Danon, L., & Brooks Pollock, E. (2016). The need for data science in epidemic modelling: Comment on: "Mathematical models to characterize early epidemic growth: A Review" by Gerardo Chowell et al. *Physics of Life Reviews*, *18*, 102–104. https://doi.org/10.1016/j.plrev.2016.08.011

The need for data science in epidemic modelling: Comment on: “Mathematical models to characterize early epidemic growth: A Review" by Gerardo Chowell et al.

Leon Danon[1*], Ellen Brooks-Pollock[2]

[1]School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom, BS8 2BN

[2]Health Protection Research Unit, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom, BS8 2BN

*Corresponding author: Leon Danon, l.danon@bristol.ac.uk.

In their review, Chowell et al. consider the ability of mathematical models to predict early epidemic growth [1]. In particular, they question the central prediction of classical differential equation models that the number of cases grows exponentially during the early stages of an epidemic. Using examples including HIV and Ebola, they argue that classical models fail to capture key qualitative features of early growth and describe a selection of models that do capture non-exponential epidemic growth. An implication of this failure is that predictions may be inaccurate and unusable, highlighting the need for care when embarking upon modelling using classical methodology. There remains a lack of understanding of the mechanisms driving many observed epidemic patterns; we argue that data science should form a fundamental component of epidemic modelling, providing a rigorous methodology for data-driven approaches, rather than trying to enforce established frameworks. The need for refinement of classical models provides a strong argument for the use of data science, to identify qualitative characteristics and pinpoint the mechanisms responsible for the observed epidemic patterns.

Exponential growth in classical ordinary differential equation models results from the assumption that at the start of an epidemic there is an unlimited pool of susceptible individuals, due to the fact that everyone is assumed to be in contact with everyone else[2]. In contrast, models that include a notion of individual identity tied to spatial location, be it on a lattice, in a network or a metapopulation, impose local depletion of susceptible individuals that occurs much earlier in the epidemic [3,4], leading to sub-exponential epidemic growth. In practice, partially observed epidemics may obscure early growth patterns; this uncertainty is most acute during the early part of an outbreak of a novel pathogen [5,6]. This leads to uncertainty in model structure, parameter estimation and therefore model predictions that wanes as the epidemic grows, and estimates of burden become more accurate [7]. Novel data streams, such as social media platforms or web searches [8] hold the promise of timely and accurate estimates, but only when appropriately used.

Behavioural data sources can be incorporated into existing frameworks, or be used to inspire new modelling structures; spatial and movement data readily inform metapopulation models; social network data can be used in individual based models; temporally explicit data can inform behaviour-change models, and so on. Recently, detailed, individually explicit behavioural datasets, such as mobile phone records or cattle tracing systems have been used to provide some of this information and bring into focus the role of the individual [9]. But, repurposed data of this type often leaves gaps in understanding that can only be filled with targeted data collection. The current trend of digitising our lives through wearable devices, ubiquitous computing, and digital city initiatives is providing vast quantities of data on human behaviour on an ever-increasing scale. Data analytic tools that can handle such volumes are required, but may introduce further uncertainties into modelling predictions which need to be managed with validation across data sources[10,11].

How to compare the predictions of data-hungry models with uncertain incidence remains a significant barrier to identifying the mechanisms responsible for early epidemic patterns. Recent developments in complex model fitting and model choice may hold the promise of picking apart the most likely mechanisms. Complex data combined with complex models inevitably leads to challenges in robust model fitting and parameter estimation, however. Likelihood methods are the ‘gold standard’, but they can be difficult or impossible to implement, so simulation-based approximation methods are increasingly being used [12,13]. Computational overhead can be a limiting factor for complex model fitting and there are exciting developments involving Bayesian emulation [14] and Laplacian approximations [15] that are making complex model fitting feasible.

Together, advances in data collection and analysis, model development and fitting can provide the evidence needed to go beyond phenomenological descriptions of early epidemic growth and disentangle the driving mechanisms, but not without a trans-disciplinary effort. Combining approaches from data science with classical epidemiology is an exciting research direction and has the potential to revolutionise public health care.

Acknowledgements
LD was funded by PEPFAR modelling supplement and Canadian Institute of Health Research; EBP was funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Evaluation of Interventions. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

[1] Chowell G, Sattenspiel L, Bansal S, Viboud C. Mathematical models to characterize early epidemic growth: A Review. Physics of Life Reviews (2016) [in this issue].

[2] Kermack WO, McKendrick AG. A Contribution to the Mathematical Theory of Epidemics. Proc R Soc London A 1927;115:700–21.

[3] Keeling MJ, Eames KTD. Networks and epidemic models. J R Soc Interface 2005;2:295–307. doi:10.1098/rsif.2005.0051.

[4] Danon L, Ford AP, House T, Jewell CP, Keeling MJ, Roberts GO, et al. Networks and the epidemiology of infectious disease. Interdiscip

Perspect Infect Dis 2011;2011:284909. doi:10.1155/2011/284909.

[5] Lipsitch M, Riley S, Cauchemez S, Ghani AC, Ferguson NM. Managing and reducing uncertainty in an emerging influenza pandemic. N Engl J Med 2009;361:112–5. doi:10.1056/NEJMp0904380.

[6] Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, et al. Transmission dynamics and control of severe acute respiratory syndrome. Sci (New York, NY) 2003;300:1966–70.

[7] Baguelin M, Hoek AJ Van, Jit M, Flasche S, White PJ, Edmunds WJ. Vaccination against pandemic influenza A/H1N1v in England: a real-time economic evaluation. Vaccine 2010;28:2370–84. doi:10.1016/j.vaccine.2010.01.002.

[8] Althouse BM, Scarpino S V, Meyers LA, Ayers JW, Bargsten M, Baumbach J, et al. Enhancing disease surveillance with novel data streams: challenges and opportunities. EPJ Data Sci 2015;4:17. doi:10.1140/epjds/s13688-015-0054-0.

[9] Keeling MJ, Danon L, Vernon MC, House TA. Individual identity and movement networks for disease metapopulations. Proc Natl Acad Sci 2010;107:8866–70.

[10] Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider CM, Blondel V, et al. On the Use of Human Mobility Proxies for Modeling Epidemics. PLoS Comput Biol 2014;10:e1003716. doi:10.1371/journal.pcbi.1003716.

[11] Read JM, Edmunds WJ, Riley S, Lessler J, Cummings DAT. Close encounters of the infectious kind: methods to measure social mixing behaviour. Epidemiol Infect 2012;140:2117–30. doi:10.1017/S0950268812000842.

[12] Conlan AJK, McKinley TJ, Karolemeas K, Pollock EB, Goodchild A V., Mitchell AP, et al. Estimating the Hidden Burden of Bovine Tuberculosis in Great Britain. PLoS Comput Biol 2012;8:e1002730. doi:10.1371/journal.pcbi.1002730.

[13] Brooks-Pollock E, Roberts GO, Keeling MJ. A dynamic model of bovine tuberculosis spread and control in Great Britain. Nature 2014;511:228–31. doi:10.1038/nature13529.

[14] Andrianakis I, Vernon IR, McCreesh N, McKinley TJ, Oakley JE, Nsubuga RN, et al. Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda. PLoS Comput Biol 2015;11:e1003968. doi:10.1371/journal.pcbi.1003968.

[15] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc Ser B (Statistical Methodol 2009;71:319–92. doi:10.1111/j.1467-9868.2008.00700.x.