of the Poisson model (see Supplement C.2). The corrected standard error is then obtained by multiplying expression (3) with $\sqrt{\hat{\phi}}$. We note that quasi-Poisson regression is no full probability model, making it unsuitable e.g., for sampling future trajectories.

3.3 Negative binomial regression

The most common full probability model to generalize the Poisson assumption is the negative binomial distribution. We here parameterize it by its mean and an overdispersion parameter. The latter quantifies the overdispersion relative to a Poisson distribution (which loosely speaking corresponds to an overdispersion parameter of zero). The overdispersion parameter can be handled in two ways (Cameron and Trivedi, 1998, p.73), which we call *NegBin-L* and *NegBin-Q*, as they imply *linear* and *quadratic* mean variance relationships. Estimates and standard errors are obtained using standard regression methodology and account for overdispersion.

NegBin-L. In The NegBin-L renewal equation

$$X_t \mid \text{past} \sim \text{NegBin}(\text{mean} = \mu_t, \text{overdisp} = \xi/\mu_t), \tag{5}$$

the overdispersion parameter is coupled to the mean. This implies a linear mean-variance relationship $\sigma_t^2 = \xi \times \mu_t$ as in the quasi-Poisson model, and can be motivated by overdispersed individual offspring distributions (Ho et al., 2023). While some differences between NegBin-L and quasi-Poisson regression exist (Hilbe, 2011), in our applications we found similar behavior with almost identical point estimates.

NegBin-Q. We therefore focus primarily on NegBin-Q, defined as

$$X_t \mid \text{past} \sim \text{NegBin}(\text{mean} = \mu_t, \text{overdisp} = \psi), \tag{6}$$

and implying a quadratic mean-variance relationship $\sigma_t^2 = \mu_t + \psi \times \mu_t^2$. This version, employed e.g., by Held et al. (2005) and Brockhaus et al. (2023), can be motivated by variation of R over the estimation window. Despite the mean structure $\mu_t = R \times \Lambda_t$ mirroring the Poisson version, the ML estimator $\hat{R}_{\text{NBQ}}$ is not identical to $\hat{R}_{\text{Po}}$ from (2). No closed expression exists, but if overdispersion is sufficiently strong, the approximation

$$\hat{R}_{\text{NBQ}} \approx \frac{1}{T} \times \sum_{t=1}^{T} \frac{X_t}{\Lambda_t} \tag{7}$$

holds. Unlike in (2), the ratios $X_1/\Lambda_1, \ldots, X_T/\Lambda_T$ contribute equally to the estimate. This is because they are all equally informative under a quadratic mean variance relationship, irrespective of the magnitude of Λ_t and X_t (see discussion in Ver Hoef and Boveng 2007). This also explains why the approximate standard error

$$\widehat{\text{se}}(\hat{R}_{\text{NBQ}}) \approx \hat{R}_{\text{NBQ}} \times \sqrt{\hat{\psi}/T} \tag{8}$$

features the number of observations T rather than the total incidence as in equation (3). It moreover depends on the estimated overdispersion parameter $\hat{\psi}$, which reflects the noisiness of the time series and is computed from the model residuals (Supplement C.3). It generally holds that $\widehat{\text{se}}(\hat{R}_{\text{NBQ}}) \geq \widehat{\text{se}}(\hat{R}_{\text{Po}})$, i.e., the estimated standard error is at least equal to its Poisson counterpart. The approximations (7) and (8) are assessed empirically in Supplementary Figure S4.

Quasi-Poisson regression generally has certain advantages over the negative binomial versions if the mean-variance relationship is mis-specified (Wooldridge, 2010). In our setting, however, both (2) and (7) remain unbiased and consistent in this case, and only their efficiency is reduced. How strongly the two estimators differ is an empirical question addressed in the next section.

4 Illustration for three outbreaks

We illustrate the impact of the distributional choice in applications to three data sets covering different pathogens and locations. These are based on previous applications of `EpiEstim` summarized in Table 1. We use shifted gamma serial interval distributions with means and standard deviations retrieved from the respective references. For the influenza and COVID-19 examples, we also use the corresponding estimation window length. For the Ebola data, we present a variation of the analysis from Green et al. (2022). We aggregate the daily data to weekly totals, rescale the serial interval distribution and employ an estimation window of 4 weeks. This illustrates our point in a weekly setting for a disease with a longer serial interval.

The Poisson and quasi-Poisson models were fitted using the `stats::glm` function in `R` (R Core Team, 2013), while `gamlss:gamlss` (Stasinopoulos and Rigby, 2007) was employed for the negative binomial models. In all cases we show Wald confidence intervals.

Figure 1 shows the three incidence time series and R estimates based on the different distributional assumptions. The second row compares the Poisson and quasi-Poisson models. While the point estimates agree by definition, the Poisson version yields much narrower 95% confidence intervals. For influenza and COVID-19, they are visually indiscernible. For Ebola they are wider, as suggested by equation (3) for lower incidence counts.

Table 1: Summary of the case studies.

Reference	Disease	Location	Year	Resolution	Serial interval mean (sd)	Estimation window width T
Nash et al. (2023), using data from Riley et al. (2016)	Pandemic influenza	US (military personnel)	2009–2010	Daily	3.6 (1.6)	7 days
Richter et al. (2022)	COVID-19	Austria	2021–2022	Daily	3.37 (1.83)	13 days
Green et al. (2022)	Ebola	Guinea	2014–2015	Weekly	2.19 (1.33)	4 weeks

Nonetheless, they rarely include $R = 1.0$, implying high confidence about epidemic growth or decline. The intervals from the quasi-Poisson model, on the other hand, often contain values above and below $R = 1.0$.

The third row overlays results from the NegBin-L model (which behaves similarly to the quasi-Poisson model; Supplementary Figure S3), and the NegBin-Q model. The confidence intervals are comparable for Ebola and COVID-19. For influenza, the NegBin-Q intervals are wider, and there are considerable differences in the point estimates. These are due to the different weighting of the observations in estimators (2) and (7), in conjunction with weekday effects. We discuss this in Supplementary Section A.

Ignoring weekday effects can be argued to yield a mis-specified model (see also Nash et al. 2023). The resulting unexplained variability is absorbed by the overdispersion parameters, leading to confidence intervals which may become overly wide. Despite these issues, we kept the influenza example to illustrate the difference between NegBin-L and NegBin-Q, as well as the difficulty of fitting time-homogeneous models to daily data.

5 Discussion

We argued that unjustified Poisson assumptions lead to overconfident estimation of reproductive numbers and discussed options to account for overdispersion. The quasi-Poisson and NegBin-L versions agreed well in terms of point estimates and uncertainty intervals. The NegBin-Q approach yielded somewhat different results, with major discrepancies potentially indicating general model mis-specification issues. In applications to influenza, COVID-19 and Ebola, all three extensions yielded considerably widened uncertainty intervals relative to the Poisson model, which included the threshold $R = 1.0$ considerably more often. This can substantially change the interpretation of results.

For reliable uncertainty quantification we therefore strongly recommend to account for overdispersion. While the conjugate prior approach from Cori et al. (2013) is difficult to extend, standard regression functions can be adapted. We are moreover aware that an MCMC-based extension of `EpiEstim` to account for overdispersion in a Bayesian framework is under development, and we recommend it to future users.

Alternative tools like `EpiNow2` (Abbott et al., 2020) already contain this feature. `EpiNow2` also avoids the need for estimation windows and accounts for delays between infection events and reporting. This makes estimates more aligned with the infection dynamics. Due to its increased complexity `EpiNow2`, requires more tuning and computational effort, which may present an obstacle to some users.

We conclude by discussing some limitations of our arguments. By using off-the-shelf generalized linear models, we conflate different sources of overdispersion, including superspreading and underreporting. It is thus difficult to assign any mechanistic interpretation to the estimated overdispersion parameters, and they cannot serve to derive implications on disease emergence and control (Lloyd-Smith et al., 2005). Estimating the overdispersion parameter along with R from short time windows may lead to instabilities. Especially for negative binomial regression, it is moreover known that small-sample biases can occur (Kenne Pagui et al., 2022), with a tendency to underestimate standard errors. Lastly, we ignored numerous other complexities including unobserved initial generations (Brizzi et al., 2022) and changes in testing volumes (Goldstein et al., 2023).

A promising regression-based approach replacing windowing by splines has recently been proposed by Nouvellet (2025), and can remedy some weaknesses. It estimates the temporal variation in R flexibly, accounting for covariates like weekday effects. It currently only supports the Poisson distribution, but extensions could easily be implemented, potentially with time-varying dispersion parameters (Stasinopoulos and Rigby, 2007).

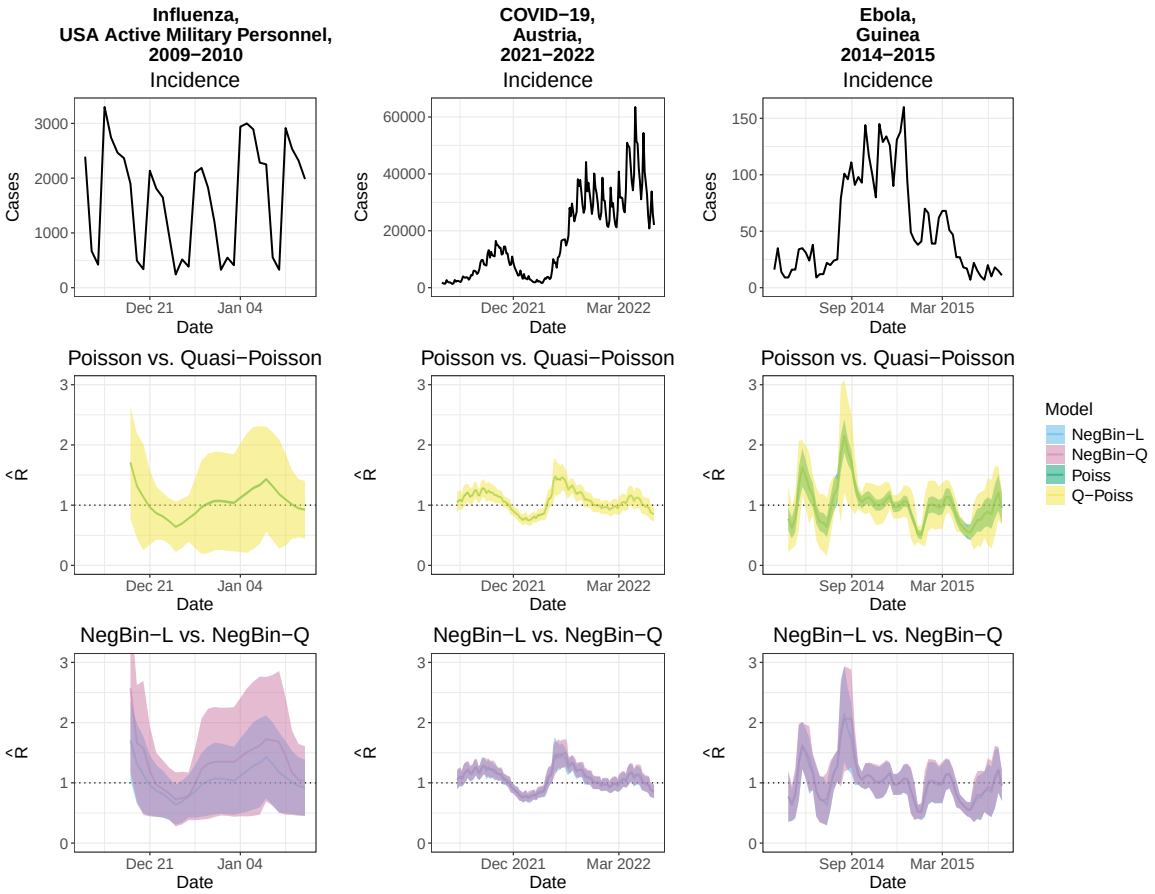

Figure 1: Incidence and estimated effective reproductive numbers for three outbreaks: influenza among the active military personnel in the USA, 2009–2010 (left column); COVID-19 in Austria, 2021–2022 (middle); and Ebola in Guinea, 2014–2015 (right). The first row shows case incidences, which are in daily resolution for influenza and COVID-19 and in weekly resolution for Ebola. The second row shows a comparison of the R estimates from the Poisson and quasi-Poisson models. All estimates are aligned with the last day of the respective estimation window. The third row shows a comparison of the R estimates from the NegBin-Q and NegBin-L models. An additional plot overlaying the quasi-Poisson and NegBin-L versions as well as the estimated overdispersion parameters are available in Supplementary Figures S2 and S3.

Reproducibility. The repository `https://github.com/barbora-sobolova/epiestim_overdispesion/` contains code and data to reproduce our analyses.

Funding statement. Barbora Němcová and Johannes Bracher were supported by the German Research Foundation (DFG), project 512483310. Barbora Němcová was moreover supported by the Helmholtz Association under the joint research school HIDSS4Health – Helmholtz Information and Data Science School for Health. Isaac H. Goldstein was supported by a Stanford Center for Computational, Evolutionary and Human Genomics Fellowship. Volodymyr M. Minin acknowledges support from NIH grant R01AI170204.

References

S Abbott, J Hellewell, RN Thompson, K Sherratt, HP Gibbs, NI Bosse, JD Munday, S Meakin, EL Doughty, JY Chun, YWD Chan, F Finger, P Campbell, A Endo, CAB Pearson, A Gimma, T Russell, S Flasche, AJ Kucharski, RM Eggo, and S Funk. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Research*, 5(112), 2020.

L Alvarez, M Colom, JD Morel, and Jean-Michel Morel. Computing the daily reproduction number of COVID-19 by inverting the renewal equation using a variational technique. *Proceedings of the National Academy of Sciences*, 118(50):e2105112118, 2021.

A Brizzi, M O'Driscoll, and I Dorigatti. Refining reproduction number estimates to account for unobserved generations of infection in emerging epidemics. *Clinical Infectious Diseases*, 75(1):e114–e121, 02 2022. ISSN 1058-4838.

EK Brockhaus, D Wolffram, T Stadler, M Osthege, T Mitra, JM Littek, E Krymova, AJ Klesen, JS Huisman, S Heyder, LM Helleckes, M an der Heiden, S Funk, S Abbott, and J Bracher. Why are different estimates of the effective reproductive number so different? A case study on COVID-19 in Germany. *PLOS Computational Biology*, 19(11):1–27, 11 2023.

AC Cameron and PK Trivedi. *Regression Analysis of Count Data*. Cambridge University Press, 1998.

A Cori, NM Ferguson, C Fraser, and S Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9):1505–1512, 2013.

L Fahrmeir, T Kneib, and S Lang. *Regression*. Springer, 2009.

IH Goldstein, J Wakefield, and VM Minin. Incorporating testing volume into estimation of effective reproduction number dynamics. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 187(2):436–453, 12 2023.

KM Gostic, L McGough, EB Baskerville, S Abbott, K Joshi, C Tedijanto, R Kahn, R Niehus, JA Hay, PM De Salazar, J Hellewell, S Meakin, JD Munday, NI Bosse, K Sherrat, RN Thompson, LF White, JS Huisman, J Scire, S Bonhoeffer, T Stadler, J Wallinga, S Funk, M Lipsitch, and S Cobey. Practical considerations for measuring the effective reproductive number, R_t. *PLOS Computational Biology*, 16(12):1–21, 12 2020.

WD Green, NM Ferguson, and A Cori. Inferring the reproduction number using the renewal equation in heterogeneous epidemics. *Journal of The Royal Society Interface*, 19(188):20210429, 2022.

O Gressani, J Wallinga, C L. Althaus, N Hens, and C Faes. EpiLPS: A fast and flexible Bayesian tool for estimation of the time-varying reproduction number. *PLOS Computational Biology*, 18(10):1–27, 10 2022.

L Held, M Höhle, and M Hofmann. A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, 5(3):187–199, 2005.

JM Hilbe. *Negative Binomial Regression*. Cambridge University Press, 2nd edition, 2011.

F Ho, KV Parag, EHY Adam, DC.and Lau, BJ Cowling, and TK Tsang. Accounting for the potential of overdispersion in estimation of the time-varying reproduction number. *Epidemiology*, 34(2):201–205, 2023.

EC Kenne Pagui, A Salvan, and N Sartori. Improved estimation in negative binomial regression. *Statistics in Medicine*, 41(13): 2403–2416, 2022.

JO Lloyd-Smith, SJ Schreiber, PE Kopp, and WM Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438:355–359, 2005.

RK Nash, P Nouvellet, and A Cori. Real-time estimation of the epidemic reproduction number: Scoping review of the applications and challenges. *PLOS Digital Health*, 1(6):1–17, 2022.

RK Nash, S Bhatt, A Cori, and P Nouvellet. Estimating the epidemic reproduction number from temporally aggregated incidence data: A statistical modelling approach and software tool. *PLOS Computational Biology*, 19(8):1–14, 08 2023.

P Nouvellet. Rt.GLM: Unifying estimation of the time-varying reproduction number, R_t, under the Generalised Linear and Additive Models, 2025. URL `https://www.medrxiv.org/content/early/2025/06/24/2025.06.24.25330176`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2013.

L Richter, A Chalupka, D Schmid, A Chakeri, S Maritschnik, S Pfeiffer, and E Stadlober. Epidemiologische Parameter des COVID19 Ausbruchs - Update 01.04.2022, Österreich, 2020–2022. Technical report, Austrian Agency for Health and Food Safety (AGES), 2022. `https://www.ages.at/fileadmin/Corona/Epidemiologische-Parameter/updates/Update_Epidemiologische_Parameter_des_COVID19_Ausbruchs_2022-04-01.pdf`.

P Riley, AA Cost, and S Riley. Intra-Weekly Variations of Influenza-Like Illness in Military Populations. *Military Medicine*, 181 (4):364–368, 2016.

DM Stasinopoulos and RA Rigby. Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23(7):1–46, 2007.

JM Ver Hoef and PL Boveng. Quasi-Poisson vs negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772, 2007.

JM Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, 2010.

A Intuitive illustration of maximum likelihood estimation

To strengthen intuition we provide a graphical illustration of likelihood inference under different models. The top panel of Figure S1 shows a 7-day window (3–9 January 2010) taken from the US influenza data discussed in Section 4. The incidence is overlaid with Λ_t from equation (1). The remaining panels show the log-likelihood contributions of the 7 individual observations (gray) and the overall log-likelihood functions (black) for the Poisson and negative binomial models, all on a relative scale. For the negative binomial versions, the overdispersion parameters are fixed at the maximum likelihood estimates.

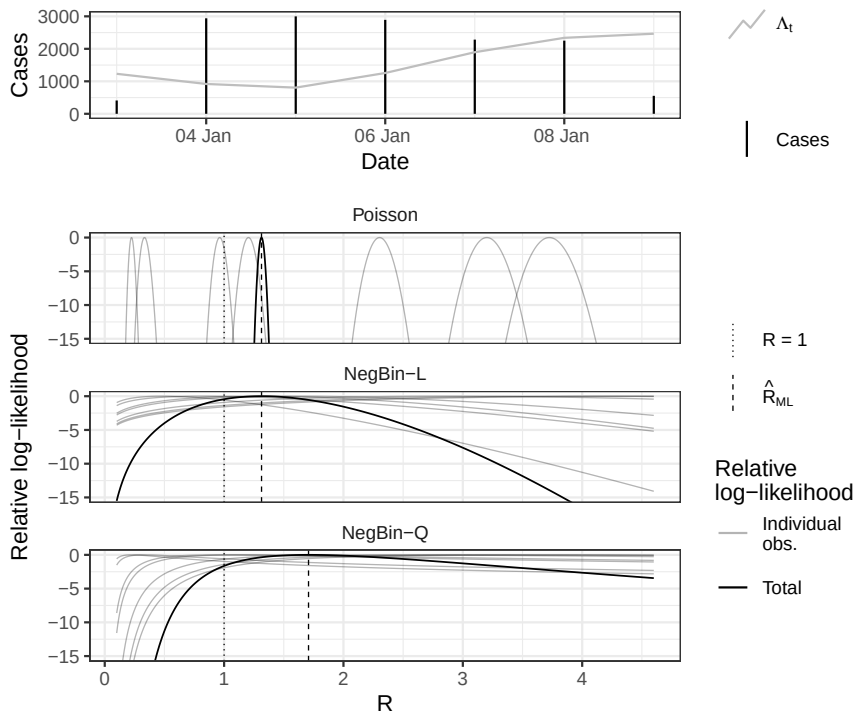

Figure S1: Illustration of maximum likelihood estimation in the Poisson, NegBin-L and NegBin-Q models. In the three bottom panels, light grey lines show the individual likelihood contributions, while black lines show the overall likelihood functions (i.e., the sums of the respective contributions). All of these are shown on a relative scale such that their peak value is zero. Maximum likelihood estimates are highlighted by vertical dashed lines, while the threshold value $R = 1.0$ is marked by a dotted line.

Due to the equidispersion assumption and high observed counts, the Poisson log-likelihood contribution functions are narrow, and the overall log-likelihood function exhibits a sharp peak. The curves for the negative binomial models are flatter as the model adapts to the higher degree of dispersion. Since $\widehat{\text{se}}(\hat{R}_\cdot)$ reflects the inverse curvature of the log-likelihood function at its peak (see Supplement C.1.1), it is clear that the Poisson model issues much more confident estimates. Notably, the Poisson likelihood clearly favors a value of $R > 1$, while for the negative binomial models, the relative likelihood is rather high for $R = 1$.

We moreover notice a shift in the point estimates from different models (dashed vertical lines). While they are almost identical for the Poisson and NegBin-L models ($\hat{R}_{\text{Po}} \approx \hat{R}_{\text{NBL}} \approx 1.31$), we obtain a higher value for NegBin-Q ($\hat{R}_{\text{NBQ}} \approx 1.71$). This is a result of the particular alignment of Λ_t and X_t, caused by weekday effects, and the different weights the $T = 7$ observations receive in the two schemes. To understand this, note that both estimators are weighted averages of the daily ratios $X_t/\Lambda_t, t = 1, \ldots, T$. For the NegBin-Q model, this average is unweighted as in equation (7). For the Poisson model, equation (2) can be written as

$$\hat{R}_{\text{Po}} \approx \frac{1}{T} \times \sum_{t=1}^{T} \frac{\Lambda_t}{\sum_{t=1}^{T} \Lambda_t} \times \frac{X_t}{\Lambda_t},$$

and also applies in good approximation for the NegBin-L model (McCullagh and Nelder, 1989, p. 199). Here, the ratios X_t/Λ_t thus enter with weights proportional to Λ_t. As seen in the top panel of Figure S1, the highest values of X_t/Λ_t occur for the lowest values of Λ_t. These differences in weighting thus explain why the NegBin-Q estimate is considerably higher than its counterparts from the Poisson and NegBin-L models.

B Details on case studies

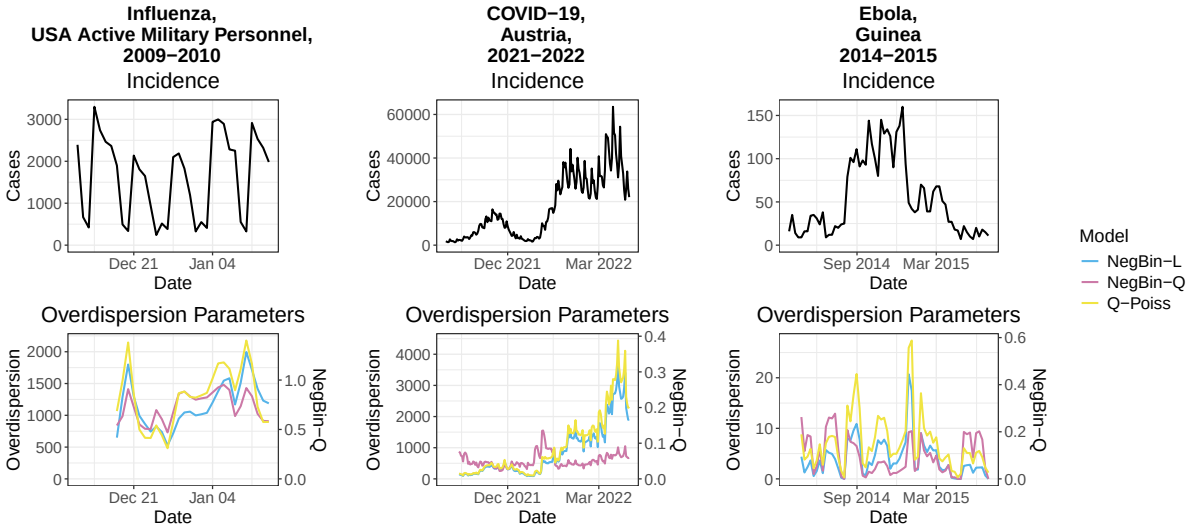

Figure S2: Estimated overdispersion parameters in the three case studies from Figure 1. The left vertical axis refers to the quasi-Poisson and NegBin-L models, while the right vertical axis refers to the NegBin-Q model.

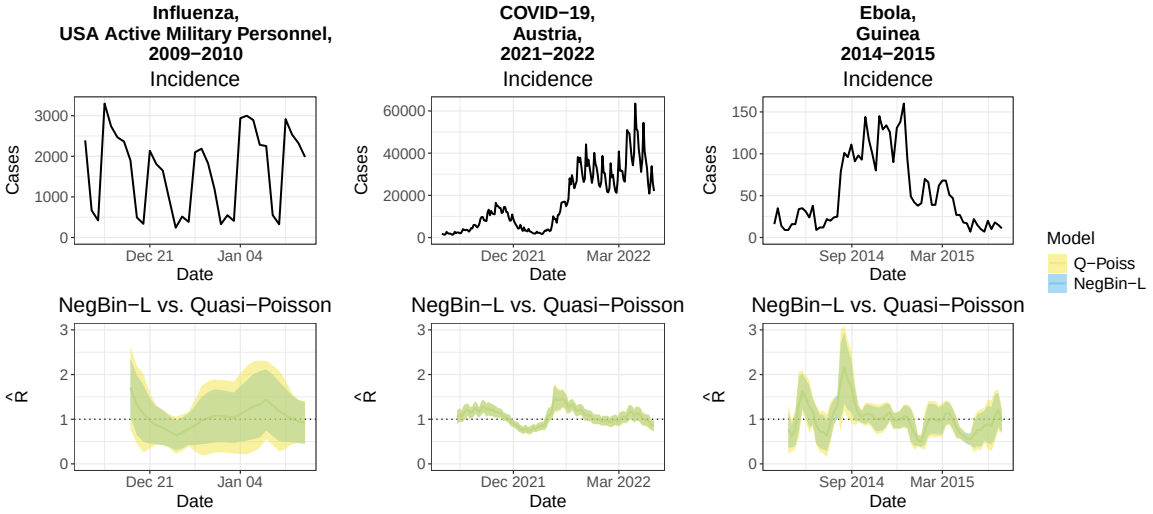

Figure S3: Comparison of R estimates from the quasi-Poisson and NegBin-L models in the case studies from Figure 1.

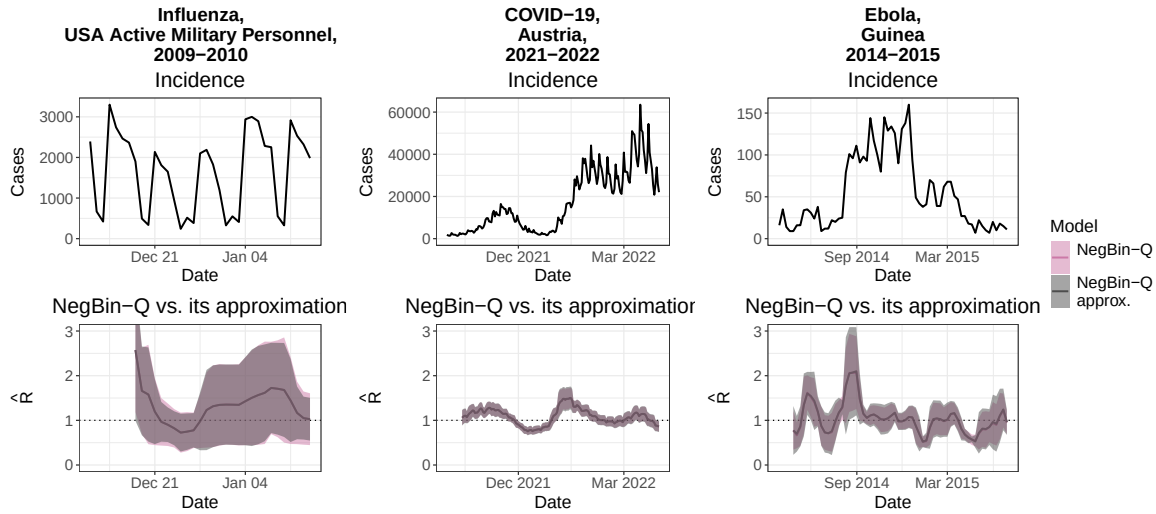

Figure S4: Comparison of R estimates and associated Wald confidence intervals from the NegBin-Q model obtained using the `gamlss` package and approximate formulas (7), (8) and (S7). Agreement is overall close, with minor discrepancies for the Ebola case study.

C Mathematical details

C.1 Details on the Poisson model, equation (1) in the main text

C.1.1 Maximum likelihood estimation (equations (2) and (3) in the main text)

Here we derive the formulas (2) and (3) using standard maximum likelihood theory. In the following we denote the vector of observed values in the estimation window by $\mathbf{X} = (X_1, \ldots, X_T)$ and substitute $\mu_t = R\Lambda_t$. Due to the D-th order Markov structure of the Poisson renewal equation model (1), the log-likelihood is given by a sum of T Poisson log-likelihoods,

$$l(R; \mathbf{X}) = \sum_{t=1}^{T} \log\left\{\frac{(R\Lambda_t)^{X_t}}{X_t!} \times \exp(-R\Lambda_t)\right\} = \sum_{t=1}^{T} \left\{X_t \log(R) - R\Lambda_t + X_t \log(\Lambda_t) - \log(X_t!)\right\}. \tag{S1}$$

Taking the derivative with respect to R we obtain the score function

$$\frac{\partial l(R; \mathbf{X})}{\partial R} = \frac{\partial}{\partial R}\left\{\sum_{t=1}^{T} \left\{X_t \log(R) - R\Lambda_t + X_t \log(\Lambda_t) - \log(X_t!)\right\}\right\} = \sum_{t=1}^{T} \left\{\frac{X_t}{R} - \Lambda_t\right\}.$$

Next we set the score function equal zero and solve for R,

$$\sum_{t=1}^{T} \left\{\frac{X_t}{R} - \Lambda_t\right\} \equiv 0. \quad \Rightarrow \quad \hat{R}_{\mathrm{Po}} = \frac{\sum_{t=1}^{T} X_t}{\sum_{t=1}^{T} \Lambda_t},$$

which results in the maximum likelihood estimator. In order to find an estimate of its variance, we compute the observed Fisher information, which involves taking the second derivative of the log-likelihood function,

$$\mathcal{J}(R) = -\frac{\partial^2 l(R; \mathbf{X})}{\partial R^2} = \sum_{t=1}^{T} \frac{X_t}{R^2}. \tag{S2}$$

We then obtain the formula of the standard error of $\hat{R}_{\mathrm{Po}}$ by inverting the observed Fisher information at the maximum likelihood estimate and taking the square root,

$$\widehat{\mathrm{se}}(\hat{R}_{\mathrm{Po}}) = \sqrt{\left(\sum_{t=1}^{T} \frac{X_t}{\hat{R}_{\mathrm{Po}}^2}\right)^{-1}} = \frac{\hat{R}_{\mathrm{Po}}}{\sqrt{\sum_{t=1}^{T} X_t}}.$$

We note that unlike in a classic generalized linear regression model (GLM), the "covariate" Λ_t is itself random, and all $\Lambda_t, X_t, t = 1, \ldots, T$ are dependent. This invalidates some of the basic assumptions underlying maximum likelihood theory for GLMs. Due to the Markov structure of the process (1), however, the log-likelihood function (S1) and the observed Fisher information (S2) are nonetheless correct, and the `glm` implementation in R can be used to evaluate them. It is non-trivial to show that maximum likelihood estimators preserve their usual properties in stochastic process models like (1), as relevant mixing properties need to be established. This has been done in related stationary models from the INGARCH family (see e.g., Fokianos et al. 2009), but in practice maximum likelihood estimators and associated standard errors are widely used and have been found to be reliable even when their properties have not been formally demonstrated (see e.g., Paul et al. 2008). For model (1), an additional challenge arises from the fact that the process is non-stationary and potentially explosive. Mixing properties in related settings have been explored by Doukhan et al. (2022), but we consider these aspects outside the scope of the present applied article.

C.1.2 Comparison to Bayesian estimation

In the `EpiEstim` package, Bayesian estimation of R in the model (1) is based on a conjugate gamma prior distribution. Specifying the prior distribution as

$$R \sim \mathrm{Gamma}(\mathrm{shape} = a, \mathrm{scale} = b),$$

the posterior distribution given the data $\mathbf{X} = (X_1, \ldots, X_T)$ is likewise a gamma distribution. Specifically, we obtain (Cori et al., 2013, Supplementary Material, page 3)

$$R \mid \mathbf{X} \sim \mathrm{Gamma}\left(\mathrm{shape} = a + \sum_{t=1}^{T} X_t, \mathrm{scale} = \frac{1}{\frac{1}{b} + \sum_{t=1}^{T} \Lambda_t}\right).$$

The first two posterior moments of R are consequently given by

$$\mathbb{E}(R \mid \mathbf{X}) = \frac{a + \sum_{t=1}^{T} X_t}{\frac{1}{b} + \sum_{t=1}^{T} \Lambda_t},$$
$$\mathrm{sd}(R \mid \mathbf{X}) = \frac{\mathbb{E}(R \mid \mathbf{X})^2}{\sqrt{a + \sum_{t=1}^{T} X_t}}.$$

This strongly resembles the maximum likelihood estimator $\hat{R}_{\mathrm{Po}}$ from (2) and its estimated standard deviation given in equation (3). Indeed, the Bayesian scheme can be seen as a regularized maximum likelihood estimate resulting from the addition of a pseudo observation with $\Lambda_0 = 1/b$ and $X_0 = a$. The default values in `EpiEstim` are $a = 1, b = 1/5$, meaning that for moderately high incidence values they have little influence on the result.

C.2 Details on the quasi-Poisson model

The most common way to estimate the overdispersion parameter $\hat{\phi}$, implemented in the `R` function `glm(..., family = quasipoisson)`, is

$$\hat{\phi} = \frac{1}{T-1} \sum_{t=1}^{T} \frac{(X_t - \hat{R}_{\mathrm{Po}}\Lambda_t)^2}{\hat{R}_{\mathrm{Po}}\Lambda_t}. \tag{S3}$$

This is simply the empirical variance of the Pearson residuals in the Poisson model. The standard errors in expression (3) are then corrected to

$$\widehat{\text{se}}(\hat{R}_{\text{Po}}) = \sqrt{\hat{\phi}} \times \frac{\hat{R}^2_{\text{Po}}}{\sqrt{\sum_{t=1}^{T} X_t}}. \tag{S4}$$

C.3 Details on the NegBin-Q model, equation (6) in the main text

C.3.1 Maximum likelihood estimation, equations (7) and (8) in the main text

The negative binomial distribution with mean μ and dispersion parameter ψ has probability mass function

$$f(x) = \frac{\Gamma(1/\psi + x)}{x!\Gamma(1/\psi)} \left(\frac{1/\psi}{1/\psi + \mu}\right)^{1/\psi} \left(\frac{\mu}{1/\psi + \mu}\right)^{x}.$$

Following the same logic as in Appendix C.1.1, the log-likelihood function of model (6) is

$$\begin{aligned} l(R; \mathbf{X}) &= \sum_{t=1}^{T} \log[f(X_t|R, \Lambda_t)] \\ &= \sum_{t=1}^{T} \Big\{ \log\left[\Gamma(1/\psi + X_t)\right] - \log\left[X_t!\Gamma(1/\psi)\right] + (1/\psi) \times \log(1/\psi) \\ &\qquad - (X_t + 1/\psi) \times \log(1/\psi + R\Lambda_t) + X_t \log(R\Lambda_t) \Big\}. \end{aligned}$$

Now we differentiate with respect to R to obtain the score equation

$$\begin{aligned} \frac{\partial l(R; \mathbf{X})}{\partial R} &= \sum_{t=1}^{T} \left\{ (X_t + 1/\psi) \times \frac{-\Lambda_t}{R\Lambda_t + 1/\psi} + \frac{X_t}{R} \right\} \\ &= \ldots = \\ &= \sum_{t=1}^{T} \frac{X_t - R\Lambda_t}{\psi R(R\Lambda_t + 1/\psi)} \equiv 0, \end{aligned}$$

which is equivalent to

$$\sum_{t=1}^{T} \frac{X_t - R\Lambda_t}{R\Lambda_t + 1/\psi} \equiv 0. \tag{S5}$$

We now assume that there is substantial overdispersion, which we formalize as

$$1/\psi \ll \mu_t = R\Lambda_t \quad \text{for} \quad t = 1, \ldots, T,$$

implying that

$$\sigma_t^2 \approx \psi \mu_t^2.$$

Equation (S5) then simplifies to the estimator from equation (7) as

$$\begin{aligned} & \sum_{t=1}^{T} \left(\frac{X_t}{R\Lambda_t} - 1 \right) \approx 0 \\ \Leftrightarrow \quad & \frac{1}{R} \times \sum_{t=1}^{T} \frac{X_t}{\Lambda_t} \approx T \\ \Rightarrow \quad & \hat{R}_{\text{NBQ}} \approx \frac{1}{T} \times \sum_{t=1}^{T} \frac{X_t}{\Lambda_t}. \end{aligned} \tag{S6}$$

Next we again compute the observed Fisher information with respect to R,

$$\begin{aligned}\mathcal{J}(R) = -\left(\frac{\partial^2 l(R;\mathbf{X})}{\partial R^2}\right) &= -\frac{\partial}{\partial R}\left(\sum_{t=1}^{T}\frac{X_t - R\Lambda_t}{\psi R(R\Lambda_t + 1/\psi)}\right)\\ &= \sum_{t=1}^{T}\frac{\Lambda_t\psi R(R\Lambda_t + 1/\psi) + (X_t - R\Lambda_t)\psi(2R\Lambda_t + 1/\psi)}{\psi^2 R^2(R\Lambda_t + 1/\psi)^2}\\ &= \ldots = \sum_{t=1}^{T}\frac{-R^2\Lambda_t^2 + X_t(2R\Lambda_t + 1/\psi)}{\psi R^2(R\Lambda_t + 1/\psi)^2},\end{aligned}$$

Now again assuming that $1/\psi \ll \mu_t = R\Lambda_t$ this simplifies to

$$\begin{aligned}\mathcal{J}(R) &\approx \sum_{t=1}^{T}\frac{-R^2\Lambda_t^2 + 2X_tR\Lambda_t}{\psi R^4\Lambda_t^2},\\ &= \frac{1}{\psi R^2}\times\left[\sum_{t=1}^{T}-1 \quad + \quad \frac{2}{R}\times\sum_{t=1}^{T}\frac{X_t}{\Lambda_t}\right]\end{aligned}$$

Now evaluating this at the (approximate) maximum likelihood estimator (S6) we obtain

$$\begin{aligned}\mathcal{J}(\hat{R}_{\text{NBQ}}) &= \frac{1}{\psi\hat{R}_{\text{NBQ}}^2}\times\left[-T \quad + \quad \underbrace{\frac{2}{\sum_{t=1}^{T}\frac{X_t}{\Lambda_t}}\times\sum_{t=1}^{T}\frac{X_t}{\Lambda_t}}_{=2T}\right]\\ &= \frac{T}{\psi\hat{R}_{\text{NBQ}}^2}.\end{aligned}$$

As before, the standard error of $\hat{R}_{\text{NBQ}}$ then results as

$$\widehat{\text{se}}(\hat{R}_{\text{NBQ}}) = \sqrt{\frac{1}{\mathcal{J}(\hat{R}_{\text{NBQ}})}} = \hat{R}_{\text{NBQ}}\times\sqrt{\frac{\psi}{T}}.$$

C.3.2 Estimation of the overdispersion parameter ψ

The overdispersion parameter ψ can be estimated as

$$\hat{\psi} = \frac{1}{T-1}\sum_{t=1}^{T}\frac{(X_t - \hat{R}_{\text{NBQ}}\Lambda_t)^2}{(\hat{R}_{\text{NBQ}}\Lambda_t)^2}, \tag{S7}$$

which resembles equation (S4), but is different due to the square in the denominator. This estimator, which to our knowledge has not been proposed previously, is motivated by the fact that under model (6) we have

$$\sigma_t^2 = \text{Var}(X_t \mid \mu_t) = \mu_t + \psi\mu_t^2.$$

Under our additional assumption $\psi \gg 1/\mu_t$ this simplifies to

$$\begin{aligned}&\text{Var}(X_t \mid \mu_t) = \psi\mu_t^2\\ \Leftrightarrow\ &\text{Var}\left(\underbrace{\frac{X_t - \mu_t}{\mu_t}}_{\text{denote this by } r_t} \;\middle|\; \mu_t\right) = \psi.\end{aligned}$$

Now note that we moreover have

$$\mathbb{E}\left(r_t \mid \mu_t\right) = 0.$$

Due to the Markov property of $\{X_t\}$, the $r_1, \ldots, r_T$ are hence independent random variables with mean zero and identical variance ψ. The empirical variance (S7) of $r_1, \ldots, r_T$ is thus a natural estimator of ψ.

Additional references

P Doukhan, A Leucht, and MH Neumann. Mixing properties of non-stationary INGARCH$(1,1)$ processes. *Bernoulli*, 28(1):663 – 688, 2022.

K Fokianos, A Rahbek, and D Tjøstheim. Poisson autoregression. *Journal of the American Statistical Association*, 104(488):1430–1439, 2009.

P McCullagh and John A Nelder. *Generalized Linear Models, Second Edition*. CRC Press, 1989.

M. Paul, L. Held, and A. M. Toschke. Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine*, 27(29):6250–6267, 2008.